Sous la direction du RECIF

Réseau d'Epidémiologie Clinique International Francophone

Recherche Clinique

Penser, Réaliser, Publier

- Ouvrage de référence pour les étudiants et les professionnels de la santé
- Bases méthodologiques des principaux types d'études
- Finalités cliniques illustrées par des exemples

Editions Abatos

Préfaces et avant-propos

Préface

Pendant le week-end du 11 novembre 1986, Charles MERIEUX avait pris l'initiative de convoquer un séminaire international aux Pensières, lieu magique et symbolique pour lui, afin de faire connaître à la communauté médicale lyonnaise l'existence d'une démarche née outre-Atlantique qui était intitulée **clinical epidemiology**. Il avait un contact avec l'**IN**ternational **CL**inical **E**pidemiology **N**etwork (INCLEN) -organisation soutenue par la Fondation Rockefeller et qui assurait la diffusion internationale de cette démarche. A ce séminaire étaient convoqués, d'une part un certain nombre d'universitaires responsables de l'enseignement et responsables administratifs d'activité lyonnaise, et en face, une délégation venant des Etats-Unis d'Amérique qui souhaitait détailler le processus.

Seul Charles Mérieux était capable de monter une réunion comme celle-ci et de dominer très vite l'intérêt qu'il y avait à se rapprocher de la structure internationale.

Assez rapidement, beaucoup d'entre nous fûmes convaincus par l'intérêt de cette approche pédagogique, tout en regrettant très fortement l'intitulé. En effet, épidémiologie clinique ne veut pas dire grand-chose aux yeux d'un français qui, par contre, gardait en mémoire la méthode de la MEDECINE EXPERIMENTALE définie par Claude BERNARD comme un élément essentiel de la méthodologie en expérimentation clinique.

Car c'est bien de cela qu'il s'agit : **l'épidémiologie clinique** correspond à la méthodologie de la recherche clinique.

Très vite, nous pûmes sortir de cette querelle sémantique et concevoir combien les choses pouvaient être utiles en pratique, notamment pour l'avenir de la recherche dans l'entité hospitalo-universitaire de Lyon.

La démarche s'est assez rapidement déclinée en termes de formateurs. En effet, la médecine expérimentale avait disparu du cadre de l'Université depuis une dizaine d'années et, par conséquent il fallait repartir à zéro. La seule façon de faire était d'envoyer un certain nombre de jeunes collègues pleins d'allant et d'avenir se former dans les différents centres, que ce soit au Canada, aux Etats-Unis d'Amérique ou même en Australie.

Le nerf de la guerre manquait et, comme toujours, généreusement, la Fondation Marcel Mérieux a injecté des financements pour faire démarrer les choses.

La chance a fait que des circonstances favorables ont permis au Ministère de l'Education Supérieur et de la Recherche de s'intéresser au problème. Ainsi, 11 futurs professeurs de l'université ont pu accéder à cette formation complémentaire pendant des périodes dépassant souvent 12 mois.

Les choses se sont mises en place, après le retour des diplômés. Les Hospices Civils de Lyon ont apporté une contribution significative au développement de la nouvelle discipline, parallèlement à l'Université, dans le cadre du Département d'Information Médicale (DIM).

Je ne reviendrai pas sur les années qui ont suivi et qui ont vu essaimer cette dynamique de formation, hors de l'hexagone, notamment en Afrique du Nord et dans les pays francophones de l'Est de l'Europe.

Un corps de doctrine s'est peu à peu mis en place et a donné lieu à l'édition d'un premier ouvrage en 1995. Il était absolument essentiel de remettre dans l'actualité tous les chapitres déjà traités et de les compléter. Ce fut le travail accompli par Yves MATILLON, le Président du Réseau d'Epidémiologique Clinique International Francophone et par Hélène PELLET qui est l'animatrice permanente du RECIF. Je les remercie ainsi que tous ceux qui ont contribué par leurs prestations à ce type d'action.

Doyen René Mornex
Membre de l'Académie Nationale de Médecine

PREFACE de l'ouvrage *La Recherche Clinique, de l'idée à la publication* de 1995

C'est avec enthousiasme que j'introduis ce premier volume d'épidémiologie clinique qui, avec le diplôme universitaire de l'Université Claude Bernard, est un élément de vie publique du RECIF — Réseau d'Epidémiologie Clinique International Francophone —créé par la Fondation Marcel Mérieux en 1987.

Cette nouvelle activité complète celle que nous poursuivons depuis 10 ans aux Pensières, Centre Collaborateur OMS, pour former des épidémiologistes de terrain — les "épitériens" — en liaison avec le CDC d'Atlanta : l'IDEA (Institut d'Epidémiologie Appliquée) est organisé dans ce but avec l'Ecole de Rennes et la Direction Générale de la Santé Publique ; c'est ainsi que nous avons formé plusieurs centaines d'épidémiologistes. Ces épitériens sont en poste dans les structures françaises de la santé, notamment le Réseau National de Santé Publique.

A l'origine du programme d'épidémiologie clinique, le rapport de John Evans rendait compte en 1980 de la distance entre les préoccupations de santé publique et la pratique clinique quotidienne des médecins, et Kerr White préconisait un nécessaire rapprochement entre l'école de santé publique et l'école de médecine. De plus, dans de nombreux articles, Thomas Chalmers estimait à 10 ans, le délai d'intégration des retombées de la recherche dans la pratique clinique courante.

Les besoins de rapprochement étaient ressentis. La recherche médicale clinique et biologique devait engendrer une pratique médicale s'appuyant sur des démarches diagnostiques et thérapeutiques validées. Des connaissances plus multidisciplinaires, notamment en économie et en sociologie, devaient permettre à la pratique clinique de s'exercer avec aussi une préoccupation de santé publique. Posséder un langage commun, plus objectif, plus statistique devenait alors encore plus nécessaire afin de comparer et d'évaluer les observations. Finalement, une méthodologie d'épidémiologie clinique regroupant l'ensemble de ces savoirs était proposée par les pionniers de l'Université Mc Master au Canada.

Les pays anglophones mettaient sur pied des enseignements d'épidémiologie clinique au sein même des écoles de médecine, inclus dans les programmes de chaque spécialité médicale. Et avec l'aide de la Fondation Rockefeller, quatre centres universitaires dispensaient cet enseignement, à Mc Master (Canada), à Philadelphie (Pennsylvanie), à Chapel Hill (Caroline du Nord), et à Newcastle (Australie). Concrétisés dans le réseau INCLEN (INternational CLinical Epidemiology Network) et grâce à un soutien important logistique et financier des pays anglophones, ces programmes étaient disséminés dans les écoles de médecine de 17 pays du monde anglophone. C'est dans ce contexte que la Fondation Marcel Mérieux a eu la chance d'intervenir et de proposer le développement de l'enseignement de l'Epidémiologie Clinique pour les facultés de médecine francophones.

C'est actuellement à l'Université Claude Bernard que se développe ce programme spécifique. Je remercie nos confrères lyonnais dont ce livre reflète le rayonnement en souhaitant qu'il s'étende au centre méditerranéen d'épidémiologie. Je m'étais engagé à Tunis à le créer dans le cadre de l'année Louis Pasteur sans oublier qu'il y a 100 ans, Marcel Mérieux était à ses côtés pour organiser avec Elie Metchnikoff et Alexandre Yersin les cours de microbie technique sous la direction d'Emile Roux.

Docteur Charles Mérieux

NOTE DES AUTEURS ET REMERCIEMENTS

Cet ouvrage constitue une introduction à la recherche clinique. Il présente les bases méthodologiques des principaux types d'études avec leurs finalités cliniques, illustrées de manière pratique par des exemples. Il est destiné aux professionnels de la santé et aux étudiants souhaitant faire de la recherche clinique.

Dans une démarche progressive, il se compose de quatre parties :

- Formulation de la question de recherche, en justifiant le projet à l'aide de la revue de la littérature et en envisageant les aspects éthiques ;
- Description des grands types d'études à la disposition du clinicien pour élaborer son protocole de recherche ;
- Aspects techniques de l'élaboration du protocole ;
- Exécution du protocole et présentation des résultats.

La recherche documentaire et l'analyse critique de la littérature doivent intéresser particulièrement les étudiants qui ont dorénavant ce thème nouveau à leur programme. Les biostatistiques, incontournables, sont présentées de façon aussi simple que possible, soutenues par des exemples. Le chapitre *Ethique* envisage les données les plus récentes -y compris dans le domaine international. Une attention particulière est portée aux finalités de la recherche clinique : points clés de la validité interne d'une étude, applicabilité des résultats et valorisation de la recherche.

Les auteurs ont mis leur expérience dans cet ouvrage qu'ils souhaitent au service de la recherche clinique. Les suggestions sont bienvenues. Les coordonnées des collaborateurs de l'ouvrage sont données au début du livre.

Ce livre doit beaucoup aux personnes et institutions suivantes :

- Le Professeur Mornex, le Professeur Zech (Université Claude Bernard), le Docteur Charles Mérieux et le Docteur Caroline Dupuy (Fondation Marcel Mérieux), le Professeur Hélène Pellet et le Professeur Yves Matillon, ainsi que madame Nicole Sallet (Université Claude Bernard), qui ont été à l'origine de la création du Réseau d'Épidémiologie Clinique International Francophone et de sa mise en œuvre ;

- Les soutiens à la création et financeurs du RECIF : les principaux ministères, en particulier le ministère de la Recherche et le ministère des Affaires étrangères, l'Université Claude Bernard à Lyon, les Hospices Civils de Lyon, la ville de Lyon, le Conseil régional du Rhône-Alpes et la Fondation Marcel Mérieux ;

- les membres des universités affiliées à l'INCLEN qui ont été à l'initiative de ce programme pédagogique à l'origine du mouvement de l'*evidence based medicine*, et en particulier le Docteur Halstedt (Fondation Rockefeller), les Docteurs Eisenberg et Ström (Université de Pennsylvanie), les Docteurs Suzanne et Robert Fletcher et les Docteurs Harris et Laura Sadowski (Université de Caroline du Nord), les Docteurs Tugwell et Browman (Université MacMaster, Ontario, Canada), les Docteurs Kellerman et Heller (Université de Newcastle, Australie), ainsi que le Docteur Claire Bombardier (Université de Toronto, Ontario, Canada).

- les membres actuels du réseau et plus particulièrement le Doyen Honoraire Hélène Pellet et Madame Charlotte Barthe notamment pour la relecture des épreuves, les Doyens des universités médicales d'Algérie, du Maroc et de Roumanie.

AUTEURS

ANDREJAK Claire, Maître de Conférences des Universités - Praticien Hospitalier en Pneumologie, CHU Nord, Place Victor Pauchet, 80000 AMIENS

BORY Eric-Nicolas, Praticien Hospitalier en odontologie, Centre Hospitalier Spécialisé Le Vinatier, 95 Boulevard Pinel, 69677 Bron Cedex, INCLEN Fellow 1991, CERTC Hamilton, Ontario, Canada

BOUËSSEAU Marie-Charlotte, Médecin responsable de l'unité Ethique et Santé, Organisation mondiale de la Santé, Ethics, equity, Trade and Human Rights, IER / ETH, 20 av Appia, CH 1211 Genève Suisse, http://www.who.int/ethics

BURON Catherine, Docteur en Economie de la santé, Cellule Innovation (DRCI) - Unité d'Evaluation médico-économique (Pôle IMER), Hospices Civils de Lyon - 162 Avenue Lacassagne - 69424 LYON Cedex 03

CHAPUIS François, Praticien Hospitalier en Epidémiologie, Economie de la Santé, Prévention, Unité de Recherche Clinique des Hospices Civils de Lyon, 162 Avenue Lacassagne - 69424 LYON Cedex 03. INCLEN Fellow, 1991, CERTC Philadelphie, Pennsylvanie, Etats-Unis.

CHAPURLAT Roland, Professeur des Universités - Chef du Service de Rhumatologie de l'Hôpital E Herriot, Lyon, Directeur de l'unité INSERM 831, Directeur du Centre National de Référence sur les Dysplasies Fibreuses des Os, Pavillon F, Hôpital E Herriot, 69437 Lyon cedex 03

COLIN Cyrille, Professeur des Universités - Praticien Hospitalier en Epidémiologie, Economie de Santé, Prévention, Responsable du Pôle Information Médicale Evaluation Recherche - Hospices Civils de Lyon, 162 avenue Lacassagne, 69003 Lyon

DELAHAYE François, Professeur des Universités - Chef du service de Cardiologie A, Hôpital Louis Pradel, 28, avenue du Doyen Lépine, 69677 - Bron Cedex. INCLEN Fellow, 1989, CERTC Chapel Hill, Caroline du Nord, Etats-Unis

DUCROIX Jean-Pierre, Professeur des Universités - Chef du service de Médecine Interne, CHU Nord, Place Victor Pauchet, 80000 AMIENS

DUHAUT Pierre, Professeur des Universités - Praticien Hospitalier en Médecine Interne, CHU Nord, Place Victor Pauchet, 80000 AMIENS. INCLEN Fellow, 1991, CERTC Chapel Hill, Caroline du Nord, Etats-Unis

ECOCHARD René, Professeur des Universités - Chef du service de Biostatistique, Praticien Hospitalier en Biostatistique et informatique médicale, 162 Avenue Lacassagne - 69424 LYON Cedex 03

HODGKINSON Isabelle, Praticien Hospitalier en Médecine Physique et Réadaptation Pédiatrique aux Hospices Civils de Lyon, Hôpital Femme Mère Enfant, 59 Boulevard Pinel, 69677 Bron Cedex

JAISSON-HOT Isabelle, Praticien Hospitalier, Responsable de l'Unité Hospitalière d'Information Médicale, d'Evaluation et de Recherche de l'Hôpital Edouard Herriot, Pôle Information Médicale Evaluation Recherche des Hospices Civils de Lyon, 162 Avenue Lacassagne, 69424 Lyon Cedex 03

LANDRIVON Gilles, Praticien Hospitalier en gynéco-obstétrique, actuellement en poste à Niamey pour l'Organisation mondiale de la Santé, où il est responsable des programmes de santé maternelle pour le Niger et les pays de la sous région. INCLEN Fellow, 1990, CERTC Newcastle, Australie

MAISONNEUVE Hervé, Professeur associé de Santé publique, Université Paris-Sud 11. Il a été rédacteur en chef de European Science Editing (2000-2006). Il anime un blog sur la rédaction médicale www.h2mw.eu

MATILLON Yves, Professeur des Universités - Praticien Hospitalier, Directeur de l'Institut des Sciences et Techniques de la Réadaptation de l'Université Lyon1, 8 avenue Rockefeller, 69373 Lyon Cedex 08, Directeur de recherche de l'EA4129 Hôtel Dieu, Chargé de mission ministériel pour le développement de l'évaluation des compétences médicales. INCLEN fellow, 1988, CERTC Université de Toronto, Ontario, Canada

MOATTI Jean-Paul, Professeur d'Economie de la Santé à l'Université de la Méditerranée (Aix-Marseille II) et Directeur de l'Unité Mixte de Recherches 912 SE4S (Sciences économiques et sociales, Systèmes de Santé, Sociétés) de l'INSERM/IRD/Université, 23 rue Stanislas Torrents, 13006 MARSEILLE, http://www.se4s-orspaca.org

PELLET Hélène, Ancien Professeur des Universités - Praticien Hospitalier en Histologie, Embryologie et cytogénétique, Doyen Honoraire de la Faculté de Médecine Lyon Grange Blanche, Responsable du RECIF, Université Claude Bernard Lyon 1 - 8 av Rockefeller - 69373 Lyon Cedex 08

RABILLOUD Muriel, Maître de Conférences des Universités - Praticien Hospitalier en biostatistique, Service de Biostatistique des Hospices Civils de Lyon, Laboratoire Biostatistique Santé, 162 Avenue Lacassagne, 69424 Lyon Cedex 03

SCHMIDT Jean, Praticien Hospitalier en Médecine Interne, CHU Nord, Place Victor Pauchet, 80000 AMIENS

SCHOTT Anne-Marie, Professeur des Universités - Praticien Hospitalier en Epidémiologie, Economie de la Santé, Prévention, Pôle Information Médicale, Evaluation, Recherche, Hospices Civils de Lyon, 162 avenue Lacassagne, 69424 Lyon Cedex 03. INCLEN fellow, 1992, CERTC Toronto, Ontario, Canada

SORENSEN Henrik, Professeur des Universités, spécialiste en médecine interne et gastroentérologie, Department of Clinical Epidemiology, Aarhus University Hospital, Olof Palmes Allé 43 – 45, DK-8200 Aarhus N, Denmark

TERMOZ Anne, Attachée de Recherche Clinique, Pôle Information Médicale Evaluation Recherche - Hospices Civils de Lyon, 162 Avenue Lacassagne 69424 Lyon Cedex 03

VUILLEROT Carole, Praticien Hospitalier en Médecine Physique et Réadaptation

Pédiatrique aux Hospices Civils de Lyon, Hôpital Femme Mère Enfant, 59 Boulevard Pinel, 69677 Bron Cedex

Avant-propos

Tous les pays développés sont confrontés à des difficultés économiques similaires, en particulier celles qui sont liées au système de protection sociale. Les sociétés demandent légitimement des comptes pour savoir si les ressources sont utilisées de manière appropriée pour délivrer des soins de la meilleure qualité possible.

La pratique de la médecine évolue. Ce changement est tellement profond qu'on peut le considérer comme un véritable changement de paradigme dans les modalités de la pratique de la médecine. Les bases du paradigme nouveau reposent sur les développements de la recherche clinique et l'intégration des résultats de cette recherche dans la pratique quotidienne du médecin, et des autres professionnels de santé. Pour cela, il paraît nécessaire de développer une approche plus objective de la décision médicale.
Rendre crédibles des études de recherche clinique et d'évaluation, par la clarté des objectifs, la rigueur de la méthode, l'utilité des résultats, nécessite temps, compétences et expérience pour les promoteurs.

Pour faire face à tous ces enjeux, la formation initiale des médecins est déterminante, du fait de la reproduction professionnelle des attitudes décisionnelles apprises initialement.

Il y a une vingtaine années, l'épidémiologie était une notion totalement nouvelle dont les buts et les méthodes étaient inconnus des médecins. Du fait de l'enthousiasme, de la compétence et de l'activité des épidémiologistes, les notions de raisonnement causal de l'épidémiologie descriptive et analytique, étayées par des bases méthodologiques solides, sont maintenant largement entrées dans le domaine public de la médecine.

L'enseignement médical traditionnel doit préparer et donner des exemples de cette médecine « fondée sur des faits » (*l'* « *évidence-based-medicine* » des Anglo-saxons). Il pratique par étapes successives : définition précise du problème du malade, informations nécessaires pour résoudre ce problème, interrogation efficace de la littérature médicale et sélection des meilleures études concernant le problème, détermination du niveau de preuves qui les qualifie, extraction du message clinique et application de ce message au problème du malade, capacité de présenter à des collègues d'une manière pertinente la logique de son raisonnement avec ses forces et ses faiblesses.

Cet ouvrage reprend cet enchaînement logique et actualise toutes les données, en tenant compte de l'expérience pratique de ceux qui travaillent dans ce domaine de la recherche clinique, mais aussi de leur expérience pédagogique (formation des professionnels de santé, formation post-doctorale, et lors des cursus de master…)

Plusieurs originalités apparaissent dans cette édition et notamment :

- l'élargissement « extra-lyonnais » des contributions des auteurs à la rédaction des chapitres de l'ouvrage et des sujets traités ;
- l'intérêt du contenu pour les étudiants en médecine, compte tenu de la mise en œuvre d'épreuves de « lecture critique d'articles » dites « LCA ».

Les auteurs sont des cliniciens, des méthodologistes ayant des expériences variées. Ils ont reçu pour la majorité d'entre eux, il y a quelques années, une formation universitaire complémentaire en France et à l'étranger. Cette dernière expérience revêt l'intérêt d'avoir été mise en œuvre dans des centres universitaires nord-américains (USA, Canada) et australiens dans le cadre d'une expérience pédagogique unique et importante, par ses enjeux et ses objectifs : le programme INCLEN (INternational CLinical Epidemiology Network). L'Université Claude Bernard et la Fondation Marcel Mérieux, ainsi que d'autres institutions, ont favorisé le

développement d'une expérience identique pour les pays francophones en constituant le RECIF : Réseau d'Épidémiologie Clinique International Francophone.

Notre souhait est que ce livre permette une meilleure compréhension de ces concepts, qu'il favorise l'adhésion des cliniciens à cette démarche, et qu'il permette la mise en œuvre de ce nouveau paradigme à travers un apprentissage socratique des objectifs, méthodes et discussion des résultats de la recherche clinique. Puisse l'enseignement initial dans les facultés de médecine françaises et francophones en bénéficier.

Professeur Yves MATILLON
Professeur d'épidémiologie clinique
à l'Université Claude Bernard – Lyon

Professeur Hélène PELLET
Doyen Honoraire de la faculté de Médecine
Lyon Grange Blanche
Université Claude Bernard

Première Partie : Justification et Formulation de la question de recherche

PREMIERE PARTIE: JUSTIFICATION ET FORMULATION DE LA QUESTION DE RECHERCHE

Si l'argent est le nerf de la guerre, il est le bulbe rachidien de la recherche, dont le clinicien est le cortex. L'argent existe. Il faut le trouver. Pour l'obtenir, il faut argumenter. C'est pourquoi cette première partie: "justification et formulation de la question de recherche" est fondamentale. Il faut se convaincre, d'abord, puis convaincre son entourage, les collaborateurs potentiels, les bailleurs de fond, que le projet de recherche qu'on a en tête mérite d'être réalisé. Il y a dans la justification une démarche de commercialisation car on doit prouver que la réponse à la question de recherche, apportée par le protocole que l'on souhaite exécuter, trouvera preneur sur le marché de la médecine. Et si c'est le cas, on peut trouver un soutien pour l'étude.

On doit donc insister sur le caractère innovant du projet, l'importance du problème considéré, la taille de la population considérée, mais aussi l'impact en termes de conséquences sociales, politiques et économiques. Un projet de recherche clinique a d'autant plus de chances de trouver des commanditaires qu'il présente un caractère multi-disciplinaire.

La logistique, tout comme la méthodologie, est une chose. L'éthique en est une autre. Quatre principes fondamentaux guident le chercheur, pour le plus grand respect des sujets participant à la recherche. Ce sont les principes de l'intérêt et du bénéfice de la recherche, de son innocuité, du respect des personnes, et enfin de la justice. Il est naturel de se demander si le processus de randomisation est éthique pour le patient. Ne risque-t-il pas d'être privé d'un bénéfice éventuel parce qu'il s'est trouvé dans le "mauvais" groupe ? Mais alors il doit être aussi naturel de s'interroger sur le caractère éthique de pratiques médicales qui n'ont jamais fait l'objet de protocoles d'évaluation ou de recherche. Si l'éthique est un garde-fou de la recherche, elle en est aussi par elle-même une justification.

QUESTION DE RECHERCHE, HYPOTHESE, DIFFERENTS TYPES D'ETUDES.

Roland Chapurlat

Tout projet de recherche clinique naît d'une question que se pose un investigateur. La première étape d'un projet est d'avoir une idée de recherche et de la formuler de façon correcte. Le passage de l'idée à la formulation de la question consiste simplement à exprimer ce qu'on a l'intention de faire et pourquoi.

Il faut d'abord vérifier que le projet de recherche est utile, nouveau, éthique, et faisable. Dans la formulation de la question de recherche doivent figurer la nature de la population à étudier, le facteur étudié et son critère de jugement, la nature de la comparaison s'il y en a une, et le type d'étude envisagé : étude transversale, étude cas-témoins, étude de cohorte, essai randomisé.

ANATOMOPHYSIOLOGIE DE LA RECHERCHE : COMMENT CELA FONCTIONNE-T-IL ?

La structure du projet de recherche est définie dans le protocole, qui doit toujours être rédigé avant de commencer l'étude. Le protocole est indispensable pour demander des crédits, la promotion de l'étude, mais il est aussi vital pour obliger le chercheur à organiser ses idées de manière logique et efficace avant de commencer le projet.

La question de recherche

C'est l'incertitude que le chercheur veut lever, c'est l'objectif de l'étude. La question est souvent générale au début, mais doit être progressivement raffinée. Par exemple :

Les Français devraient-ils manger plus de fromage ?

La question doit être davantage centrée. Ainsi :

Le fait de manger plus de fromage réduit-il le risque de fracture ?

Est-ce que le fromage n'apporte pas trop de cholestérol ?

Est-ce que des suppléments de calcium n'apportent pas le même bénéfice anti-fracturaire que le fromage?

Est-ce que les suppléments en calcium ne font pas sentir le fromage ?

La justification (ou le contexte)

C'est la première partie du protocole, qui justifie la question : on fait le point sur le domaine, grâce à la littérature, et on met en évidence les questions qui se posent encore.

Organisation : les différents types d'études

On distingue les études observationnelles et les études d'intervention.
Les **études observationnelles** peuvent être **transversales**, lorsque l'observation des sujets est faite en une fois. Les **études de cohorte** observent les sujets au cours du temps. Les études de cohorte sont prospectives si on recueille des données au départ puis on suit les sujets. Elles sont rétrospectives si le recueil des données se fait après coup. Les **études cas-témoins** permettent de comparer des sujets qui ont une maladie avec un autre groupe qui ne l'a pas.
Les **études d'intervention**, ou **essais cliniques**, testent l'effet d'un médicament ou d'une intervention non médicamenteuse.

Les sujets
Les critères d'inclusion ou d'exclusion définissent la population cible de l'étude. Le nombre de sujets à enrôler est également un élément important.

Les variables

Ce sont les informations à recueillir. Dans une étude analytique, il y aura des variables prédictives (qui ont le potentiel d'être causales) et des variables de type critère de jugement (c'est le résultat).

Les statistiques

Il faut calculer la taille de l'échantillon nécessaire avant d'avoir écrit tout le protocole, car si l'étude n'est pas faisable à cause d'une taille inaccessible, il ne faut pas perdre son temps. Le plan d'analyse statistique doit être prévu à l'avance.
Cette analyse repose sur la notion d'hypothèse, qui doit être clairement énoncée dès le départ, en général sous la forme de la question de recherche.

LA QUESTION DE RECHERCHE

La question provient des problèmes que l'on peut rencontrer dans sa propre pratique clinique, ou émerger de travaux de recherche antérieurs personnels ou d'autres chercheurs. Souvent un jeune chercheur n'aura pas suffisamment de recul pour concevoir une question et aura besoin de l'aide d'un sénior.

Les origines de la question

Connaître la littérature
Avant de se lancer dans une nouvelle étude, il faut connaître la littérature sur le sujet, et dans l'idéal réaliser une revue systématique. Il faut assister à des réunions dans le domaine et discuter avec des experts de la question.

Etre ouvert aux nouvelles idées, nouvelles techniques
Il faut écouter des conférences et garder un esprit critique à l'égard des idées communément admises. Par exemple, on a longtemps considéré que la sciatique par hernie discale devait être traitée par le repos au lit, de façon à diminuer la pression sur le disque inter-vertébral. En fait, plusieurs essais cliniques contrôlés ont montré que cela n'améliorait pas le pronostic.

Garder son imagination en mouvement
L'observation clinique des patients apporte souvent des idées. L'enseignement, lors de sa préparation ou grâce aux questions inquisitrices de certains étudiants, peut également être la source de nouvelles idées. Il faut ensuite être créatif pour appliquer l'idée, tenace car les difficultés pratiques sont nombreuses et ne pas redouter la critique.

Choisir un mentor
Il apportera l'expérience et la connaissance d'un sujet. Il évitera de commettre certaines erreurs et facilitera l'accès aux financements nécessaires.

Concevoir la bonne question
Une bonne question doit être **FINE** : **F**aisable, **I**ntéressante, **N**ouvelle, **E**thique.

Faisable

Nombre de sujets
Trop d'études ne peuvent pas répondre à la question car la taille d'échantillon n'est pas suffisante. Le calcul doit être fait d'emblée, de façon réaliste, et il ne faut pas hésiter à reformuler la question de façon moins ambitieuse, ou abandonner le sujet si l'on ne peut pas répondre.

Aspects techniques
Il faut avoir l'expérience suffisante, l'équipement technique pour mesurer les variables d'intérêt et analyser les données. Travailler avec des co-investigateurs qui sont plus spécialisés, notamment un biostatisticien, se révèle souvent indispensable, de façon à disposer de toutes les compétences techniques souhaitables.

Coût
Il faut soigneusement estimer le coût de chaque composante du projet, sachant que les coûts réels dépassent souvent les estimations.

Rester concentré
Il ne faut pas chercher à répondre à trop de questions à la fois.

Intéressante

Il ne faut pas être le seul à trouver la question intéressante. Il en parler à ceux qui sont familiers du domaine. Il faut ensuite imaginer ce que vont apporter les résultats, dans quelle mesure ils feraient progresser la connaissance, influenceraient les pratiques, ou guideraient de nouvelles recherches. Quand cette pertinence n'est pas évidente, il faut rediscuter la question.

Nouvelle

On voit trop d'études qui ont déjà été conduites… Une étude doit apporter une contribution nouvelle. C'est l'étude de la littérature et la discussion avec les experts qui permettra de ne pas réinventer la poudre.

Ethique

Si l'étude implique trop de risques physiques ou de violation de la vie privée, il faut revoir la conception de l'étude. En outre, une étude mal conçue, qui ne permet pas de répondre correctement à la question (taille d'échantillon insuffisante…) n'est pas éthique non plus.

L'HYPOTHESE

C'est une supposition, née d'une réflexion ou d'une observation, qui conduit à des prédictions, éventuellement réfutables. Une étude va déterminer si cette supposition est une description exacte de la relation qui existe entre les facteurs que l'on étudie.
En général, on pose une hypothèse nulle : il n'y a pas d'association entre le facteur étudié et le critère de jugement. L'hypothèse alternative est qu'il y a une relation entre le facteur étudié et le critère de jugement.

Caractéristiques d'une bonne hypothèse

Une bonne hypothèse est basée sur une bonne question. Elle doit être simple, spécifique, et énoncée à l'avance.

Simple par rapport à complexe
Une hypothèse simple comprend une variable prédictrice et un critère de jugement : *la ménopause précoce s'associe à un risque accru de fracture par fragilité*.
Une hypothèse complexe contient plus d'une variable prédictrice : *la ménopause précoce, l'âge et la minceur sont associés à un risque accru de fracture par fragilité* ; ou plusieurs critères de jugement : *le tabagisme s'associe au risque cardiovasculaire et au cancer du poumon.*
Les hypothèses complexes ne peuvent pas être testées avec un test statistique simple.

Spécifique par rapport à vague
Une hypothèse spécifique ne laisse pas d'ambiguïté sur les sujets et les variables ou comment le test statistique sera appliqué.

Types d'hypothèses

Hypothèse nulle et hypothèse alternative
L'hypothèse nulle indique qu'il n'y a pas d'association entre la variable prédictrice et le critère de jugement dans la population. L'hypothèse nulle est la base formelle d'un test statistique. Celui-ci aide à estimer la probabilité que l'association observée dans une étude est due au hasard.
La proposition qu'il y a une association est l'hypothèse alternative. Elle ne peut pas être testée directement. Elle est acceptée par défaut si le test de signification statistique rejette l'hypothèse nulle.

Hypothèses alternative uni- et bilatérale
Une hypothèse unilatérale spécifie la direction de l'association entre la variable prédictrice et le critère de jugement. L'hypothèse que la ménopause augmente le risque de fracture par fragilité est unilatérale.
Une hypothèse bilatérale spécifie qu'une association existe ; elle ne précise pas la direction. L'hypothèse que la ménopause s'associe au risque de fracture est bilatérale, car elle ne précise pas la direction.

LES DIFFERENTS TYPES D'ETUDES (Figure 1)

Les études descriptives

Elles rendent compte d'un phénomène, de sa fréquence, de sa distribution et de son évolution. Elles fournissent des données quantitatives sur la répartition d'une maladie, d'un facteur de risque. Il s'agit des études transversales ou études de prévalence, dans lesquelles une mesure est faite à un temps donné.
Elles posent le problème de l'absence d'information sur la chronologie des événements. Faciles à réaliser, elles permettent de formuler des hypothèses à tester dans des études analytiques

Les études analytiques

Elles déterminent le rôle d'un ou plusieurs facteurs dans l'étiologie ou le traitement d'une maladie.

Les études d'observation

Les études cas-témoins

Les sujets présentent une maladie. Les témoins n'ont pas cette maladie. On explore ensuite les facteurs de risque ou d'exposition passés, et l'on compare leur fréquence chez les cas et les témoins. C'est une étude rétrospective, de ce fait soumise à certains biais (par exemple de mémoire), mais bien adaptée à l'étude d'événements assez rares.

Les études de cohorte

Les sujets sont sélectionnés en fonction de l'exposition à certains facteurs de risque ou d'exposition et sont suivis pendant une période de temps donnée pour observer les conséquences de cette exposition.

Les études expérimentales

On soumet des sujets à une intervention et on observe ses effets. Le modèle de référence de l'étude expérimentale est l'essai clinique randomisé, contrôlé par rapport au placebo ou à un traitement de référence. Néanmoins, il existe plusieurs modèles :

- Sans groupe témoin, ou en permutation croisée (cross-over) : le sujet est son propre témoin.
- Avec groupe témoin : l'essai randomisé est la référence, puisque la seule différence entre le groupe intervention et le groupe témoin est l'intervention, grâce à l'allocation aléatoire (randomisation). Le traitement reçu peut être connu du patient mais pas de l'investigateur (simple insu), inconnu des deux (double insu), ou inconnu du patient de et de l'évaluateur (qui est différent de celui qui traite, par exemple avec une procédure invasive ; c'est le pseudo-double insu).

On définit souvent un niveau de preuve en fonction du type d'étude réalisée, le sommet étant l'essai clinique randomisé contrôlé. Toutefois, le type d'étude dépend surtout de la question à laquelle il faut répondre, et doit donc y être adapté (Figure 2).

CONCLUSION

Quand la question est formulée, l'hypothèse soulevée et le type d'étude choisi, on peut rédiger précisément le protocole de recherche.

Chapitre II
La revue de la littérature

Gilles Landrivon, Pierre Duhaut, Hélène Pellet

L'accès à l'information est l'étape indispensable à l'élaboration et à la mise en œuvre de la question de recherche. Cette question a-t-elle déjà été envisagée, ou existe-t-il dans la littérature des données amenant à modifier la question de recherche ?
Il importe donc de trouver dans la littérature les informations pertinentes concernant la question de recherche, de s'assurer que ces informations sont exhaustives, puis de les analyser, ce qui fera l'objet d'une « lecture critique » (critical appraisal).

A. LA RECHERCHE DOCUMENTAIRE

La recherche documentaire doit être **exhaustive** (ne laissant pas échapper d'article utile à la question de recherche) et **pertinente** (évitant le plus possible les articles sans intérêt pour la recherche). Elle se fait en plusieurs étapes. Il faut :

- **Savoir ce que l'on recherche.** Il faut pour cela **définir le sujet** : évolution d'une maladie, étude de facteur(s) modifiant cette évolution…, préciser le **contexte** (type de patients concernés, âge, conditions sociales et géographiques…) et fixer les **limites dans le temps** de la recherche documentaire.
- **Choisir les sources documentaires appropriées**
- **Elaborer la stratégie de recherche,** et la mettre en œuvre.

I. Les sources documentaires

Elles ne sont pas exclusives les unes des autres (plusieurs sources peuvent en fait contenir un nombre important de revues identiques). Il importe donc d'utiliser dans un premier temps, la source documentaire la plus performante dans son domaine, avant d'élargir la recherche vers d'autres sources si nécessaire.

1. La recherche sur Internet

On utilise des moteurs généraux (Google, Yahoo…), des métamoteurs (Copernic), des répertoires (Evidence-Based Healthcare, CISmef…) ou des outils sélectifs (SUM-search) permettant d'interroger simultanément plusieurs sources. On appelle « filtres EBM » (Evidence-Based Medicine) les outils offrant des stratégies de recherche prédéfinies, et à l'aide de critères plus qualitatifs.
Pour un travail prolongé, il est généralement nécessaire de faire appel à une ou plusieurs bases payantes pour obtenir les articles dans leur texte intégral.

a. PubMed

La solution la plus performante dans les sciences du vivant et de la santé consiste à utiliser PubMed (www.pubmed.org), l'interface Internet de Medline dont l'accès est gratuit. Elaboré par la *National Library of Medicine* (Bethesda, USA), il y avait en Décembre 2007 16,8 millions de références (dont les plus anciennes remontent à 1865). Cinq cent mille références

en moyenne sont ajoutées chaque année. La majorité proviennent de la base de données Medline, couvrant la période 1966 à nos jours, et PubMed donne également accès à la base 'OldMedline', couvrant la période 1950-1966. 5200 périodiques provenant de 80 pays sont indexés. La plupart des revues référencées sont en langue anglaise, même lorsqu'elles proviennent de pays non anglophones, comme certaines revues scandinaves ou italiennes de bon niveau.

Outre sa gratuité, PubMed combine de nombreux avantages:

- Les revues référencées sont avant tout des **revues à comité de lecture**, publiant des articles revus par des experts reconnus par la revue comme compétents dans le domaine : il s'agit du 'peer-review', ou revue par les pairs.
- Les stratégies de recherche d'articles offertes par PubMed sont performantes, grâce notamment à un **riche thésaurus** et aux possibilités de combinaison de ces mots-clés.
- PubMed donne **accès aux résumés** en anglais de pratiquement tous les articles originaux référencés (moins souvent pour les revues générales ou les rapports de cas).
- PubMed donne **accès aux textes intégraux** d'articles par deux voies principales. Le texte intégral peut faire partie de la **banque d'articles PubMed**, et son accès est alors gratuit. Il peut aussi être mis en ligne par la revue, et PubMed associe à chaque référence **le lien vers la revue** et le texte intégral de l'article. Deux cas de figure peuvent alors se présenter : l'article est mis en ligne gratuitement par la revue (c'est le cas pour certaines grandes revues internationales, pour les articles publiés depuis plus de 6 mois, et pour les articles à répercussion jugée importante dans le domaine considéré) ou l'article peut n'être accessible que par voie d'abonnement payant à la revue, ou par achat en ligne de l'article considéré de façon isolée.

  Lorsque les articles sont disponibles en ligne, ils le sont souvent sous deux formats. Le **format pdf** (Adobe), (format à imprimer pour une lecture commode) et le **format html**, avec texte, graphiques, et illustrations séparés qui permet d'enregistrer facilement graphes et illustrations et de les intégrer dans une présentation de type PowerPoint. Certaines revues fournissent même pour leurs illustrations la diapositive toute faite avec texte et références, que l'on peut intégrer telle quelle dans une présentation.
- PubMed fournit souvent les **références bibliographiques** des articles trouvés, avec le lien hypertexte renvoyant à l'article correspondant et, le cas échéant, à son texte intégral s'il fait partie de la banque d'articles PubMed ou s'il est mis en ligne par la revue d'origine. L'exploration de ces références permet de trouver des articles intéressants non listés lors d'une première recherche.
- De plus en plus d'articles sont mis en ligne avant leur impression et leur publication papier : ils sont alors référencés avec la mention [**Epub ahead of print**], et peuvent être accessibles parfois 6 mois avant leur parution papier.
- Enfin, de plus en plus d'institutions (universités, structures de recherche, hôpitaux) ont actuellement des **abonnements électroniques collectifs** mis à disposition de leurs membres (étudiants, professionnels). Les publications sont alors accessibles par l'intermédiaire de sites spécialisés regroupant souvent toutes les publications d'une maison d'édition, et ces sites sont fournis en liens hypertexte par PubMed. L'accès à l'article original est dans ces conditions rapide et gratuit pour l'utilisateur s'il accède au site par l'intermédiaire d'un ordinateur appartenant au réseau de l'institution abonnée.

b. Deux outils « généraux » : Current Contents, Cochrane Library

Les Current Contents et Cochrane Library sont disponibles sur Internet, et aussi sur CD-Rom, mais avec l'inconvénient de mises à jour annuelles.

La liste de références peut être imprimée, et l'on peut rayer de cette liste les articles que l'on exclut.
Les Current Contents produits par l'*Institute for Scientific Information* (www.isinet.com) fournissent deux publications (parmi neuf) particulièrement adaptées à la recherche clinique, à publication hebdomadaire :

- CC « Life Sciences » fournissent le sommaire de 1350 périodiques,
- CC « Clinical Medicine » le sommaire de 1120 périodiques, dont certains se trouvent dans les 2 publications.

Certaines sources fournissent les références sans sélection qualitative. C'est le cas de Embase Medline et Pascal.
Les sources EBM effectuent une *sélection qualitative*. C'est le cas de la *Cochrane Library :*

- DARE (*Database of Abstracts of Reviews of Effectiveness)*
- CAT (*Critically Appraised Topics)*

c. Les stratégies de recherche prédéfinie

Clinical Queries utilisable sur Pubmed permet d'effectuer une recherche spécifique. L'inconvénient en est cependant le risque de manquer des références intéressantes.
Lorsque la recherche concerne une thématique particulière, il est bon de s'adresser à des **bases de données spécifiques**, par exemple :
- pour la **santé publique** : la *BDSP* (Banque de Données françaises en Santé Publique, (www.bdsp.tm.fr), d'accès gratuit, est gérée par l'Ecole Nationale de Santé Publique. Elle contient environ 15 000 documents indexés avec un Thésaurus.
- pour **l'économie médicale** :

▪ EED (*Economic Evaluation Database* du *National Health Service*)
▪ *Ecosanté* pour les données numériques en France
▪ Le CRD (Center for Reviews and Dissemination de York) (http://www.crd.york.ac.uk/crdweb/) présente des bases d'évaluation économique : NHS *Economic Evaluation Database* et des résumés d'analyses coût/bénéfice et coût/efficacité de pratiques médicales : base EED (NHS *Economic Evaluation Database)* consultable gratuitement.
▪ La *base CODECS (COnnaissance et DECision en économie de Santé,* (http://infodoc.inserm.fr/codecs/codecs.nsf), gérée par le Collège des économistes de la santé et l'INSERM, rapporte les études d'économie de la santé faites en France.

2. La recherche des articles dans une bibliothèque

Habituellement, les bibliothèques universitaires laissent aux chercheurs l'accès libre aux périodiques. Le chercheur peut donc consulter directement les périodiques et travailler sur place avec sa liste de références, éliminant rapidement les articles sans intérêt pour son étude, notant sur une fiche les quelques éléments à retenir, ou photocopiant les articles sur lesquels il aura à travailler plus longuement.

3. La recherche documentaire non publiée

Cet aspect doit être pris en compte. Si les connaissances en médecine ou dans d'autres secteurs d'activités scientifiques sont le plus souvent publiées, la non-publication de données n'est pas l'exception. Dans le secteur industriel, les résultats d'études cliniques, lorsqu'ils sont

négatifs, ne soulèvent pas l'enthousiasme de la part des promoteurs. Ceci est vrai pour les essais thérapeutiques négatifs pour une molécule testée. Les données produites existent cependant. Certains sujets sont sensibles au plan politique et, par exemple, les études et les données épidémiologiques qui concernent les contaminations, les modes de transmission et les modalités de prévention de maladies infectieuses ne sont pas toutes disponibles en temps réel…ou celles qui sont publiées sont parfois à analyser en tenant compte des biais de publication !
Ces données « non publiées » constituent ce que l'on appelle la *littérature « grise »* ou « grey literature ». Elles nécessitent des connaissances institutionnelles et professionnelles pour savoir chercher les données au « bon endroit ».
Sur le plan qualitatif, elles peuvent être significatives, en particulier dans le domaine de nouvelles technologies (études de « technology assessment »). Le développement de nouvelles technologies de dépistage, de diagnostic et de traitement, fait l'objet d'investigations, sous forme de rapports industriels ou de pré-rapports cliniques qui ne sont pas toujours disponibles dans les phases initiales de développement de ces technologies.

II. La stratégie de recherche documentaire

• **Il faut avant tout déterminer les mots-clés de la recherche** en utilisant des termes (ou descripteurs) faisant partie du thésaurus en anglais MeSH (disponible sur le MeSH de Browser de Pubmed, http://www.nlm.nih.gov/mesh/).
On peut utiliser pour cela l'interface « **HONselect** » de la fondation suisse *Health On the Net* (HON) qui permet une connexion rapide et fournit la traduction en français des termes du MeSH.
La demande doit être précisée en utilisant les *opérateurs booléens* :

AND (ET)
OR (OU)
NOT (SAUF)

L'utilisation de **Pubmed** nécessite un entraînement, et peut être déroutante au premier abord. L'URFIST de Paris fournit un bon mode d'emploi (http://www.hon.ch/HONselect/index_f.html), comportant une animation et des exercices corrigés.

• **La sélection des articles** se fait d'après les *titres*, puis d'après les *résumés*, que l'on peut lire sur l'ordinateur. Un résumé de bonne qualité correspond généralement à un article de bonne qualité, et permet au lecteur de juger si cet article lui apporte des informations utiles à son étude. La *lecture rapide de l'article* prend en considération l'objectif de l'article, les critères de qualité, les résultats rapportés. On imprime l'article s'il est considéré comme pertinent, la lecture approfondie se faisant sur papier.

III. L'utilisation des documents recueillis

On raye des listes de référence les articles rejetés – après avoir toutefois consulté les références qu'ils comportent. Les autres articles seront classés en deux groupes : ceux qui présentent un intérêt ponctuel, et ceux qui comportent de nombreuses données intéressantes.
A l'intérieur de chacun de ces deux groupes, on choisira un mode de classement, par exemple par nom du premier auteur (et année, si le même nom se retrouve plusieurs fois, ce qui est fréquent).

Pour le groupe « d'intérêt ponctuel », on notera la nature de ce point sur la liste, en regard de la référence.

• **Contrôle de l'exhaustivité de la recherche**

Elle se fait par une double démarche : rétrospective et prospective.

Contrôle rétrospectif : il consiste à consulter les références bibliographiques de tous les articles recueillis (y compris ceux que l'on a considérés comme à rejeter) pour s'assurer que l'on n'a pas laissé passer une référence « historique » intéressante. Ce travail, indispensable, est relativement aisé, et rapide. Après quoi, on peut définitivement rejeter les articles sans intérêt pour la question de recherche.

La démarche prospective consiste à repérer les publications nouvelles paraissant au cours même de l'étude. On fait appel pour cela aux *listes d'alertes*. Il existe de nombreuses possibilités : la plupart des périodiques proposent la réception gratuite par e-mail de la table des matières le jour même de la parution. L'accès aux articles peut être gratuit, comme pour certains articles du British Medical Journal (www.bmj.com), il peut se faire par un système de « pay-per-view » (consultation payante), par exemple pour le New England Journal of Medicine (www.nejm.com) ou de nombreuses autres revues, ou être réservé aux abonnés.

Il est facile pour de nombreux journaux d'obtenir les titres et les résumés des articles correspondant à des mots-clés sélectionnés, et particulièrement dans le cas des périodiques indexés dans des bases de données. Ce service est gratuit avec MyNCBI (http://www.ncbi.nlm.nih.gov/entrez/login.fcgi?call=so.signon.login). Il existe aussi des services gratuits, sponsorisés, dont certains fournissent également la conclusion des articles (www.mdlinx.com).

• **Détermination de la pertinence des documents sélectionnés**

Le travail de compilation de données est nécessaire mais pas suffisant, puisqu'il doit être soumis à une lecture « critique » au sens de « critical appraisal » des Anglo-Saxons.

C'est la partie la plus délicate de la documentation. Elle doit répondre à deux questions :

- l'article, et les informations qu'il contient, est-il pertinent dans le cadre de ma question de recherche ?
- l'article est-il fiable ? La réponse à cette question est fournie par l'analyse critique de la littérature médicale. En d'autres termes, les résultats fournis sont-ils de bonne qualité pour que je les utilise pour construire ma question de recherche et que mon protocole de recherche en tire bénéfice ?

B. EVALUATION DE LA PERTINENCE DE LA LITTERATURE

C'est la tâche du futur investigateur, qui désire préciser sa question de recherche clinique ou trouver la justification de son projet de protocole. C'est aussi ce que fait le clinicien, à la recherche, à travers la lecture régulière de la littérature médicale, d'une aide à la décision pour sa pratique quotidienne.

La lecture de la littérature nécessite une sélection et une évaluation. Pour cela s'est développé le concept de "**lecture critique**". Le principe en est de juger la valeur des publications, que ce soit la qualité de la recherche entreprise ou la pertinence des résultats publiés:

- Quelle est la crédibilité de la publication (validité interne) ?

Les résultats rapportés par l'auteur correspondent-ils vraiment à la réalité ? Peut-on avoir confiance dans les conclusions proposées par l'auteur ? Pour apprécier la validité de l'étude. Le lecteur doit être capable d'identifier rapidement, en fonction du type de l'étude, les différentes étapes du protocole qui l'a sous-tendue et leurs composants, en les évaluant de façon à définir le niveau de crédibilité des informations fournies.

- Quelle est l'applicabilité des informations contenues dans la publication (validité externe) ?

Si les conclusions sont considérées comme valides, sont-elles applicables à la pratique médicale du lecteur ou au projet de recherche du futur investigateur ?

Description générale de la méthode

Cette méthode de lecture repose sur l'utilisation d'une grille d'évaluation unique et standardisée. Le plan d'analyse proposé a l'avantage d'être applicable à tous les types de publications. Il s'agit d'une adaptation du "Critical Appraisal Worksheet" du Centre for Clinical Epidemiology and Biostatistics (Pr R. F. Heller) de l'Université de Newcastle (New South Wales, Australia).
Cette grille est constituée de 8 lignes, correspondant à 8 critères d'évaluation.
Chacun de ces 8 critères appelle les 3 mêmes types de questions, ce qui correspond aux 3 colonnes de la grille :
1- Est-il possible de trouver dans l'article l'information pour le critère en question?
2- La façon dont le critère en question a été abordé est-elle correcte ?
3- Si la façon d'aborder le critère en question est incorrecte, cela menace-t-il la validité de l'étude ?

Ces 8 lignes correspondent aux principales étapes de la conception d'un protocole. Toute publication est le résultat d'une étude définie par un protocole. L'analyse du protocole permet de le valider, à l'aide de la grille. En répondant aux huit questions, le lecteur a la capacité d'écarter très vite ce qui n'est pas valide. Il peut ainsi porter un regard objectif sur la qualité des résultats qui lui sont proposés.

Les huit étapes

1. Quel est l'objectif ?
Le médecin est à la recherche d'informations scientifiques concernant ses quatre grandes préoccupations : la thérapeutique, le pronostic, l'étiologie et le diagnostic. (Ce sont ces quatre catégories qui sont proposées à la rubrique des questions cliniques de la base de données en ligne NCBI PubMed. www.pubmed.org)

L'impact d'une intervention.
L'intervention médicale est le plus souvent thérapeutique, médicamenteuse ou non, mais peut aussi être de diagnostic, de dépistage ou d'éducation. L'objectif est de distinguer l'intervention utile de celle qui est inutile voire dangereuse. Les questions dont le lecteur cherche la réponse dans la littérature sont toujours les mêmes : est-on sûr de faire plus de bien que de mal ? A efficacité égale, peut-on faire moins cher ? A coût égal, peut-on faire plus efficace ?

Le risque de développer une maladie, son évolution et son pronostic.

Risque et pronostic se trouvent sur le même continuum de l'histoire de la maladie. Le facteur de risque est associé à l'acquisition de la maladie, le facteur « pronostic » au déroulement de la maladie une fois acquise. Le pronostic constitue une information capitale car il conduit à la notion de risque de base qui permet de quantifier l'effet d'une intervention et d'évaluer son intérêt.

La détermination d'une causalité (ou étiologie)
Pour le clinicien, la connaissance de l'étiologie ou de la causalité est fondamentale pour sa pratique médicale, que ce soit la prévention, le diagnostic ou le traitement. La causalité concerne l'association entre un facteur de risque et une maladie, et la force de cette association.

La validité et l'utilisation d'un nouveau test diagnostique.
Si un standard de référence existe, l'article concerne les qualités intrinsèques et la performance du test.
Si le standard de référence n'existe pas, l'intérêt de l'étude du test réside dans ses conséquences cliniques pour le patient. Il s'agit de savoir si le patient se porte mieux avec cette procédure diagnostique que sans elle. On rejoint le type d'article concernant l'évaluation d'une intervention médicale au sens général du terme.

2. Quel est le plan d'étude ?

En dehors du rapport d'un cas intéressant et inhabituel, ou de celui d'une série de cas, il existe quatre grandes sortes de plan d'étude:

- Etude transversale: description de la fréquence d'une maladie, de ses facteurs de risque ou de ses autres caractéristiques dans une population déterminée à un temps déterminé.

- Etude cas-témoins: étude d'observation, rétrospective, dans laquelle les caractéristiques de patients atteints d'une maladie (les cas) sont comparées avec celles de patients indemnes de la maladie (les témoins).

- Etude de cohorte: étude d'observation, prospective, dans laquelle un groupe de sujets exposés à des facteurs de risque d'une maladie est suivi pendant une période de temps donnée. Le taux d'incidence de la maladie dans ce groupe exposé est comparé à celui d'un groupe témoin, suivi pendant le même temps, mais non exposé aux facteurs de risque.

- Essai contrôlé: étude expérimentale dans laquelle une intervention est pratiquée dans un groupe de sujets; le résultat de cette intervention est comparé à celui dans un groupe semblable, témoin, qui ne reçoit pas l'intervention.

Le lecteur doit reconnaître le plan d'étude pour vérifier si celui-ci est le plus approprié pour la question posée. En outre, il existe une "hiérarchie" parmi ces modèles et le niveau de preuve des résultats d'une étude (et la confiance du lecteur) est variable d'un modèle à l'autre. Il va croissant, du cas ou de la série de cas à l'étude transversale, puis à l'étude cas-témoins, à l'étude de cohorte, pour être maximal avec l'essai contrôlé.

3. Quel est le facteur étudié ?

Le facteur étudié est l'exposition ou l'intervention supposée avoir des conséquences sur un problème de santé, une maladie ou un état clinique.

Le lecteur doit pouvoir savoir comment le (ou les) facteur a été mesuré, si tous les facteurs pertinents ont été pris en compte, et si la même méthode de mesure a été appliquée à tous les sujets, ainsi que d'un groupe à l'autre. Il doit aussi pouvoir apprécier la qualité de cette mesure (variabilité, mesure en "insu", …).

Si le facteur étudié est un test diagnostique, y-a-t-il une comparaison indépendante avec l'étalon ?

4. Quel est le critère de jugement ?

Le critère de jugement est l'événement ou la situation supposée être le résultat de l'influence du facteur étudié (mort, maladie, inconfort, insatisfaction, …). Le lecteur doit trouver les mêmes renseignements que pour le facteur étudié (définition précise, méthode de mesure, …).

5. Quelle est la population étudiée et de quel échantillon s'agit-il ?

La population de référence, ou population à étudier, est le groupe auquel les résultats de l'étude, s'ils sont valides, vont s'appliquer. L'échantillon est un sous-groupe de la population étudiée, sélectionné, de façon aléatoire ou non, pour représenter l'ensemble de la population étudiée, lorsqu'il n'est pas possible pour des raisons pratiques d'étudier celle-ci dans sa globalité.

La sélection est-elle correcte ? Y-a-t-il randomisation ? Les groupes diffèrent-ils par des caractéristiques autres que les facteurs étudiés ? Quelle est la proportion de sujets atteignant la fin du suivi ? A-t-on, si le facteur étudié est un test diagnostique, pris en compte un large éventail de patients ?

Il s'agit aussi, à cette étape, de juger la validité externe de l'étude: les conclusions, en admettant qu'elles soient valides, peuvent-elles s'appliquer à une population plus large que le simple échantillon étudié ?

6. Y-a-t-il des biais et des facteurs de confusion ?

Un biais est une erreur systématique qui contribue à produire des estimations systématiquement plus élevées ou plus basses que la valeur réelle des paramètres à estimer. Il intervient par exemple au niveau de la sélection des patients, ou de la mesure des paramètres étudiés.

Un facteur de confusion est un facteur qui modifie les effets du facteur étudié sur le critère de jugement, du fait de son lien à la fois avec le facteur étudié et avec le critère de jugement.

Sont-ils tous envisagés et pris en compte ? Si tel n'est pas le cas, la validité de l'étude peut être mise en doute.

7. Quels sont les résultats ?

La confiance dans le résultat : l'intervalle de confiance

La recherche clinique est réalisée la plupart du temps sur des échantillons, pour des raisons évidentes de faisabilité. Les résultats publiés sont les valeurs observées dans l'échantillon. La vraie valeur (de l'effet du traitement, par exemple), qui correspond à la vérité dans la population, se situe quelque part au voisinage de cette valeur observée. Le rôle des statistiques est de définir par calcul un intervalle de valeurs où se trouve 95 fois sur 100 la vraie valeur,

permettant ainsi le passage de l'échantillon à la population générale. Cet ensemble de valeurs constitue l'intervalle de confiance à 95 %.
Le lecteur doit savoir que la vérité se trouve entre les deux bornes de l'intervalle. Plus cet intervalle est petit, mieux la vérité est cernée, car plus proche de la vérité se trouve la valeur observée. Et cet intervalle est d'autant plus petit que l'échantillon étudié est grand et que le nombre d'événements étudiés est grand.

La valeur de petit p
Le lecteur doit démystifier la fameuse valeur de p. La valeur de p est la probabilité, calculée par le test statistique construit à partir des données collectées, d'obtenir par pur hasard une différence plus grande ou égale à celle qui est observée. Le risque consenti d'affirmer qu'il y a une différence entre les deux groupes, quand en réalité il n'y en a pas, est dénommé seuil de signification. Un seuil de signification de 5 % signifie qu'on a décidé de prendre un risque de 5 chances sur 100 de se tromper en affirmant qu'il y a une différence. Les seuils de signification habituellement choisis sont 5 % et 1 %, ce qui veut dire qu'on a 5 chances ou 1 chance sur 100 seulement que la différence observée ne soit due qu'au seul hasard et non au facteur étudié.

La significativité statistique ou clinique ?
On peut se laisser abuser par la magie de ce petit p, et croire, par exemple, qu'un $p < 00001$, c'est mieux qu'un $p < 0,05$, et que c'est l'argument définitif pour valider et accepter les résultats.
Une différence statistiquement significative n'est pas nécessairement pertinente cliniquement. C'est le problème de la signification clinique et de la signification statistique. De petites différences peuvent être statistiquement significatives si elles sont observées sur de grands échantillons, mais peuvent n'avoir que peu d'importance sur le plan clinique. Si une association très forte existe entre un facteur étudié et le critère de jugement choisi, il suffit d'un petit échantillon pour la démontrer. Au contraire, si cette association existe mais est de faible amplitude (augmentation de 10 % de la survie à 10 ans, par exemple), il faut alors un très grand échantillon pour le démontrer.
Cette approche doit permettre au lecteur de garder son sang-froid quand « c'est statistiquement significatif », et de ne pas perdre tout espoir quand « ce n'est pas statistiquement significatif ».

La puissance
Une différence statistiquement significative n'est intéressante que quand elle est cliniquement pertinente.
Lorsque la différence est non significative, le lecteur dira que le résultat est négatif. Mais est-ce bien un « vrai négatif » ? Ce peut être en effet un « faux négatif ». La différence ou l'effet recherché existe peut-être bien, mais la taille de l'échantillon était insuffisante pour le mettre en évidence. Ou bien il existe réellement une différence ou un effet, mais moins important que ne le voulait l'hypothèse, et là encore la taille de l'échantillon était insuffisante pour la mettre en évidence (on parle de manque de puissance de l'étude). L'erreur de type II est d'affirmer qu'il n'y a pas de différence ou d'effet alors qu'en réalité il y en a. C'est l'erreur β, et la puissance de l'étude est $1 - \beta$. On peut faire le parallèle entre cette situation et celle des faux négatifs des tests diagnostiques.

8. Synthèse de la lecture critique

A chacune des étapes précédentes, la validité de l'étude était-elle menacée, gravement ou faiblement (validité interne) ? Quelles sont les conclusions des auteurs ? Répondent-ils aux questions ? Les résultats sont-ils applicables à la population étudiée (validité externe) ?

Surtout, les résultats sont-ils acceptables pour la propre pratique du lecteur ? Vont-ils changer son comportement et améliorer l'état de ses patients ?

Cette approche exige du lecteur la connaissance des notions de base en méthodologie, il s'agit cependant d'une démarche essentiellement clinique. A travers une lecture critique, le lecteur démonte et évalue le protocole que les auteurs ont conçu pour réaliser l'étude présentée dans la publication.

Conclusion

La critique ne doit être ni systématique ni paranoïaque. Si la recherche clinique repose sur des règles rigoureuses, elle a pour but d'analyser et de quantifier des phénomènes biologiques et humains dont l'investigation peut trouver ses limites pour des questions de méthodologie (suivi nécessaire trop long de cohortes trop vastes, effet escompté trop faible…), de budget ou d'éthique. La perfection n'est pas de ce monde.

L'utilisation de cette technique d'analyse de la littérature est une démarche essentiellement pragmatique qui donne au lecteur les moyens de croire, ou de ne pas croire, et d'appliquer, ou de ne pas appliquer, ce qui pourrait lui être utile. Cette approche nécessite une connaissance des notions de base de méthodologie. Elle a le mérite de s'adapter aux principaux types de publications. Elle permet d'apprécier le niveau de preuves apportées par les auteurs. S'il n'y a pas de preuve absolue, il y a certainement pour un même sujet des articles plus convaincants que d'autres. Au lecteur de chercher les meilleurs arguments avec l'outil qui lui est proposé !

Références

Americain Medical Association. Guyatt GH, Rennie D. Users' Guides to the Medical Literature. AMA, 2002.

ANAES Guide d'analyse de la littérature et gradation des recommandations. ANAES, Paris, 2000, www.anaes.fr

Bazi R. La recherche documentaire. *In* : Matillon Y, Maisonneuve H, éd. *L'évaluation Médicale. Du concept à la pratique.* 3e éd. Paris, Flammarion Médecine-Sciences, 2007, p.143-157.

Glanville J, Wilson P, Richardson R. Accessing the online evidence: a guide to key sources of research information on clinical and cost effectiveness. Qual Saf Health Care, 2003, *12*: 229-231.

Greenhaglh T. Savoir lire un article médical pour décider. La médecine fondée sur les niveaux de preuve (evidence-based medicine) au quotidien. Traduit de l'anglais par Broclain D, Doubovetzky J. Editions RanD, Meudon, 2000.

Haynes RB, McKibbon KA, Fitzgerald D et al. How to keep up with the medical literature: I. Why try to keep up and how to get started. Ann Intern Med, 1986, *105*: 149-153.

Huguier M, Flahaut A. Biostatistiques au quotidien. Elsevier, Paris, 2000.

Junod AF. Décision médicale ou la quête de l'explicite. Editions Médecine et Hygiène, Genève, 2003.

Landrivon G. Méthode globale de lecture critique d'articles médicaux à l'usage de l'étudiant et du praticien. Frison Roche, Paris, 2002 et 2009.

Lorette G, Grenier B. La lecture d'articles médicaux, Doin, Paris, 2002.

Mouillet E. *La recherche bibliographique en médecine et santé publique. Guide d'accès*. Paris, Elsevier, 2005.

Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessments reports carried out for the National Institute for Clinical Excellence appraisal system. Health Technol Assess, 2003, *7*: 1-64.

Weightman AL, Williamson J. The value and impact of information provided through library services for patient care: a systematic review. Health Info Libr J, 2005, *22*: 4-25.

Annexe 1 : La grille de lecture

L'information existe-t-elle pour chacune des 8 questions?	**La façon d'aborder la question est-elle correcte ?**	**Si non, cela menace-t-il la validité de l'étude ?**
1 - Objectif - pronostic - évolution - test diagnostique - impact d'une intervention - étiologie - causalité	* Y a-t-il une hypothèse ?	
2 - Type d'étude - rapport de cas - série de cas - étude transversale - étude cas-témoin - étude de cohorte - essai contrôlé	* Le type de l'étude est-il approprié à la question posée ?	* Si non, les résultats de l'étude sont-ils totalement inutiles ?
3 - Facteur(s) étudié(s) - exposition - intervention - test diagnostique	Sont-ils bien décrits ? Comment sont-ils mesurés ? * Même méthode de mesure chez tous les sujets ? dans tous les groupes ? * Méthode à l'aveugle ? Y a-t-il une comparaison indépendante avec le standard de référence ?	* Si non ce biais de mesure menace-t-il la validité de l'étude ? * Idem Si non ce biais menace-t-il la validité de l'étude ?
4 - Critère(s) de jugement	Comment sont-ils mesurés ? * Même méthode de mesure chez tous les sujets ? dans tous les groupes ? * Méthode à l'aveugle ? Tous les critères de jugement pertinents sont-ils évalués ?	*Sinon ce biais de mesure menace-t-il la validité de l'étude ? *Si non, ceux qui ont été oubliés sont-ils importants ?
5 - Population source et sujets étudiés	* La sélection est-elle correcte ? * Y a-t-il randomisation ? * Les groupes différent-ils par des caractéristiques autres que les facteurs étudiés ? * Quelle est la proportion de sujets atteignant la fin du suivi ? * Y a-t-il, pour le test, un large éventail de patients ?	* Si non, ce biais menace-t-il la validité externe? * Si non, ce biais menace-t-il la validité interne ? * Si elle n'est pas optimale, la validité interne est-elle menacée? * Si non, ce biais menace-t-il la validité externe?
6 - Facteurs de confusion potentiels et **biais**	* Sont-ils tous envisagés ? * Sont-ils bien contrôlés ?	* Si non, cela invalide-t-il l'étude ?
7 - Analyses statistiques et résultats Intervalle de confiance ? Test statistique ? * si résultats positifs * si résultats négatifs Force de l'association Calcul rapports de vraisemblance	* Taille de l'échantillon suffisante ? * Cliniquement intéressant ? * Puissance du test, taille de l'échantillon ?	* Si non, les résultats sont-ils inutiles ? * Si non, l'étude est-elle utile ? * Si insuffisant, l'étude est-elle utile ou non concluante ?
8 - Conclusions des auteurs? Réponses aux questions ? Vérification de l'hypothèse ? Objectif atteint ?	* Les conclusions répondent-elles à l'objectif ?	<u>En somme :</u> * *Les résultats sont-ils acceptables appliqués à la population-source ?* = **VALIDITE** * *Les résultats sont-ils acceptables pour votre propre pratique ?* = **APPLICABILITE**

Annexe 2 : **Exemple d'utilisation de la grille de lecture critique à propos d'un article fictif « Consommation d'alcool et risque de cancer du sein »**

RESUME

Une étude cas-témoins a été réalisée pour déterminer si la consommation d'alcool augmente le risque de cancer du sein. On a interrogé 1594 femmes âgées de 22 à 56 ans, avec un diagnostic récent de cancer du sein, et 1663 femmes du même âge, sélectionnées au hasard, à partir de la population générale. Les femmes consommatrices d'alcool ne présentaient pas un risque supérieur de développer un cancer du sein par rapport aux femmes qui ne buvaient pas d'alcool: risque relatif: 1,0; intervalle de confiance à 95%: 0,8 à 1,2. Le risque de cancer du sein n'était associé ni à la quantité moyenne d'alcool consommée par semaine, ni au type de boissons alcoolisées consommées. En comparaison avec les femmes qui ne buvaient pas, les risques relatifs de développer un cancer du sein pour des femmes qui avaient bu de la bière, du vin ou des spiritueux étaient respectivement de 1,0, 0,8 et 0,9.

INTRODUCTION

L'étude de Hutchinson et Bergounian avait suggéré que les femmes qui consomment de l'alcool ont un risque de développer un cancer du sein 1,5 à 2 fois supérieur à celui des femmes qui n'ont jamais consommé d'alcool; cette augmentation de risque était associée à la consommation de tous les types d'alcool (bière, vin et spiritueux).

Le cancer du sein est une cause majeure de décès dans la plupart des pays industrialisés, et la consommation d'alcool est très commune chez les femmes dans ces pays. Si les résultats de Hutchinson et Bergounian — un risque 2 fois supérieur de développer un cancer du sein chez les femmes qui boivent de l'alcool — s'appliquent aux femmes américaines, dont 60% boivent de l'alcool et 7% développent un cancer du sein, alors on peut estimer qu'une proportion non négligeable des cancers du sein est à mettre sur le compte de la consommation d'alcool. Il est donc important de clarifier la relation entre consommation d'alcool et cancer du sein.

SUJETS ET METHODES

Les sujets participant à l'étude proviennent de 8 zones géographiques (les zones urbaines de Zorgrad, Zorgburg, Zorgcity et Zorgtown dans l'état du Zorgland, et les 4 comtés urbains du Zorgshire).

Un questionnaire standard pré-testé a été distribué aux femmes participant à l'étude, à domicile. Le questionnaire insistait sur les antécédents gynéco-obstétricaux et les antécédents contraceptifs, les antécédents familiaux, le passé médical, les caractéristiques personnelles et les habitudes, et collectait les informations concernant la quantité et la fréquence de la consommation de bière, de vin et de spiritueux pendant les 5 années passées.

Les critères d'inclusion pour les cas étaient les suivants:
- âge: 22 à 56 ans,
- cancer du sein primitif, histologiquement confirmé, diagnostiqué entre le 1er janvier 1991 et le 30 avril 1992,
- résidant dans l'une des 8 zones décrites précédemment.

De plus, les femmes devaient être disponibles pour les interrogatoires.

Nous avons ainsi inclus 1594 femmes (83,7% des femmes atteintes d'un cancer du sein qui satisfaisaient aux critères d'inclusion). Les raisons de non inclusion ont été la maladie (3,4%), le refus de la patiente (3,2%), le refus du médecin traitant (2,9%), et l'impossibilité de contacter ou de conduire un interrogatoire dans les 6 mois qui ont suivi la date du diagnostic (6,8%).

Les témoins ont été des femmes identifiées par la méthode de sélection téléphonique de Schprountz habitant dans les mêmes zones géographiques que les cas. Environ 94% des foyers ont le téléphone et les échantillonnages réalisés par appels téléphoniques au hasard sont représentatifs de la population. Une proportion appropriée de témoins par tranches d'âge de 5 ans a été sélectionnée pour être appariée avec les cas de cancer du sein en respectant la distribution de l'âge. Parmi les témoins sélectionnées et disponibles pour un interrogatoire, 1663 (84,9%) femmes acceptèrent de participer; 10,5% des témoins sélectionnées ont refusé de participer et 4,6% avaient changé de lieu de résidence ou n'ont pas pu être contactées.

On a demandé aux femmes si elles avaient eu l'occasion de boire une quelconque boisson alcoolique ou spécifiquement de la bière, du vin ou des spiritueux durant les 5 années précédentes. Les femmes qui ont répondu non ont été considérées comme non consommatrices d'alcool. Aux femmes qui ont répondu oui, on a demandé le nombre moyen de jours par semaine où elles ont bu de la bière, du vin ou des spiritueux, et la quantité qu'elles avaient l'habitude de boire ces jours là. Pour chaque femme, nous avons utilisé les données concernant la quantité et la fréquence de consommation pour estimer le nombre moyen de boissons qu'elles ont eu chaque semaine et nous avons multiplié ce nombre moyen par 12,6 (le poids en grammes de la quantité absolue d'éthanol par boisson) pour estimer l'ingestion hebdomadaire d'éthanol pur pour chaque femme.

Nous avons estimé le risque relatif par la méthode de Cornfield et son intervalle de confiance à 95% par le test de Miettinen.

Les variables suivantes ont été retenues comme facteurs de confusion potentiels parce qu'elles constituent des facteurs de risque classiques de cancer du sein ou bien parce qu'elles sont fortement liées à la consommation d'alcool:
- antécédents de maladie bénigne du sein,
- antécédents familiaux de cancer du sein,
- âge à la première grossesse menée à terme,
- statut ménopausal,
- niveau éducatif,
- âge lors du diagnostic de cancer du sein ou âge lors de l'interrogatoire,
- religion,
- nombre de cigarettes consommées,
- et index de Quetelet (poids/taille^2, mesure de l'adiposité).

Nous n'avons pas inclus l'utilisation des contraceptifs oraux parce qu'il a été prouvé récemment que ceux-ci ne constituaient pas un facteur de risque du cancer du sein.

Une régression logistique a été utilisée pour contrôler simultanément tous ces facteurs de confusion potentiels et pour calculer l'estimation du risque relatif pour l'association entre la consommation d'alcool et le risque de cancer du sein.

RESULTATS

La distribution des âges et des races était la même pour les femmes présentant un cancer du sein (les cas) et pour les témoins. Il y avait plus de nullipares chez les cas que chez les témoins, les cas de cancer du sein étaient plus âgées lors de la naissance de leur premier enfant et présentaient plus d'antécédents familiaux de cancer du sein ainsi que d'antécédents personnels de maladie bénigne du sein. Un pourcentage plus important de cas était en période pré-ménopausique alors qu'un plus grand pourcentage de témoins avait eu une ménopause chirurgicale.

Comparées à celles qui ne buvaient pas, les femmes qui buvaient des boissons alcoolisées avaient un risque relatif de développer un cancer du sein de 1,1 (intervalle de confiance à 95%: de 0,9 à 1,3) (tableau 1). On n'a pas mis en évidence d'influence de la consommation alcoolique moyenne hebdomadaire sur l'apparition du cancer du sein. Les femmes qui affirmaient boire l'équivalent de plus de 300 grammes d'alcool par semaine avaient un risque relatif ajusté de développer un cancer du sein de seulement 1,1 (intervalle de confiance à 95%: de 0,6 à 1,8).

Ni le type de boissons alcoolisées, ni la quantité consommée ne sont apparus comme un sur-risque de développer un cancer, même après ajustement sur la consommation d'autres types de boissons alcoolisées (tableau 2). Le risque relatif associé avec des antécédents de forte consommation de bière, de vin ou de spiritueux était respectivement de 0,8, 1,2 et 1,1.

Aucune association significative n'a été trouvée entre le risque de cancer du sein et la consommation d'alcool pour des femmes appartenant à différents groupes religieux ou dans différents groupes d'âge; cependant, en général, un risque plus bas était observé pour des femmes plus jeunes. On n'a pas observé de relation entre consommation d'alcool et risque de cancer du sein qu'il y ait eu ou non des antécédents personnels de maladie bénigne du sein ou des antécédents familiaux de cancer du sein.

DISCUSSION

Nos résultats concordent avec tous ceux qui n'ont pas pu confirmer l'augmentation du risque de cancer du sein associé avec la consommation d'alcool décrite par Hutchinson et Bergounian. On pourrait expliquer les résultats de Hutchinson et Bergounian par l'inclusion dans leur étude de sujets présentant d'autres troubles liés à la consommation d'alcool; des femmes présentant un cancer de l'ovaire et un cancer de l'endomètre ont constitué dans leur étude un groupe témoin. Actuellement, nous travaillons sur l'hypothèse que l'augmentation du risque de développer un cancer du sein observée par Hutchinson et Bergounian puisse être due à l'effet protecteur de l'alcool sur le cancer de l'endomètre plutôt qu'à son effet direct sur le développement du cancer du sein.

Dans cette étude, Hutchinson et Bergounian disposaient de données limitées pour étudier une relation dose-réponse entre consommation d'alcool et risque de cancer du sein. Il y avait des informations sur la fréquence mais pas sur la quantité d'alcool consommée. Dans notre étude, nous avions à la fois des informations sur la quantité et sur la fréquence de la consommation d'alcool et nous avons pu estimer l'ingestion hebdomadaire moyenne d'alcool.

Hutchinson et Bergounian ont mis en évidence une augmentation du risque de cancer du sein pour la bière, le vin et les spiritueux bien que ces estimations de risque soient fondées sur de petits nombres. Nous n'avons pas mis en évidence d'augmentation du risque de cancer du sein associé avec la consommation de chacun de ces types de boissons alcoolisées lorsque nous avons ajusté sur chacun des principaux facteurs de risque de cancer du sein ainsi que sur chacun des autres types de boissons alcoolisées. De plus, nous n'avons pas trouvé de relation

dose-réponse entre le risque de cancer du sein et la quantité de consommation des différents types de boissons alcoolisées. Il est peu probable que des biais soient intervenus dans nos résultats. Le biais de sélection était certainement très réduit du fait que les participantes à l'étude ont été incluses uniquement très précocement après la mise en évidence du diagnostic et dans les 8 zones géographiques, et du fait que les témoins ont été sélectionnées dans la population provenant de ces mêmes zones. Il est peu probable que de mauvaises descriptions de consommation de l'alcool par les participantes à l'étude expliquent le manque d'association entre la consommation d'alcool et le risque de cancer du sein parce que les cas comme les témoins ont rapporté des niveaux de consommation alcoolique légèrement supérieurs à ceux qui sont rapportés dans les études nationales. Si la période critique d'exposition pour le développement d'une tumeur du sein est supérieure à 5 ans avant que le diagnostic de cancer du sein ne soit posé, alors notre classification en statut de buveuses et de non buveuses sur la base d'une consommation durant les 5 années précédentes pourrait assimiler des patientes ayant consommé de l'alcool à des patientes non consommatrices d'alcool. Cette mauvaise classification pourrait cacher une véritable association entre consommation d'alcool et cancer du sein si la consommation d'alcool pendant cette période critique augmentait réellement le risque de développer un cancer du sein. Quoi qu'il en soit, l'amplitude de cette mauvaise classification n'est certainement pas supérieure à 5%.

Tableau 1 - Risque de cancer du sein selon la consommation d'alcool hebdomadaire moyenne

	Cas	Contrôles	Risque relatif (95%)
n'a jamais bu	286	300	1,0
a bu (g/semaine)	1308	1363	1,1 (0,9-1,3)
<50	722	759	0,9 (0,7-1,2)
50-149	342	377	0,9 (0,7-1,2)
150-199	93	87	1,1 (0,7-1,7)
200-249	56	52	1,1 (0,7-1,9)
250-299	40	37	1,0 (0,5-1,7)
≥ 300	55	51	1,1 (0,6-1,8)

Tableau 2 - Risque de cancer du sein par type de boissons alcoolisées consommées

Consommation moyenne (g/semaine)	Cas	Contrôles	Risque relatif (95%)
n'a jamais bu de bière	856	896	1,0
a bu de la bière	738	767	1,0 (0,9-1,2)

< 50	618	629	1,1 (0,9-1,3)
50-149	82	91	0,9 (0,6-1,3)
≥ 150	38	47	0,8 (0,4-1,3)
n'a jamais bu de vin	481	456	1,0
a bu du vin	1113	1207	0,8 (0,7-1,1)
< 50	841	959	0,8 (0,6-1,0)
50-149	188	184	0,9 (0,6-1,2)
≥ 150	84	64	1,2 (0,8-1,9)
n'a jamais bu de spiritueux	507	510	1,0
a bu des spiritueux	1087	1153	0,9 (0,7-1,2)
< 50	846	897	0,9 (0,7-1,2)
50-149	164	179	0,8 (0,6-1,2)
≥ 150	77	77	1,1 (0,7-1,7)

Lecture critique:

L'objectif de cette étude est de fournir des informations concernant l'étiologie - causalité.

L'hypothèse est celle de l'association entre la consommation d'alcool et le développement du cancer du sein. Attention, "association" ne veut pas dire "relation de cause à effet".

Le type d'étude est une étude cas - témoins. Deux groupes de femmes ont été constitués:
- un groupe de cas de cancer du sein: 1594 femmes;
- un groupe de femmes témoins, indemnes de cancer du sein: 1663 femmes.

Chez les cas comme chez les témoins, les investigateurs sont remontés dans leur passé pour rechercher et mesurer la consommation d'alcool, et la comparer entre les deux groupes.

Ce type d'étude est bien approprié à la question posée.

Seul le modèle rétrospectif est concevable. On imagine la difficulté de conception d'une étude prospective sur ce sujet: partir d'un groupe de femmes "alcooliques" et les suivre dans le futur de nombreuses années, pour collecter les cas incidents de cancer du sein qui seraient comparés à ceux qui surviennent chez des femmes suivies en parallèle, mais "non alcooliques".

L'essai contrôlé est bien entendu impensable.

Quant aux séries de cas et études transversales, elles seraient nécessairement non conclusives du fait de l'absence de groupe témoin.

Le facteur étudié (l'exposition ou l'intervention qui est supposée avoir des conséquences sur un problème de santé, une maladie ou un état clinique) est la consommation d'alcool au cours des 5 années précédentes: alcool, vin, bière, spiritueux.

Celle-ci est mesurée par questionnaire standard pré-testé distribué à domicile. On a apparemment évité le risque de poser les questions de façon différente selon que l'on a affaire à un cas de cancer du sein ou une femme témoin. Cette situation pourrait en effet conduire à une surestimation de la consommation d'alcool chez les cas, et donc à une surestimation de l'association entre alcool et cancer du sein.

Le problème se situe au niveau de la quantité d'alcool ingérée, en absolu comme en fonction des différents types de boisson, dans les deux groupes. La méthode de mesure n'est pas assez précise. Il faut définir ce qu'est une unité de vin, de bière, de spiritueux, et connaître le degré d'alcool de chaque boisson.

Le critère de jugement (l'événement ou la situation supposés être le résultat de l'influence du facteur étudié) est le cancer du sein. Est-ce le diagnostic de cancer du sein ou est-ce la mortalité par cancer du sein ?

Si l'anatomo-pathologie définit de façon irréfutable un cas, comment peut-on être sûr qu'un témoin, dans cette étude, est indemne de cancer du sein ? La présence de femmes en réalité atteintes de cancer du sein dans le groupe témoin conduirait à une sous-estimation de l'association entre alcool et cancer du sein.

La population

Dans l'exemple, la population de référence est celle des femmes qui consomment de l'alcool.

Dans la population étudiée, les témoins sont identifiées par téléphone. Si des femmes ne sont pas accessibles par ce moyen, ou si des femmes refusent de participer à l'étude, est-ce pour des raisons socio-économiques ou psychologiques, qui pourraient expliquer aussi une consommation alcoolique supérieure à la normale ? Dans ce cas, cela ne risque-t-il pas d'augmenter artificiellement le taux des "alcooliques" chez les cas par rapport aux témoins, et d'aller dans le sens d'une fausse association entre alcool et cancer ? Le même phénomène peut se produire si les cas sont recrutés dans un hôpital drainant une population d'un niveau socio-économique particulier, et différente de celle des témoins. Les 1663 femmes témoins sont celles qui ont accepté de répondre. Elles représentent 84,9% des femmes identifiées. Les 15,1% qui n'ont pas répondu sont-elles différentes des autres ?

Même question pour les cas: 16,3 % des cas de cancer du sein identifiés n'ont pas pu participer à l'étude. Sont-elles systématiquement exposées de façon différente au facteur de risque ?

Les facteurs de confusion et les biais

Interroger sur leur consommation d'alcool les femmes atteintes de cancer avec plus de soin que les témoins constituerait un biais de mesure. Cela pourrait tendre à mettre en évidence une différence entre les deux groupes alors qu'elle n'existe pas. C'est ce qui se produirait

également si les femmes ininterrogeables du groupe témoin étaient systématiquement plus alcooliques. Ce serait un biais de sélection.

De même, il faut prendre en compte dans une telle étude, et c'est ce qui a été fait, tous les autres facteurs connus comme étant des facteurs de risque de cancer du sein (âge, statut ménopausal, …). En effet, s'il sont significativement plus fréquents dans le groupe des cancers, par exemple, on ne peut pas savoir si une éventuelle différence en taux de cancer entre les deux groupes est à mettre sur le compte de ces facteurs de risque ou sur le compte de la consommation alcoolique elle-même. Ces facteurs sont des facteurs de confusion.

Les analyses statistiques ont consisté en une estimation du risque relatif et de son intervalle de confiance. Les intervalles contiennent 1. Il n'y a donc pas d'association. Ces intervalles sont par ailleurs très réduits autour de 1. Cela veut dire que l'on est assez sûr que ce résultat négatif est un vrai négatif. Mais s'il s'agissait d'un faux négatif, on manquerait seulement une association extrêmement faible, un sur-risque de 1,2 ou une protection de 0,9, qui ne seraient pas forcément pertinents cliniquement.

En conclusion, dans l'exemple d'une étude cas-témoins sur la consommation alcoolique et le cancer du sein, c'est à la troisième étape que la validité interne de l'étude semble la plus menacée, du fait de la difficulté à mesurer le facteur étudié dans ce cas particulier. En ce qui concerne la validité externe d'une telle étude, on a vu que celle-ci était menacée par la proportion et la nature des femmes inaccessibles à l'étude.

Enfin, une étude de ce type réalisée sur une population urbaine pauvre d'Amérique du Nord, ou bien chez des Scandinaves, est-elle pertinente pour l'ensemble de la population française ?

Annexe 3 : Quelques sites

❖ **Moteurs de recherche de la littérature scientifique, bibliothèques en ligne :**

http://www.ncbi.nlm.nih.gov/pubmed/: correspond au site du moteur de recherche PubMed, explorant la base de données Medline, OldMedline, et quelques revues indexées de façon plus ancienne.
http://www.nlm.nih.gov/mesh/: correspond à la base de données PubMed des mots-clés en anglais.
http://www.crd.york.ac.uk/crdweb/: correspond à la bibliothèque électronique de l'université de York (Grande-Bretagne), avec notamment l'accès aux résumés de la base Cochrane et à des articles d'économie de la santé.
http://thomsonreuters.com/products_services/science/?view=Standard (anciennement isinet.com) : site complexe couvrant de larges domaines de la publication scientifique, dont le journal des facteurs d'impact et les Current Contents

❖ **Sites spécifiques (sites thématiques, maisons d'édition, exemples de revues)**
www.thecochranelibrary.com: site de la Collaboration Cochrane
www.bdsp.tm.fr: site de la Banque de Données en Santé Publique, organisme public indépendant de l'industrie pharmaceutique.
www.sciencedirect.com: regroupe toutes les publications du groupe Elsevier, soit environ 2500 périodiques.
www.bmj.com : site du British Medical Journal
www.nejm.com: site du New England Journal of Medicine

❖ **Sites institutionnels :**
http://www.hon.ch/HONselect/index_f.html : site de l'organisation non gouvernementale HON, reconnue par le Conseil Economique et Social des Nations Unies. Son but est de promouvoir l'accès à l'information médicale au sens large du terme (politique de santé comprise).
www.has-sante.fr: site de la Haute Autorité de Santé française
http://www.ahrq.gov/ : site de 'Agency for Healthcare research and quality', dépendant du ministère américain de la santé.

❖ **Sites en relation avec l'industrie pharmaceutique :**
http://infodoc.inserm.fr/codecs/codecs.nsf: site du collège des économistes de la santé parrainé par l'INSERM (site français), largement financé par l'industrie pharmaceutique.
www.mdlinx.com: site sélectionnant et résumant des articles publiés dans des revues à comité de lecture, avec présentation des résumés des articles ainsi sélectionnés. Son directeur est un dirigeant de la principale firme japonaise de marketing de spécialités pharmaceutiques.

CHAPITRE III

ETHIQUE DE LA RECHERCHE IMPLIQUANT DES ETRES HUMAINS

Marie-Charlotte Bouësseau, Gilles Landrivon

Chercheurs et promoteurs doivent avoir présentes à l'esprit les questions éthiques que toute recherche impliquant des êtres humains est susceptible de soulever. Il est donc de leur responsabilité d'anticiper ces questions avant même la rédaction du protocole. Un comité d'éthique de la recherche (CER) indépendant évaluera le protocole avant son autorisation et sa mise en œuvre. Il peut solliciter des amendements au protocole initial. Il sera également tenu informé du déroulement de l'étude par l'investigateur principal. Ce chapitre présente les principes fondamentaux de l'éthique auxquels une recherche impliquant des êtres humains doit se référer. L'objectif n'est pas d'entrer dans le détail de la constitution et du fonctionnement d'un comité d'éthique, ni de commenter la réglementation sur la recherche clinique actuellement en vigueur mais d'évoquer les questions que tout chercheur devra se poser avant de conduire une recherche.

Plan du chapitre

APPLICATION A LA RECHERCHE DES PRINCIPES ETHIQUES FONDAMENTAUX

A - Le principe de bienfaisance consiste à maximiser les bénéfices de la recherche

B - Le principe hippocratique de non malfaisance (*primum non nocere*) consiste à minimiser les risques inhérents à toute recherche

C - Le principe du respect de la dignité des personnes qui acceptent de participer à la recherche s'applique en particulier au respect de leur autonomie de décision dont découle le processus de consentement libre et éclairé et au respect de la vie privée dont découle la confidentialité des données personnelles.

D - Le principe de justice dont l'application à la recherche implique de:

- développer des priorités de recherche nationales répondant aux priorités de santé des populations
- favoriser le partage des bénéfices en particulier lorsqu'il s'agit de recherche internationale
- respecter la transparence et l'accès à des données scientifiques de qualités (registres, publications, etc.)
- assurer la participation de tous les acteurs concernés (associations de patients, etc.)

E - La recherche internationale

CONCLUSIONS ET PERSPECTIVES

APPLICATION A LA RECHERCHE DES PRINCIPES ETHIQUES FONDAMENTAUX

Toute recherche impliquant des êtres humains (sains ou malades), qu'il s'agisse d'essais cliniques, de recherches épidémiologiques, d'étude en psychologie ou en sciences sociales, peut soulever des questions éthiques. Celles-ci devront toujours être envisagées par le chercheur et le promoteur de la recherche dès la phase initiale d'élaboration du protocole. Il existe dans la plupart des pays des régulations (législations, lignes directrices, etc.) qui encadrent ces activités de recherche ; il existe également des normes internationales et des codes de déontologie professionnelle (voir annexes). L'investigateur devra en prendre connaissance et les respecter ; une formation à l'éthique de la recherche est donc souhaitable pour les professionnels qui participent à ce type de travaux. Le chercheur principal doit également se référer à un CER (Comité d'Ethique de la Recherche) pour une évaluation préalable du protocole de recherche, dans chacun des pays où celle-ci doit être mise en œuvre.

L'évaluation éthique des protocoles de recherche et les cadres normatifs internationaux et nationaux repose sur des principes éthiques fondamentaux, formulés au lendemain de la deuxième guerre mondiale dans le Code de Nuremberg puis la déclaration d'Helsinki ainsi que de nombreuses lignes directrices telles que celles des Conseils des Organisations Internationales des Sciences médicales ou de l'Organisation mondiale de la Santé (voir annexe).

A - Maximiser les bénéfices de la recherche

La recherche est justifiée par l'hypothèse d'un bénéfice pour la santé de la population ; il peut s'agir de nouvelles données scientifiques, nouvelles stratégies diagnostiques, nouveaux traitements, nouveaux vaccins, etc. Le protocole de recherche est conçu pour fournir des données scientifiquement valides et généralisables. Le bénéfice attendu doit être comparé avec les risques encourus par les sujets impliqués dans l'étude et pour la communauté à laquelle ils appartiennent (rapport risques/bénéfices).

Dans les cas où le bénéfice est essentiellement escompté pour la collectivité (ex. essais cliniques de phase I, voir chapitre 7), le risque pour les participants devra être d'autant plus faible que le bénéfice individuel est proche de zéro. Dans d'autres cas, on peut espérer un bénéfice pour les participants de l'étude, il sera comparé aux risques potentiels. Dans tous les cas, le bénéfice attendu de la recherche doit être clairement expliqué à tous les participants (voir processus de **consentement libre et éclairé**). Il arrive que le bénéfice d'un nouveau traitement apparaisse avant la fin de l'essai, celui-ci devra alors être arrêté pour

permettre à tous les participants y compris ceux du groupe contrôle de bénéficier de ce nouveau traitement.

B - Minimiser les risques inhérents à toute recherche

Les risques encourus sont de différents types:

- dommages physiques, dus aux complications d'un nouveau produit pharmaceutique ou d'une nouvelle technologie,

- dommages psychologiques, comme le stress, l'atteinte à la vie privée par rupture de la confidentialité et la discrimination sociale qui peut s'ensuivre.

Les risques doivent être distingués des contraintes et désagréments tels que le fait de subir un prélèvement sanguin ou de donner du temps pour répondre à un questionnaire. Les dommages soufferts par un participant sont sous la responsabilité du promoteur de la recherche qui devra les pallier. C'est le rôle des "*Data Safety Monitoring Boards*" (DSMB) (Conseils de contrôle des données et de la sécurité –voir annexes) de détecter au plus vite les effets indésirables de nouveaux médicaments sur la base des informations transmises par les investigateurs principaux, afin qu'ils soient corrigés dans les meilleurs délais. Dans certains cas, l'essai clinique sera interrompu pour des raisons de sécurité.

Les désagréments sont le plus souvent prévisibles et peuvent faire l'objet d'une indemnisation des participants (ex. prise en charge des frais de transport et indemnisation du temps passé). En aucun cas une indemnisation ne sera considérée comme un bénéfice de la recherche pour le participant.

Risques et désagréments font l'objet d'une information claire, transmise aux participants potentiels avant leur inclusion dans l'étude.

Il s'agit donc d'**optimiser le rapport risques/bénéfices pour les participants à la recherche et la population à laquelle ils appartiennent**. Une attention particulière sera portée aux personnes se trouvant en situation de vulnérabilité en raison de leur âge (personnes mineures ou âgées), leur condition de santé (personnes affectées de troubles mentaux ou de conscience), leur situation sociale (personnes privées de liberté, marginalisées, analphabètes ou en situation de grande pauvreté); plus généralement on devra assurer une protection particulière des personnes dont l'autonomie de décision est insuffisante pour assurer la validité du processus de consentement libre et éclairé.

L'évaluation éthique d'un protocole de recherche commence par la vérification de sa pertinence scientifique. Selon l'adage "ce qui n'est pas scientifique n'est pas éthique" on évitera d'exposer des participants aux risques d'une recherche dont la pertinence scientifique n'est pas avérée ; on évitera aussi de consacrer des ressources humaines et financières à des recherches futiles ou dont l'impact probable sur la santé est négligeable.

C - Respecter la dignité des personnes

- Respect de l'**autonomie de décision**: processus de consentement libre et éclairé

Le processus de consentement libre et éclairé doit être à la fois dynamique et interactif, il doit permettre une décision libre de toute coercition, reposant sur une information complète, transparente, compréhensible et adaptée à la personne à laquelle elle s'adresse. Ce processus et les moyens de support utilisés (document écrit, support visuel, etc.) doivent être décrits et évalués par le CER (voir annexe). La décision est authentifiée le plus souvent par un document écrit, signé par le participant; en tout état de cause le consentement devra être exprès (exprimé de manière formelle). Lorsque le participant à la recherche est une personne mineure, son assentiment sera recherché sur la base d'une information adaptée ; le consentement des parents ou du tuteur légal étant requis. En cas de personnes analphabètes ou illettrées, le recours à un tiers sera nécessaire. Bien entendu l'information et la décision de la personne de participer ou non seront exprimés dans une langue que celle-ci maîtrise.

Le processus de consentement libre et éclairé doit tenir compte du contexte socioculturel particulier dans lequel la recherche se déroule. Par exemple, dans certains cas, l'information et l'avis de la communauté devront précéder la décision de l'individu. Le temps nécessaire à ce processus est variable, il doit être suffisant pour permettre une bonne compréhension:

- des risques et bénéfices potentiels,
- des modalités de prise en charge d'éventuels dommages dus à l'étude,
- de la durée, des modalités de déroulement, des contraintes et éventuelles indemnisations,
- des objectifs de la recherche et des alternatives de traitements,
- de la méthodologie (randomisation, double aveugle, utilisation de placebo, etc.),

- de la possibilité pour le participant de modifier sa décision et retirer son consentement à n'importe quel moment sans préjudice pour lui,
- des mesures garantissant la confidentialité des données personnelles,
- de la possibilité de recherches ultérieures, par exemple de l'utilisation d'échantillons de prélèvements et la constitution de biobanques,
- et des sources de financement de la recherche.

- Respect de la **vie privée**: confidentialité des données personnelles

La confidentialité des données personnelles est un principe de l'éthique médicale déjà explicité par le serment d'Hippocrate. Les données personnelles collectées lors d'activités de recherche ne seront partagées qu'avec des personnes ayant connaissance des mesures de confidentialité qui doivent leur être appliquées. Différentes mesures peuvent être envisagées : anonymisation, codification, accès limité ; l'identification des sujets lors de la publication des résultats de l'étude ne doit pas être possible. La destruction des données après la fin de l'étude peut être exigée dans certains cas. Pour maximiser les bénéfices de la recherche, il sera parfois nécessaire de faire le lien entre des données recueillies et la personne concernée. C'est le cas, par exemple, des résultats de tests VIH qui devront être communiqués aux personnes pour qu'elles puissent profiter d'une prise en charge adaptée. Ces mesures seront évaluées par le CER et expliquées aux participants potentiels de l'étude.

Une brèche de confidentialité peut parfois entraîner des conséquences graves pour la personne, par exemple dans certaines affections psychiatriques, certaines maladies infectieuses particulièrement discriminantes, dans certains contextes socioculturels, telles que le VIH SIDA ou la tuberculose ; les données concernant le comportement sexuel ou des activités illégales seront recueillies avec beaucoup de prudence. On veillera à ne divulguer aucune information concernant la santé d'une personne à une compagnie d'assurance, une autorité judiciaire, un employeur ou même un proche, l'autorisation préalable de la personne concernée étant nécessaire au partage de cette information, par exemple avec un proche.

D - Application à la recherche du principe de justice

1. Développer des **priorités de recherche** nationales répondant aux priorités de santé des populations

Depuis l'année 2000, l'attention de la communauté internationale a été attirée sur la nécessité d'améliorer la cohérence entre priorités de recherche et priorités de santé, en particulier dans les pays émergents. La recherche internationale sera plus "juste" si elle permet de mieux répondre aux questions que pose la prise en charge des maladies affectant les populations les plus vulnérables.

2. Favoriser le **partage des bénéfices** de la recherche internationale

Le nombre d'essais cliniques réalisés dans les pays émergents suit une courbe exponentielle. En revanche, l'impact des bénéfices de ces recherches en termes d'accès aux nouvelles connaissances, aux nouvelles stratégies diagnostiques ou de prévention et aux nouveaux traitements reste insuffisant et la distribution de ces bénéfices est trop souvent inéquitable, favorisant les pays promoteurs de la recherche au détriment des pays hôtes. Les promoteurs et chercheurs ont la responsabilité de chercher à maximiser les bénéfices de leurs recherches pour les populations dans lesquelles elles sont conduites.

3. Respecter la **transparence** et l'accès à des données scientifiques de qualité

Cet aspect de l'éthique de la recherche est à l'origine de diverses initiatives comme par exemple le développement des registres nationaux d'essais cliniques, soutenu par l'OMS[①]. Les publications scientifiques ont un rôle important dans la diffusion des résultats de la recherche. Cette diffusion requiert que les recherches se soient déroulées dans le respect des meilleurs standards éthiques internationaux. Si la recherche est justifiée et les résultats attendus pertinents, ils devront être publiés et rendus accessibles à la communauté scientifique et même au public. Il est à noter que les résultats négatifs sont rarement publiés alors qu'ils peuvent avoir un intérêt scientifique important ; ce type de publication doit être encouragé par souci de transparence.

4. Assurer la **participation des tous les acteurs** concernés

Les associations de patients et de nombreuses organisations non gouvernementales contribuent à une meilleure information des citoyens et particulièrement des personnes les plus concernées par la recherche. Les chercheurs et promoteurs ont donc de plus en plus d'interactions avec la société civile, contribuant ainsi à maintenir la confiance du public.

[①] voir ICTRP http://www.who.int/ictrp/en/index.html

E- La recherche internationale

La mise en œuvre de protocoles de recherche dans des pays émergents et donc dans des contextes socioéconomiques très divers pose des questions éthiques complexes. Ainsi le processus de consentement libre et éclairé préalable à la mise en œuvre d'un même protocole de recherche dans des pays différents pourra suivre des modalités différentes en fonction de la culture et du contexte social de chaque pays. Il peut surgir alors une tension entre les principes éthiques à visée universelle et des valeurs culturelles particulières. Les CER locaux doivent prendre en considération cette tension et suggérer des modalités pratiques qui n'aillent pas à l'encontre des principes universels.

La question du partage des bénéfices de la recherche évoquée précédemment se pose avec plus d'acuité dans le cadre de la recherche internationale. Le débat international s'est focalisé sur cette question depuis une décennie, permettant des avancées non seulement en matière de régulations internationales mais aussi pour la mise en place de modalités d'accord préalable entre les différents acteurs de la recherche visant à augmenter les bénéfices de celle-ci et leur pérennité pour les populations concernées.

De nombreuses voix se sont élevées pour dénoncer un double standard éthique, suscitant une vive polémique notamment autour de l'utilisation du placebo. De fait, certaines populations se trouvent en situation de vulnérabilité. Pour cela de nombreuses initiatives visent à renforcer les CER locaux et à harmoniser les régulations nationales ; l'Organisation mondiale de la Santé, en collaboration avec d'autres organisations internationales gouvernementales ou non, travaille dans de nombreux pays en développement à la mise en place de systèmes d'évaluation éthique de la recherche assurant la promotion des droits fondamentaux des personnes.

CONCLUSIONS ET PERSPECTIVES

Le questionnement éthique doit être proactif, visant à améliorer la qualité et l'impact de la recherche sur la santé des populations. Une des difficultés vient de la diversité des acteurs concernés : chercheurs, promoteurs, participants, décideurs politiques, institutions de recherche et instances de régulation. La diversité des contextes dans lesquels la recherche se déroule rend celle-ci encore plus complexe. Il est donc difficile de trouver des points de

consensus sur des questions telles que les conditions d'utilisation du placebo ou le partage équitable des bénéfices de la recherche. Cependant, une meilleure approche de ces questions est facilitée par la formation des acteurs de la recherche. La consolidation des CER, leur interaction avec les équipes de chercheurs et les autorités de régulation de la recherche devraient permettre d'améliorer la qualité de l'évaluation et de suivi éthique des protocoles. La mise en place de réseaux internationaux contribue à l'harmonisation des normes éthiques et des méthodes de travail des CER, des propositions sont faites dans un nombre croissant de pays pour mettre en place un mécanisme d'agrément des comités d'éthique. Les Organisations internationales de la famille onusienne telles que l'Organisation mondiale de la Santé ou l'UNESCO, des organisations régionales telles que le Conseil de l'Europe travaillent de concert pour soutenir les pays dans cette tâche. Tous ces efforts seraient vains sans la compétence et l'intégrité des équipes de recherche. Une formation à la recherche scientifique et en particulier à l'éthique de la recherche doit être proposée à tous les chercheurs pour garantir la rigueur scientifique et la qualité éthique de la recherche en santé.

Annexes

1 - Les textes de référence en France

Lorsqu'un projet de recherche comporte des essais ou expérimentations pratiqués sur l'être humain (médicament, matériel, instrument, technique), la loi définit les conditions dans lesquelles ces projets de recherche sont autorisés, pour garantir la qualité et la sécurité de leur déroulement, et notamment pour protéger les personnes qui se prêtent aux essais.

La loi de référence est la loi n° 88-1138 du 20 décembre 1988, dite loi Huriet (révisée en 2009)

Pour les autres lois, décrets et circulaires, on peut se référer au site http://www.legifrance.gouv.fr *Protection des personnes se prêtant à des recherches biomédicales et dispositions connexes.*

Certaines institutions ont également adopté des chartes, par exemple : ANRS, Institut Pasteur

2 - Les normes internationales (contraignantes ou non)

- Déclaration d'Helsinki (2008)
- CIOMS (recherche incluant des êtres humains 2002 et recherche épidémiologique 2009)
- Lignes directrices de l'Union Européenne (2001)
- Protocole additionnel à la Convention d'Oviedo (Conseil de l'Europe),
- Déclaration universelle de l'UNESCO sur la Bioéthique et les droits de l'Homme
- Lignes directrices de l'OMS et ONUSIDA

Voir aussi le site OMS http://www.who.int/ethics/research/en/index.html

3 – Les sept conditions éthiques pour une recherche impliquant la participation de sujets humains

(ref : Ezekiel J. Emanuel, David Wendler & Christine Grady, « What Makes Clinical Research Ethical ?" JAMA 283 (2000) : 2701-11)

- La valeur sociale, scientifique ou clinique.
- La validité scientifique.
- Une sélection juste des sujets.
- Un rapport risque-bénéfice favorable.
- Une évaluation indépendante.
- Un consentement éclairé.
- Le respect des sujets recrutés.

4 – Quelques définitions

(ref : Networking for Ethics on Biomedical Research in Africa, Sixth Framework Programme (2002 – 2006), Science and Society)

Autonomie

Capacité à envisager les alternatives, à faire des choix, et à agir sans l'influence ou l'interférence indue d'autrui.

Bienfaisance

Principe éthique se rapportant à l'obligation d'optimiser les bénéfices et de minimiser les maux.

Comité d'éthique de la recherche (CER)

Groupe destiné à protéger les droits, la dignité et le bien-être des participants à la recherche en décidant d'approuver ou rejeter un protocole de recherche, ou bien d'y apporter des modifications.

Conseil de contrôle des données et de la sécurité (Data and Safety Monitoring Board – DSMB)

Comité de scientifiques, médecins, statisticiens et autres, chargés de recueillir et d'analyser les données au cours d'un essai clinique afin de surveiller l'apparition éventuelle d'effets néfastes et autres tendances. Les DSMB ont l'autorité requise pour exiger la modification ou

la cessation d'une étude ou la communication d'information complémentaire aux participants de l'étude.

Equitable

Juste. Dans le contexte de la recherche, ce terme est souvent utilisé pour indiquer que le bénéfice et les charges de la recherche sont équitablement distribués parmi les différents groupes de la société.

Indépendance

Appliqué aux comités d'évaluation éthique, ce terme se réfère à la capacité qu'a le comité de prendre ses propres décisions. Un comité n'est pas indépendant si sa conduite est dictée par des agents gouvernementaux ou s'il comprend trop de membres ayant des liens avec des intervenants particuliers tels que les promoteurs de recherche.

Justice

Principe éthique nécessitant une distribution équitable des charges et des bénéfices, souvent exprimé comme le fait de traiter de façon similaire les personnes ayant des circonstances ou des caractéristiques semblables.

Processus de consentement éclairé

Processus par lequel une personne décide de participer ou non à un protocole de recherche. Ce processus inclut typiquement la transmission d'informations écrites et une entrevue en face-à-face, afin d'assurer que les participants éventuels sont dûment informés et comprennent bien les risques, bénéfices et alternatives possibles.

Promoteur (d'un essai médicamenteux)

Toute personne ou entité initiant une recherche clinique sur un médicament – habituellement le fabricant du médicament ou l'institut de recherche ayant développé le médicament. Le promoteur n'effectue pas la recherche mais distribue le nouveau médicament à des chercheurs et des médecins chargés de mener les essais cliniques.

Protocole

Document qui définit l'objectif, les conditions de réalisation et le déroulement de l'essai.

Risque

Probabilité et étendue d'un préjudice ou d'une blessure (sur le plan physique, psychologique, social ou économique) survenant du fait de participer ou d'avoir participé à un projet de recherche. La probabilité et l'étendue possible du préjudice peuvent varier du négligeable au significatif.

Sujet de recherche

Personne dont les caractéristiques et les réactions physiques ou comportementales sont étudiées dans le cadre d'un projet de recherche.

Transparence

Principe éthique incitant les instances décisionnelles à rendre disponible et accessible au public le processus de décision, ceci par le biais d'une communication claire et fréquente d'informations concernant la manière dont les décisions sont prises et pour quelles raisons.

Volontaire

Sans coercition, menace ou contrainte indue. Terme utilisé dans le contexte de la recherche pour qualifier la décision que prend une personne de participer (ou continuer à participer) à un projet de recherche.

Deuxième partie : Différents Types d'Etudes

DEUXIEME PARTIE: LES DIFFERENTS TYPES D'ETUDES

La question de recherche est valide, on est convaincu de l'utilité d'y répondre. Il est justifié d'écrire un protocole, dont la réalisation dépendra de sa qualité.

Différents types d'études sont à la disposition du chercheur pour élaborer son protocole. Chacune d'elles présente des avantages et des faiblesses en termes de niveau de preuve, de contraintes méthodologiques ou de difficultés logistiques, de coût ou de durée.

Seront successivement envisagés les études transversales, les études de cohorte, les études cas-témoins, les essais cliniques, la méta-analyse, les études de stratégies diagnostiques, les études économiques et les analyses de décision.

Il n'y a pas de bons et de mauvais types d'études. Toute étude apporte des informations utiles dans la mesure où son type est approprié à la question posée et où son protocole est bien conçu et bien exécuté. Si un bel essai randomisé apporte une réponse claire et nette à un problème thérapeutique, une modeste série de cas peut, par sa description, soulever une hypothèse intéressante qui ne demande qu'à être vérifiée par une étude analytique ultérieure. Le prospectif n'est pas de principe meilleur que le rétrospectif. La randomisation n'est pas toujours réalisable car elle peut se heurter aux règles de l'éthique.

L'important pour le chercheur est de trouver la meilleure adéquation entre l'objectif de sa recherche et le type d'étude choisi, mais aussi le meilleur compromis entre le niveau de preuve optimal lié à un plan d'étude très élaboré et la faisabilité du projet.

Le but de l'étude est de donner l'estimation la plus correcte de la vérité sur le phénomène observé.

CHAPITRE IV

LES ETUDES TRANSVERSALES

Pierre Duhaut, Jean Schmidt

En clinique, on aborde la plupart des questions en faisant référence à la fréquence de l'événement considéré. Ce sont des fractions, ou des proportions, qui donnent une idée de la fréquence de ces événements cliniques, avec au numérateur le nombre de cas, et au dénominateur la population d'où sont issus ces cas.

La première des mesures de fréquence, dont traite ce chapitre, est la prévalence.

En médecine, compter les événements, qu'ils soient bénéfiques ou adverses (mort, maladie, handicap, inconfort, insatisfaction et leur inverse), est le prélude indispensable à toute analyse ou interprétation ultérieure. C'est l'objectif fondamental de l'étude de prévalence, ou étude transversale, dont les avantages et les limites apparaîtront en fin de chapitre. Il est en effet difficile, au delà de l'observation et du dénombrement, d'établir la séquence temporelle des événements considérés. De plus, l'estimation d'une causalité est toujours une démarche hasardeuse dans ce contexte.

Mais c'est une bonne base de départ.

Plan du chapitre

Les études transversales, encore appelées études de prévalence sont ainsi nommées car elles analysent la présence d'un facteur donné ou d'une maladie particulière dans une population P à un moment précis t, sans référence au passé et sans suivi dans le futur. Elles représentent l'équivalent d'un sondage rigoureusement et scientifiquement construit, ou de l'instantané photographique d'une situation précise dans la population étudiée.

Les études transversales sont avant tout descriptives, et non pas analytiques comme le sont les études cas-témoins, les études de cohorte ou les essais randomisés. Elles sont particulièrement utiles pour apporter des connaissances quantitatives précises sur la répartition d'une maladie ou d'un facteur de risque dans une population, sa fréquence, les sous-groupes de la population plus particulièrement affectés.

Les résultats des études transversales sont donc importants dans deux principaux domaines d'applications:

- La mise en oeuvre de programmes de santé publique, préventifs ou curatifs, en permettant de circonscrire les groupes de la population dans lesquels le programme doit être appliqué en priorité (tranches d'âge, population urbaine ou population rurale, hommes ou femmes, zones géographiques, …). La définition optimale du champ d'application du programme permet une allocation optimale des ressources humaines et matérielles qui y sont consacrées et représente une des conditions indispensables à son efficacité. Par exemple, l'étude de la prévalence des formes résistantes et non-résistantes du paludisme au niveau des diverses régions du monde permet la mise en place adéquate des programmes de lutte contre le paludisme de l'OMS, et est à l'origine du type de conseil prophylactique donné aux voyageurs.

- La constatation dans une étude transversale d'associations entre un état pathologique et une ou des conditions pouvant être supposées causales conduit à la formulation d'hypothèses étiologiques à tester dans d'autres études, de nature différente (biologiques ou épidémiologiques). Par exemple, l'association entre séropositivité pour l'hépatite B et hépatocarcinome en Asie du Sud-Est a conduit à la réalisation des études cas-témoins, puis des études de cohorte qui ont prouvé la relation de cause à effet, en conjonction avec les études biologiques montrant l'intégration du génome du virus dans l'ADN des cellules néoplasiques [1].

Parmi les études descriptives, on distingue:

- d'une part les rapports de cas, les séries de cas et les études écologiques;

- d'autre part les études de prévalence, qui représentent un type particulier d'étude descriptive, à la frontière des études analytiques (que sont les études cas-témoins, les études de cohorte et, sur le versant expérimental, les essais randomisés).

I - CAS RAPPORTES, SERIES DE CAS ET ETUDES ECOLOGIQUES

A - Les cas rapportés

Les cas rapportés, en décrivant une observation inhabituelle, constituent souvent la première étape de la reconnaissance d'une nouvelle maladie ou d'un nouveau facteur de risque. Par exemple, l'association thrombo-embolie et prise d'oestro-progestatifs a été rapportée une

première fois en 1961 chez une patiente, largement discutée sur des séries plus importantes [2] avant d'être l'objet des multiples études, notamment cas-témoins, prouvant sa réalité [3].

B - Les séries de cas

Elles représentent l'étape suivante en regroupant différentes observations similaires et en établissant ainsi l'existence probable d'une entité pathologique. Elles peuvent dans certains cas suggérer très fortement un facteur étiologique.

Par exemple, Thomas Hodgkin en 1832 avait identifié 7 patients atteints d'anomalies tumorales similaires de la rate et des ganglions, 70 années avant que la cellule de Sternberg soit décrite comme pathognomonique de la maladie et que l'entité nosologique puisse ainsi être formée. Plus près de nous, le diagnostic de pneumonie à *Pneumocystis carinii* avec candidose buccale chez 4 sujets jeunes sans antécédents particuliers, homosexuels de sexe masculin, a conduit à la découverte du SIDA et portait déjà en germe la reconnaissance d'un des facteurs de risque de la maladie [4] .

Les séries de cas et les cas rapportés, cependant, traduisent avant tout le plus souvent l'expérience et l'observation d'un auteur et ne permettent pas de tirer de conclusion que l'on puisse généraliser à d'autres cas. De plus, et malgré leur indiscutable utilité, les séries de cas et *a fortiori* les cas rapportés ne permettent pas d'établir la fréquence d'une maladie: une étude d'incidence ou de prévalence serait nécessaire pour cela. Les séries de cas ne permettent pas non plus d'apprécier de manière statistique l'importance d'un facteur de risque qu'elles peuvent éventuellement suggérer. Un groupe de comparaison serait là nécessaire.

C - Les études de corrélation

Les études de corrélation, ou études écologiques, permettent une analyse à plus vaste échelle. Elles établissent la comparaison entre l'importance (ou la fréquence) d'un facteur de risque supposé au sein d'une population et la prévalence ou l'incidence de la maladie supposée secondaire, à partir de données déjà disponibles au niveau de cette population (l'incidence du cancer pulmonaire dans diverses régions du globe est-elle proportionnelle à la quantité de cigarettes fumées dans ces régions ? L'incidence des maladies cardio-vasculaires est-elle proportionnelle à la quantité de graisses animales ingérées ?). Les études écologiques utilisent donc non pas des données recueillies à l'échelle individuelle, mais des moyennes calculées au niveau d'une population.

Les études écologiques introduisent d'une part la notion de comparaison: il faut posséder les données de différentes populations pour pouvoir établir une corrélation entre l'importance d'un facteur de risque et l'importance de la maladie étudiée dans chaque population. Elles utilisent d'autre part la notion de fréquence du facteur de risque et de la maladie. Elles sont transversales dans la mesure où elles superposent deux types de données (fréquence du facteur de risque et fréquence de la maladie) recueillies dans une même période de temps. Elles sont enfin faciles à réaliser en un temps limité, car elles font appel à des données de statistique descriptive déjà collectées et publiées. Elles permettent d'émettre des hypothèses intéressantes, comme le rôle éventuel des pesticides dans la pathogénie des cancers de prostate dont l'incidence diffère selon le degré d'exposition en Martinique [5].

Elles présentent cependant des défauts importants qui en rendent l'interprétation hasardeuse:

Les données sont des moyennes décrivant les caractéristiques d'une population. Elles ne permettent pas de savoir si la personne exposée est effectivement celle qui a développé la maladie, et par conséquent si le facteur de risque est à considérer comme tel. S'il existe une corrélation entre concentration des herbicides et cancer de la prostate, est-ce vraiment les personnes –au niveau individuel- les plus soumises aux pesticides qui développent le plus de cancers de la prostate ?

Les études écologiques ne permettent pas, non plus, le contrôle de facteurs de confusion, même si une analyse multivariée peut être réalisée. Par exemple, on peut imaginer que la quantité de pesticides présents dans la terre et l'eau peut modifier l'environnement microbiologique, et favoriser l'émergence d'un carcinogène viral ou autre. Où sera le carcinogène impliqué dans la pathogénie du cancer de la prostate ? Dans les herbicides, les facteurs modifiés de l'environnement microbiologique, voire tout autre facteur environnemental au sens large du terme, y compris les facteurs alimentaires, les facteurs toxiques, les indicateurs de niveau de vie –incluant tout un ensemble de déterminants sociaux au sens large du terme, dont certains de consommation-, tous corrélés aux précédents ? Quelles sont les interactions entre ces différents facteurs (fig. 1) ?

Enfin un facteur de risque réel peut ne pas être identifié par une étude écologique s'il est "dilué" par les caractéristiques propres de la population. Par exemple, la relation entre ingestion de graisses saturées et maladie coronaire est facile à mettre en évidence dans des populations relativement âgées, où l'incidence des maladies coronaires est importante. La même relation dans des populations à moyenne d'âge plus jeune, consommant des graisses saturées en grande quantité, pourrait passer inaperçue par effet de "dilution" du groupe d'âge à risque dans les autres couches de la population.

Les études écologiques, autorisant des comparaisons à grande échelle, souffrent cependant du manque d'information individuelle: que et qui étudie-t-on précisément ? Quels sont les facteurs de confusion pouvant expliquer l'association observée par leur relation avec chacune des variables associées ? Dans quelle mesure l'association de moyennes calculées dans une population décrit-elle l'association facteur de risque - maladie effectivement présente au niveau de l'individu atteint ?

II - ETUDES DE PREVALENCE

A - Incidence *versus* prévalence

La prévalence indique le pourcentage de personnes présentant la maladie à un moment donné dans une population donnée:

$$\text{Prévalence} = \frac{\text{Nombre de personnes malades à l'instant t}}{\text{Population totale considérée}}$$

L'incidence indique le pourcentage de nouveaux cas diagnostiqués au cours d'une période de temps dans une population donnée:

$$\text{Incidence} = \frac{\text{Nouveaux cas diagnostiqués durant une période p}}{\text{Population totale considérée}}$$

On exprime le plus souvent la prévalence en nombre de cas / 100.000 habitants, et l'incidence en nombre de cas / 100.000 habitants / an.

Les études transversales ne prennent pas en compte la variable temps. Les études déterminant l'incidence d'une maladie ne sont pas, par définition, des études transversales. L'incidence est déterminée par les études de cohorte.

La prévalence est supérieure à l'incidence en cas de maladie chronique, et inférieure en cas de maladie aiguë (curable ou non).

Exemples:

- La prévalence de la polyarthrite rhumatoïde est de loin supérieure au nombre de nouveaux cas diagnostiqués chaque année (incidence). Elle donne par conséquent une meilleure idée du poids de la maladie et de ses conséquences sociales et individuelles que l'incidence.

- Le nombre de personnes intoxiquées par l'amanite phalloïde le 30 Avril (prévalence) est en revanche inférieur à l'incidence annuelle de l'intoxication, et ne donne qu'une idée très incomplète, voire sans valeur, de l'importance du problème. Cependant, la prévalence dans une région donnée permet d'estimer le nombre de lits de réanimation nécessaires pour accueillir les sujets atteints.

- Le nombre de fumeurs en France (prévalence) permet de mesurer pleinement l'importance sociale du phénomène, élément de décision majeur dans la discussion de l'opportunité d'une campagne anti-tabac. Le nombre de personnes commençant à fumer durant les 6 mois précédents et suivant une campagne de mise en garde contre les effets nocifs du tabac (incidence) permet de mesurer une tendance et d'apprécier l'efficacité de la campagne.

Prévalence et incidence ont donc chacune leur utilité propre, et apportent une information différente et complémentaire.

Les études de prévalence combinent les qualités respectives des séries de cas et des études écologiques, tout en éliminant certains de leurs défauts:

- le recueil des données se fait au niveau des individus, et permet d'identifier des possibles facteurs de confusion et de les contrôler;

- les conditions de sélection du groupe étudié permettent des calculs de statistique descriptive;

- il existe un groupe de comparaison, dont on verra cependant qu'il faut se méfier.

B - Constitution d'une étude de prévalence

Elle procède par étapes successives, similaires à celles de toute étude épidémiologique:

- quelle est la question ? Comment poser la question ?

- à quelle population s'adresse-t-elle ?

- comment sélectionner un échantillon représentatif de cette population ?

- comment quantifier et analyser les données ?

- comment interpréter les résultats ?

- au vu des réponses apportées à ces questions ou des problèmes qu'elles soulèvent: en définitive, la structure d'une étude de prévalence est-elle adaptée au problème à résoudre ? Sinon, quel autre type d'étude épidémiologique vaudrait-il mieux choisir ?

1 - Question et population, échantillonnage, biais

La question peut être simple (quelle est la prévalence de la maladie X au sein de la population Y) ou double (quelle est la prévalence de la maladie X dans la population Y, et y-a-t-il une association avec le facteur Z ?).

La population doit être définie très précisément tant sur le plan géographique, que sur le plan des caractéristiques individuelles (âge, sexe, …).

Exemples:

- Etude de la prévalence de l'hémochromatose dans la Région Picardie. Toute modification de la zone géographique peut être à l'origine de résultats différents: la distribution des gènes du système HFE, plus importante dans les populations d'origine nordique, voisine de celle observée en Bretagne, peut ne plus être la même si l'on étend l'étude à la Champagne ou l'Ile-de-France voisine [6].

- Etude de la prévalence de l'asthme allergique chez l'enfant dans la région parisienne. L'inclusion ou l'exclusion des départements limitrophes ruraux est susceptible de modifier profondément les chiffres, car la distribution des allergènes n'est pas ubiquiste: la végétation diffère considérablement dans les départements à forte dominance agricole (Picardie), d'élevage (Normandie) ou principalement urbains (Ile-de-France).

Si la population considérée est importante, un échantillonnage devient nécessaire (fig. 2). Il doit être représentatif de la population initiale et de taille suffisante pour permettre des conclusions valables. L'estimation de la taille dépend de la question posée et de la prévalence supposée des facteurs mesurés.

La représentativité de l'échantillon ne peut être assurée que par tirage au sort des personnes, à condition que toute personne de la population initiale ait une probabilité égale à celle de toute autre personne d'être tirée au sort. Ceci suppose de disposer d'une liste complète et actuelle de la population, où chaque individu ne figure qu'une seule fois, sous un seul numéro ou code d'identification. Ces listes "parfaites" sont rares en pratique dès lors que l'on s'intéresse à la population générale. Les bases de données habituelles (annuaire téléphonique, données de recensement, liste des assurés sociaux, …) ne font que s'en approcher. Elles sont encore plus difficiles à obtenir si l'on veut étudier un sous-groupe sociologique particulier. La liste des salariés d'une entreprise représente un exemple privilégié de liste parfaite lorsque l'on étudie une maladie professionnelle.

L'étape suivante consiste à définir les cas, et les difficultés ne sont pas propres aux études transversales. Le problème de la définition de l'exposition se pose si l'étude ne se limite pas à définir une prévalence, mais cherche à mesurer l'association de la maladie avec un facteur de risque présumé. Là encore, les difficultés ne sont pas propres aux études transversales: savoir quand un sujet est soumis à un facteur de risque procède des mêmes interrogations que dans les études cas-témoins et les études de cohorte.

Le type de la variable "exposition" en revanche rend l'interprétation des résultats plus ou moins hasardeuse, et trois cas de figure peuvent se présenter:

- L'exposition est fixe dans le temps et n'influence pas *a priori* l'âge d'apparition et la longueur d'évolution de la maladie.

  Ces facteurs de risque doivent être des caractéristiques présentes à la naissance et ne subissant pas de modification au cours de la vie.

  Dans l'étude des relations entre spondylarthrite ankylosante et antigène HLA B27, la question de temporalité ne se pose pas, l'antigène HLA B27 étant présent avant le début de la maladie et sa nature ne variant pas au cours de la vie. Une étude transversale peut facilement mettre en évidence la prévalence plus importante de l'antigène dans la population malade que dans la population saine.

- L'exposition est fixe dans le temps, mais peut influencer l'âge d'apparition ou la durée d'évolution de la maladie.

  Comme précédemment, le facteur de risque doit être présent à la naissance et ne pas subir de modification au cours de la vie. La question de temporalité ne se pose pas, puisque le facteur considéré précède obligatoirement la maladie (si elle est acquise au cours de la vie). Mais l'influence du facteur de risque sur l'âge d'apparition ou la durée d'évolution de la maladie peut être à l'origine d'un biais de survie sélective.

  Une étude transversale visant à tester l'association entre trisomie 21 et leucémie aiguë dans la population adulte pourrait montrer une fréquence de trisomie moindre chez les patients leucémiques et conclure à tort que la trisomie protège de la leucémie, car les enfants trisomiques, plus à risque que les enfants non trisomiques de développer une leucémie et d'en mourir jeunes, ne seraient pas inclus dans l'étude transversale du fait de leur mort précoce (fig. 3).

  L'étude de cohorte aurait montré le risque accru de leucémie aiguë chez les sujets trisomiques. L'étude transversale, ne pouvant inclure de leucémique trisomique, conclut à tort à un effet protecteur de la trisomie et fait donc disparaître le rapport de causalité.

  En pratique, il est souvent facile de savoir si le facteur de risque supposé est fixe dans le temps. Il est beaucoup plus difficile de savoir précisément s'il influence l'âge d'apparition et la durée d'évolution de la maladie, et différencier les cas 1 et 2 n'est pas toujours aisé. La suspicion de biais de survie sélective existe donc également lorsque l'on pense, sans pouvoir en être absolument sûr, se trouver dans la situation 1.

- L'exposition n'est pas fixe dans le temps.

  Il s'agit d'un facteur de risque acquis à un moment de la vie, et dont l'intensité peut être variable dans le temps. C'est le cas de la majorité des facteurs de risque étudiés en pathologie (facteurs nutritionnels, tabagisme, alcoolisme, contamination virale, bactérienne, parasitaire, exposition professionnelle, intoxication accidentelle, …).

  La relation de cause à effet est là très difficile à établir, car il se pose deux problèmes majeurs:

  - La séquence temporelle: maladie et facteur de risque supposé sont déterminés en même temps. Lequel précédait l'autre ? La réponse peut être facile si l'on peut de façon fiable reconnaître dans le passé une exposition survenue à une date précise (accident nucléaire et prévalence des malformations congénitales dans la population touchée). Elle peut être

beaucoup plus difficile, voire impossible à obtenir par une étude transversale seule, dans d'autres situations et tout particulièrement lorsque la physiopathologie d'une maladie reste mystérieuse. Dans l'association entre la présence d'anticorps anti-nucléaires et les manifestations de la maladie lupique, les anticorps anti-nucléaires sont-ils à l'origine des lésions observées, ou n'apparaissent-ils que comme conséquence de la destruction cellulaire causée par un facteur X, mettant les antigènes intracellulaires au contact du système immunitaire qui peut alors produire des anticorps contre les antigènes ainsi mis à nu ? La question n'est pas encore définitivement résolue, même s'il semble que la levée des anticorps soit annonciatrice de la reprise évolutive de la maladie, qui n'est peut-être que le stade ultime, apparent, du processus destructeur initial révélé plus tôt par la ré-ascension du taux des anticorps.

- L'exactitude de la mesure de l'exposition: lorsque l'exposition est variable dans le temps, que vaut-il mieux mesurer ? L'exposition au moment de l'étude transversale, qui peut être définie avec le maximum d'exactitude, mais qui, concomitante de l'état pathologique, n'est pas forcément celle qui a induit la maladie ? Ou l'exposition dans le passé, plus susceptible d'avoir induit la maladie, tout particulièrement lorsqu'il existe une période de latence longue, mais dont la détermination repose sur les souvenirs des sujets et est souvent imprécise ?

Exemple: prévalence des maladies cardio-vasculaires et teneur en acides gras à chaîne courte de l'alimentation. Quantifier la teneur en lipides de l'alimentation des sujets au moment de l'étude transversale est possible. Mais l'alimentation actuelle est-elle le reflet de l'alimentation des années passées, véritable facteur de risque ? A l'opposé, comment mesurer la teneur en lipides de l'alimentation des sujets 5, 10, ou 20 ans avant la réalisation de l'étude ?

Le biais de souvenir est un des facteurs limitants, très important, des études transversales.

Dans la plupart des cas enfin, le facteur de risque acquis, variable dans le temps, influence l'âge de survenue et la durée d'évolution de la maladie. Au biais de souvenir et au problème de temporalité, s'ajoute donc le biais de survie sélective, et l'interprétation de l'étude transversale en est d'autant plus aléatoire.

2 - Mesures effectuées, expression mathématique

Les résultats peuvent être exprimés sous forme de table (tableau 1).

Dans cette présentation, la prévalence (seule mesure vraiment rigoureuse autorisée par ce type d'étude), s'écrit:

$$\text{Prévalence} = \frac{a + c}{a + b + c + d}$$

On peut définir également un taux de prévalence, qui répond à la question suivante: combien de fois la maladie est-elle plus fréquente chez les sujets exposés par rapport au non-exposés, dans la population examinée dans l'étude transversale ?

Il faut calculer la prévalence de la maladie chez les sujets exposés (Prévalence 1), et chez les sujets non exposés (Prévalence 2):

$$\text{Prévalence 1} = \frac{a}{a+b} \quad \text{Prévalence 2} = \frac{c}{c+d}$$

$$\frac{\text{Prévalence 1}}{\text{Prévalence 2}} = \frac{a}{a+b} \bullet \frac{c+d}{c}$$

Il faut prendre garde au fait que le taux de prévalence n'est pas l'équivalent d'un risque relatif, qui répondrait à la question suivante: combien de fois les sujets exposés ont-ils plus de risque d'être atteints par la maladie que les sujets non exposés ? Le risque relatif calculé dans une étude de cohorte ou approché dans une étude cas-témoins mesure le "pouvoir pathogène" du facteur d'exposition.

Pour que le taux de prévalence puisse approcher le risque relatif, il faudrait:

- qu'il n'y ait pas de biais de survie sélective;

- qu'il n'y ait pas de biais de souvenir;

- qu'il y ait une véritable relation de cause à effet entre le facteur de risque supposé et la maladie, ce qui est impossible à prouver par une étude transversale isolée, du fait même de sa structure; il y a dans l'étude transversale, *juxtaposition d'un facteur de risque supposé et de la maladie, mais on ne sait pas lequel a précédé l'autre.*

- que la durée de la maladie chez les sujets exposés soit la même que la durée de la maladie chez les sujets non exposés, de telle sorte que sujets malades exposés et non exposés aient la même chance d'être inclus dans l'étude transversale en tant que patients. Si tel n'était pas le cas, nous risquerions de nous trouver dans la situation suivante (fig. 4): aucun des sujets non exposés ayant contracté la maladie ne se trouve inclus en tant que patient dans l'étude transversale, et pourtant le risque de contracter la maladie pour les non exposés (3 / 3) est le même que pour les sujets exposés (3 / 3). La comparaison des prévalences amènerait à la conclusion fausse que la maladie n'existe pas chez les sujets non exposés et par conséquent que ceux-ci ne sont pas à risque pour la maladie considérée.

En revanche, l'information intéressante, à savoir l'augmentation de la durée de la maladie chez les sujets exposés, n'est pas obtenue par l'étude transversale.

Interpréter de manière abusive un taux de prévalence conduit donc à des conclusions erronées. Les quatre conditions détaillées plus haut étant rarement réunies, il ne faut pas inférer à partir du taux de prévalence une quelconque relation causale ou l'importance d'un risque relatif. Le taux de prévalence ne répond donc qu'à la question, rapidement posée, rapidement résolue, de la fréquence relative, à un instant t, de la maladie parmi les sujets exposés et non exposés à un facteur de risque supposé. Il laisse beaucoup d'autres questions posées sans réponse fiable. La constatation d'une différence de fréquence peut cependant servir d'hypothèse à une étude cas-témoins, une étude de cohorte ou des expérimentations biologiques visant à confirmer ou à infirmer le rôle pathogène du facteur de risque.

III - AVANTAGES ET FAIBLESSES DES ETUDES TRANSVERSALES

A - Avantages

- Les études transversales sont les seules à pouvoir établir la prévalence. Cette mesure est tout particulièrement utile pour apprécier l'ampleur d'un phénomène, les répercussions sociales d'une maladie, sa distribution géographique. Elle est nécessaire pour pouvoir ajuster le nombre et la qualité des structures de soins aux besoins rencontrés dans la population.

- Elles possèdent un groupe de comparaison et permettent ainsi d'étudier l'association entre un état pathologique et un facteur de risque supposé.

- Elles permettent d'étudier simultanément l'association entre plusieurs états pathologiques et plusieurs facteurs de risque supposés. Elles servent ainsi de génératrices d'hypothèses pour des études plus élaborées de type études cas-témoins ou études de cohorte.

- Elles peuvent représenter une première étape d'une étude de cohorte (phase d'inclusion des sujets).

- Elles peuvent être réalisées dans un laps de temps relativement court, et sont donc peu coûteuses.

- Des facteurs de confusion éventuels peuvent être contrôlés, par stratification des sujets malades et sains en fonction de l'élément de confusion.

B - Faiblesses

- Elles ne permettent pas d'établir la séquence temporelle des événements. Constater une association entre un état pathologique et un facteur de risque supposé n'autorise par conséquent pas à en déduire une relation de cause à effet.

- Elles ne permettent pas d'estimer une association lorsque la maladie est rare dans la population, car elles exigeraient, pour pouvoir inclure un nombre suffisant de sujets malades, une taille d'échantillon trop importante.

- Elles sont soumises à la possibilité de biais de survie sélective.

- Elles sont soumises au biais de souvenir.

- Elles sont soumises à l'existence toujours possible de facteurs de confusion non prévus.

- La prévalence ne permet pas d'estimer l'incidence, et le rapport des prévalences ne permet pas d'estimer le risque relatif.

- Enfin et surtout, il faut se garder de toute interprétation abusive, souvent tentante.

Fig. 1 - Exemple théorique de facteur de confusion dans l'association pesticides et cancer de la prostate.

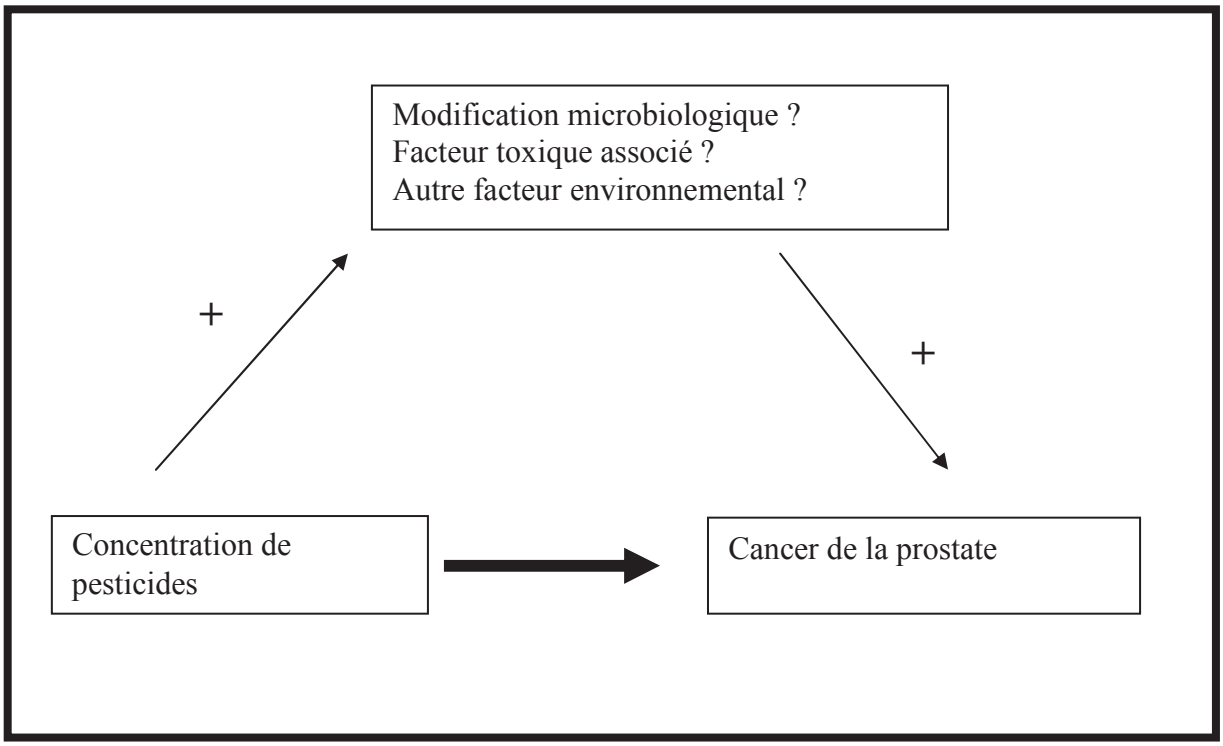


Fig. 2 - Population de référence et ses sous-groupes

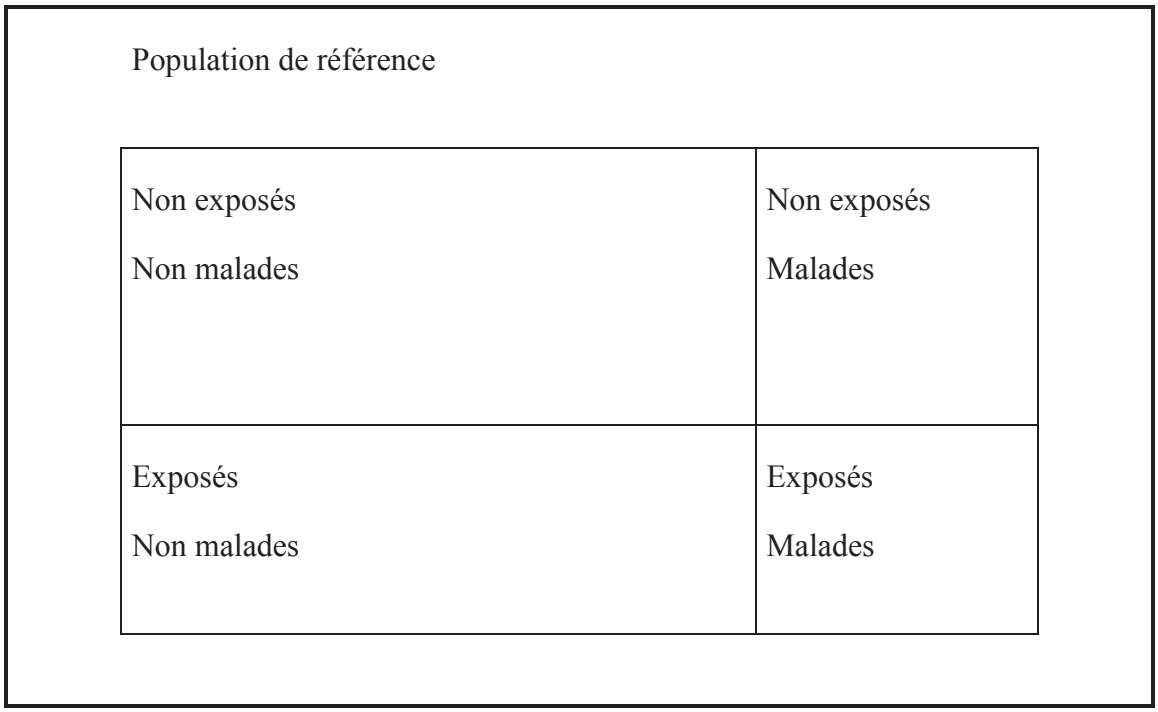


Fig. 3 - Association entre trisomie 21 et leucémie aiguë: représentation d'une étude transversale dans une population adulte (trait vertical) et d'une étude de cohorte (flèche horizontale).

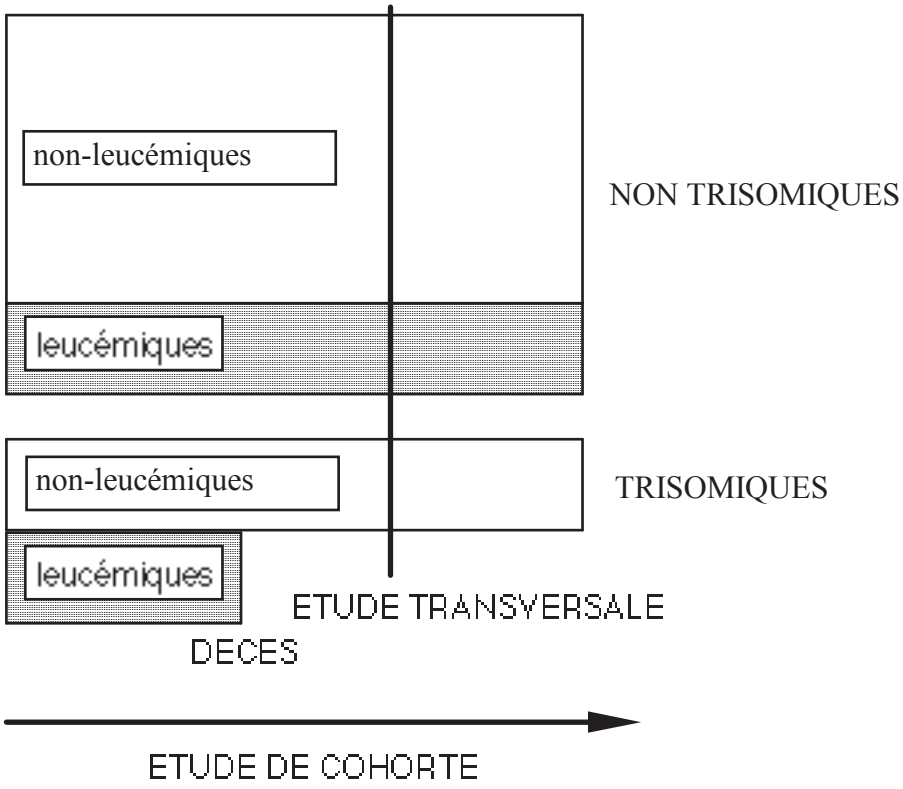


Fig. 4 - Influence de la durée d'une maladie dans une étude transversale.

———————————— Durée de la maladie chez les sujets exposés

---------- Durée de la maladie chez les sujets non exposés

Etude transversale

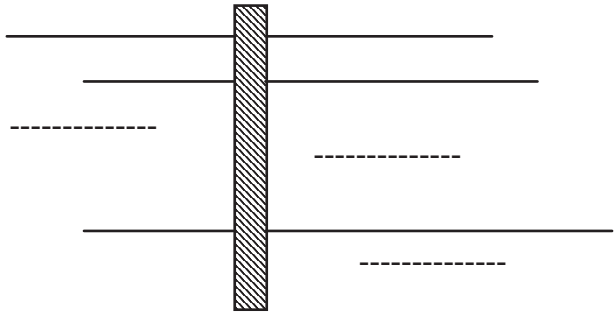


Tableau 1 - Expression des résultats d'une étude de prévalence

	Malades	Non malades	
Exposés	a	b	a + b
Non exposés	c	d	c + d
	a + c	b + d	a + b + c + d

[1] Ganem D, Prince AM. Hepatitis B infection- Natural history and clinical consequences. NEJM 2004;350:1118-1129.

[2] - Tyler ET. Oral contraception and venous thrombosis. JAMA 1963 ;185 :131-132.

[3] WHO Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Effect of different progestagens in low oestrogen oral contraceptives on venous thromoembolic disease. Lancet 1995;346:1582-1588.

[4]- Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, Saxon A. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency.NEJM 1981;305:1425-31.

[5] Belpomme D, Irigaray P, Ossondo M, Vacque D, Martin M. Prostate cancer as an environmental disease: an ecological study in the French Caribbean islands, Martinique and Guadeloupe. Int J Oncol 2009;34:1037-1044.

[6] - Merryweather-Clarke AT, Pointon JJ, Jouanolle AM, Rochette J, Robson KJ. Geography of HFE C282Y and H63D mutations. Genet Test 2000;4:183-98.

.

.

CHAPITRE V

LES ETUDES CAS-TEMOINS

Pierre Duhaut, Jean Schmidt

On a vu dans le précédent chapitre que l'étude transversale permettait, à peu de frais, de dénombrer les événements cliniques qui présentent un intérêt pour le médecin. En d'autres termes, elle permet d'apprécier l'ampleur d'un phénomène, par son caractère commun mais aussi par ses répercussions.

Il est caricatural, mais réaliste, de diviser les pathologies en deux catégories, rares et moins rares (ou fréquentes). Le monde universitaire peut confirmer l'importance de ces événements rares: il a été estimé que 90 % des programmes pédagogiques concernent 10 % de la pathologie. Si 40 % d'entre nous mourront d'une maladie cardio-vasculaire, et 20 à 25 % d'un cancer (dont seuls quelques-uns sont des maladies fréquentes), le pronostic des 40 % restants sera plus ou moins lié à une des très nombreuses maladies rares !

Parce qu'ils sont rares, ces événements sont souvent décrits et analysés par des études peu sophistiquées, les rapports de cas isolés ou les séries de cas. La faiblesse principale de ces études est l'absence de groupe témoin, et donc l'impossibilité d'effectuer une quelconque comparaison.

Telle n'est pas la situation avec l'étude cas-témoins. C'est une étude plus sophistiquée dans sa conception, dont les avantages, les inconvénients et les contraintes sont décrits dans les pages qui suivent. Elle a manifestement des atouts puisqu'elle bénéficie d'une "cote d'amour" chez les épidémiologistes: 30 à 40 % des publications en épidémiologie sont des études cas-témoins.

L'étude cas-témoins est une étude le plus souvent rétrospective. Comme l'étude de cohorte, il s'agit d'une étude d'observation, analytique, non expérimentale.

Plan du chapitre

Les études cas-contrôle, comme les études en cohorte se rangent dans la catégorie des études d'observation par opposition aux études expérimentales représentées par les essais randomisés. Etudes d'observation, car on examine, sans intervenir sur le patient, les relations possibles entre un ou plusieurs facteurs de risque et la survenue d'un ou de plusieurs états pathologiques.

Contrairement aux études en cohorte, où l'on sélectionne les sujets sur le critère de l'exposition au facteur de risque pour analyser leur devenir et les conséquences pathologiques du facteur de risque présent dans le futur, les études cas-contrôle (dites encore cas-témoins) sont fondées sur la démarche inverse : les sujets sont sélectionnés **dans le présent ou le passé** en fonction de leur statut vis-à-vis de la maladie (ils seront donc cas, sujets atteints par la maladie, ou contrôles, sujets sains) et l'on cherche à déterminer le facteur de risque potentiel qui, **dans le présent ou le passé des sujets**, diffère entre les cas et les contrôles et pourrait donc être impliqué dans la genèse de la maladie étudiée.

Cette approche a été développée en partie pour répondre aux besoins d'étude des maladies chroniques ayant une longue période de latence. Les avantages en sont évidents:

1- Une étude cas-contrôle peut être menée à terme rapidement :
La maladie est déjà déclarée et l'enquête étiologique est habituellement rétrospective. La durée de l'étude est donc indépendante de la durée d'incubation ou de la période de latence.

2- La stratégie cas-contrôle est particulièrement intéressante pour les maladies rares :
L'étude prospective en cohorte d'une maladie rare nécessiterait un très grand nombre de sujets soumis à l'exposition présumée et suivis au cours d'un temps indéterminé avant de voir apparaître un petit nombre de sujets atteints. Le risque, particulièrement important si l'étiologie présumée n'en est pas une ou si le facteur de risque incriminé ne joue qu'un faible rôle dans la pathogénie de la maladie, est de réaliser une étude longue, coûteuse et difficile pour un résultat négatif ou non-informatif.

> Exemple : l'incidence de la maladie de Horton en France est de l'ordre de 10/100 000 habitants/an[1]. Le temps de latence de la maladie est inconnu, mais elle apparaît uniquement chez des sujets de plus de 50 ans, ce qui laisse supposer la nécessité d'un vieillissement physiologique ou pathologique avant que la maladie puisse se déclarer. Il faudrait donc dans une étude de cohorte, suivre plus d'un million de personnes pour espérer rassembler, compte tenu de l'attrition obligatoire de la cohorte, une centaine de cas au terme sans doute de nombreuses années. La stratégie de la cohorte n'est pas adaptée.
> En revanche, l'étude cas-contrôle permettra de rassembler un nombre suffisant de sujets malades pour comparer de manière statistiquement satisfaisante la distribution des facteurs de risque entre malades et sujets sains.

3- On peut analyser dans une étude cas-contrôle un **nombre important de facteurs de risque présumés** que l'on collecte dans les antécédents des sujets, alors que l'on assemble les populations dans une étude en cohorte sur la base de l'exposition des sujets à **un facteur de risque précis** dont on veut comprendre le rôle. En permettant d'explorer un grand nombre d'hypothèses, l'étude cas-contrôle est particulièrement utile lorsque les connaissances sur une maladie donnée sont réduites et qu'il n'existe pas de direction préférentielle d'investigation.

4- Une étude cas-contrôle est par conséquent beaucoup moins coûteuse, tant en argent qu'en temps ou en personnel.

I- STRUCTURE DE L'ETUDE :

Elle est représentée dans la figure 1. Une étude cas-témoins dans sa structure, remonte toujours le temps du présent vers le passé (contrairement à l'étude de cohorte, qui suit le temps dans sa direction 'normale' du présent vers le futur, ou du passé vers le présent).
Trois étapes sont particulièrement importantes dans sa construction :

1- La sélection d'un échantillon d'une population de sujets malades.
2- La sélection d'un échantillon d'une population de sujets sains (témoins ou contrôle).
3- La mesure des facteurs de risque suspectés.

II- ELABORATION DE L'ETUDE :

Clarification du vocabulaire : études rétrospectives et prospectives :
L'expression "étude rétrospective" est synonyme pour beaucoup d'auteurs d'étude cas-contrôle, et l'expression "étude prospective" d'étude en cohorte.

La réalité est cependant plus nuancée. **Une étude en cohorte peut être rétrospective** si l'exposition des sujets a été déterminée de manière précise dans le passé et si les états pathologiques ont été dûment enregistrés. On peut ainsi réaliser une étude rétrospective en cohorte sur une population suivie dans le cadre de la médecine du travail, où tous les employés d'une usine A, soumis à une exposition toxique X accidentelle survenue l'année Y, seront comparés aux employés d'une usine similaire B où l'accident n'aura pas eu lieu. Les pathologies survenues entre l'année Y et le moment de l'étude auront été notées dans les registres du service de Médecine de travail lors des consultations obligatoires biannuelles ou annuelles. Simplement, le temps de l'étude *s'écoule dans le sens physiologique du passé vers le présent.*
De même, le terme prospectif peut s'appliquer aux études cas-contrôle lorsque l'étude ne porte pas sur les cas prévalents ou diagnostiqués dans le passé, mais sur les cas incidents qui seront recrutés durant une période de temps définie à partir du début de l'étude, et associés à des contrôles recrutés durant la même période. A partir du diagnostic des nouveaux cas, la recherche portera sur les facteurs de risque du passé, *et l'étude remonte ainsi le temps du présent vers le passé.*

A- Définition et sélection des cas :

Il s'agit d'un problème majeur dont la non-résolution peut être à l'origine d'un certain nombre de biais, tout particulièrement dans les études cas-contrôle de maladies rares impliquant une coopération multicentrique. Définition de la maladie et sélection des patients sont les deux volets de la question.

1- Définition de la maladie :

Etablir des critères objectifs pour aboutir à un diagnostic reproductible de la maladie peut être souvent assez difficile. Considérons, par exemple, le cas de la polyarthrite rhumatoïde : cette maladie relativement fréquente se présente avec de très nombreux signes cliniques et des tests

de laboratoire dont la sensibilité et la spécificité sont variables en fonction du laboratoire, de l'âge du patient, de la population au sein de laquelle sont faits les tests, de l'association avec des signes cliniques plus ou moins spécifiques... De plus, la variation entre les différents observateurs dans l'interprétation des signes cliniques et des résultats de laboratoire peut être importante. Très souvent, en fait, le diagnostic repose sur un faisceau d'arguments et le sentiment du clinicien formé à partir de l'évolution du patient dans le temps, de l'aspect de l'inflammation articulaire, de sa localisation, de la réponse à la thérapeutique... Le cadre nosologique admet des formes frontières avec d'autres maladies systémiques, et le diagnostic différentiel peut être difficile à établir. La variabilité des critères et la subjectivité du clinicien obligatoirement mise en jeu dans le diagnostic, peuvent rendre l'homogénéité du diagnostic très aléatoire dans une étude multicentrique.

Il faut donc établir une liste de critères diagnostiques, tant symptômes que signes physiques et tests de laboratoire et essayer de définir pour chacun d'entre eux leur **spécificité** et **sensitivité**. Il faut ensuite définir quelle combinaison de critères sera exigée pour admettre le diagnostic (Travail de l'ARA (American Rheumatology Association) pour la polyarthrite rhumatoïde, de la DSM IV (Diagnostic and Statistical Manual-Revision 4) en Psychiatrie, etc). L'établissement de critères diagnostiques est ardu, y compris pour des pathologies que l'on croit bien connaître : ainsi, il n'existe pas en 2009 de critères diagnostiques validés de pneumopathie bactérienne (hormis les critères de gravité justifiant les soins intensifs), les critères diagnostiques du syndrome de Gougerot-Sjögren, du diabète de type 2 ou du syndrome dysmétabolique évoluent de façon constante et se heurtent à des différences de conception entre les différents continents et l'on établira plus souvent –et plus modestement- des critères de classification que des critères diagnostiques. Une maladie relativement bien définie sur le plan anatomo-pathologique, mais très polymorphe sur le plan clinique comme la sarcoïdose, n'a pas à l'heure actuelle de critères diagnostiques ou de classification reconnus ou seulement ébauchés.

La dernière étape consistera à tester la **reproductibilité** et la **validité** des critères sur un échantillon d'investigateurs participant à l'étude :

- Quelle est la **variabilité inter-observateur**? Deux observateurs confrontés au même cas répondront-ils de la même façon?
- Quelle est la **variabilité intra-observateur**? Le même observateur confronté au même cas à distance dans le temps répondra-t-il de la même façon?
- **La validité -ou exactitude-** des critères peut être plus difficile à mesurer. Les critères définissent-ils ou mesurent-ils avec exactitude ce qu'ils sont censés définir ou mesurer? La réponse nécessite l'existence d'un **critère étalon**, ou **gold standard**, qui permet un diagnostic fiable auquel on puisse se référer. De tels étalons n'existent pas toujours, et il faut alors se référer à l'avis des experts, à l'expérience de ceux qui utilisent les critères diagnostiques dans leur pratique quotidienne... ce qui implique l'apparition d'une certaine subjectivité dès la première étape de l'élaboration de l'étude.

2- Sélection des cas :

a- Critères d'inclusion ou d'éligibilité des sujets :
Les sujets inclus doivent bien sûr correspondre aux critères diagnostiques précédemment fixés, mais doivent aussi être représentatifs de l'ensemble de la population des patients à laquelle on souhaite que les résultats de l'étude s'appliquent. Ils peuvent ne pas correspondre à *toute* la population atteinte par la maladie, car les patients doivent avoir une **probabilité**

raisonnable d'être atteints par la maladie **à cause** du facteur de risque étudié, sous peine de voir l'association réelle entre facteur de risque et maladie diluée par des facteurs extérieurs:
Exemple :
Une phlébite avec ou sans embolie pulmonaire peut survenir en ville chez des sujets ambulatoires, mais on sait qu'une proportion importante survient en milieu hospitalier. L'étude de l'association thrombo-embolisme et prise de contraceptifs oraux devrait exclure – ou au moins faire analyser séparément- les cas de thrombo-embolisme survenus en milieu hospitalier. Les patients hospitalisés ont plus de chance d'être alités, d'avoir subi une chirurgie thrombogène (orthopédique notamment), d'être atteints de cancer, ou tout simplement d'avoir accouché… autant de facteurs de risque importants de thrombose veineuse. Par contre, ces mêmes patients, du fait de leur âge, ou de l'accouchement récent ! ont beaucoup plus de chance de ne pas être sous contraception orale que les patientes ambulatoires… une étude effectuée sur une échantillon de patients représentatifs de toute la population atteinte de thrombose veineuse pourrait donc ne pas mettre en évidence le rôle pourtant réel de la contraception orale, par effet de dilution du groupe à risque de pilule dans un groupe *subissant d'autres facteurs de risque plus importants, et ne prenant pas, ou plus, la pilule.*

Il faut choisir enfin entre cas **prévalents** -déjà traités le plus souvent- et cas **incidents**, pour lesquels la mesure de l'exposition peut se faire sans risque de modification de celle-ci par le traitement éventuel ou l'évolution de la maladie et sans risque de confusion entre l'exposition suspectée et l'une des conséquences du processus pathologique.

Exemples :

- problèmes conjugaux comme cause ou conséquence d'une pathologie dépressive où l'interrogatoire peut être difficile et peu fiable.
- Virus de l'hépatite B : agent étiologique des carcinomes hépatiques très fréquents en Asie du Sud-Est ou simple contaminant d'un foie préalablement pathologique ? La question était restée très longtemps débattue avant d'être tranchée par les études prospectives, objectivant la séquence temporelle : 1. Infection par le virus de l'hépatite B, 2. Hépatite B chronique, et 3. Survenue tardive de la transformation cellulaire avec apparition de l'hépatocarcinome, plaidant en faveur du rôle étiologique du virus dans la cancérisation.

b- Origine des cas étudiés :

L'étude peut être limitée aux **cas hospitalisés**. Elle sera alors relativement facile à réaliser et peu coûteuse. Elle peut aussi s'étendre à **l'ensemble de la population** habitant une zone prédéterminée. Elle aura alors l'avantage d'embrasser un spectre beaucoup plus large de la maladie et d'éviter les **biais de sélection** inhérents à toute population hospitalisée. Elle sera par contre souvent beaucoup plus difficile et plus coûteuse à réaliser.

Le principal problème est celui de la **généralisation** des résultats :
Est-il raisonnable, pour la maladie en question, d'étendre les résultats de l'étude à l'ensemble de la population des malades si seuls les patients hospitalisés ont été étudiés?
Est-il utile de réaliser une étude sur patients hospitalisés seulement si le but final consiste par exemple en la reconnaissance et l'éviction d'un facteur de risque dans la population générale?
Est-il utile de réaliser une étude sur l'ensemble de la population si seuls les cas hospitalisés, potentiellement plus graves, posent un problème thérapeutique ou présentent un pronostic sévère?

Le problème n'est pas que théorique : on a généralisé le traitement des patients atteints d'hépatite virale C sur la base des complications observées en milieu hospitalier et notamment des cirrhoses avec insuffisance hépatique ou hypertension portale et des hépatocarcinomes secondaires. Cependant, une des rares études basées sur la population –en double cohorte- des patients atteints d'hépatite post-tranfusionnelle suivis sur une période moyenne de 18 ans montre que leur mortalité globale est similaire à celle des patients sans hépatite, et que le surcroît de mortalité hépatique est avant tout en relation avec l'abus d'alcool qui agit comme modificateur d'effet agoniste, et *ceci bien que tous les patients atteints d'hépatite post-transfusionnelle aient une histologie de cirrhose à la biopsie hépatique* [2]. Faut-il dès lors imposer un traitement lourd et difficile à supporter ou agir sur le facteur alcool ? La question mérite au moins d'être posée.

B- Sélection et définition des contrôles :

Ils doivent être issus de la **population d'où proviennent les cas** afin d'assurer une comparabilité maximale entre les cas et les témoins. Toute exclusion ou toute restriction appliquée aux cas s'applique par conséquent également aux témoins. La population dont les cas sont issus peut être différente de l'ensemble de la population non malade et les sujets contrôles ne sont donc pas obligatoirement représentatifs de la population de sujets sains.
Le choix de la population témoin est crucial dans une étude cas-contrôle, car les résultats de l'étude reposent précisément sur la comparaison entre cas et témoins : du choix des témoins dépendront donc les conclusions et les possibles biais dont elles seront entachées.
En pratique courante, les médecins savent diagnostiquer un cas -avec les réserves émises plus haut sur la validité des critères diagnostiques- mais prêtent souvent peu d'attention aux témoins : de ce fait, ils peuvent être plus importants que les cas pour expliquer les discordances entre études cas-témoins.

Exemple : le traitement hormonal substitutif (THS) de la ménopause a commencé à être prescrit dans les années 60 et ne comprenait que des œstrogènes non contrebalancés par des progestatifs. En 1975, paraissait la première étude cas-témoins dans le *New England Journal of Medicine*, rapportant un excès de cancer du corps utérin chez les patientes traitées. Les patientes avec cancer de l'endomètre et des personnes sans, étaient interrogées sur leur utilisation passée ou présente de THS. L'odds ratio en résultant s'élevait à 4,5 pour l'apparition d'un cancer sous THS, de façon significative[3].
La critique à cette étude vint du mode de diagnostic du cancer de l'endomètre post-ménopausique : il fallait, pour que le diagnostic puisse être fait, que la patiente présente une métrorragie motivant le curetage biopsique. Or, les œstrogènes favorisent les métrorragies… donc les patientes métrorragiques avaient plus de chance d'avoir pris des œstrogènes (biais de détection) que des patientes non métrorragiques (par conséquent non biopsiées). Ainsi l'odds ratio pouvait découler du fait que la sélection des cas n'avait pas été faite uniquement sur la présence de la maladie, mais aussi sur la présence du facteur de risque (induisant la métrorragie, et par là le diagnostic).
Une deuxième étude, pour pallier le biais potentiel de la première, n'inclut donc que des patientes venues consulter pour métrorragies, chez lesquelles un curetage était effectué. Le biais de détection par la métrorragie disparaissait. Les patientes avec cancer diagnostiqué sur curetage constituaient le groupe des cas, les patientes sans cancer, le groupe des témoins. L'odds ratio, cette fois-ci, était égal à 1, innocentait donc le THS… et prouvait que les résultats de la première étude découlaient du biais de détection[4].

La preuve ne tint pas longtemps : si les œstrogènes induisaient les métrorragies, alors les témoins de la deuxième étude *étaient aussi plus à risque d'avoir pris des œstrogènes*... et par conséquent, n'*avaient pas été sélectionnés uniquement sur la base de l'absence de cancer de l'endomètre, mais aussi sur la base de la présence de l'exposition au THS*... Il n'était donc plus possible de mettre en évidence une différence avec les cas, puisque cas et témoins avaient été sélectionnés, en fait, sur la base de leur exposition au THS !
Plus de 20 études cas-témoins correctement construites se sont succédées sur le sujet 'Cancer de l'endomètre et THS', chacune essayant de pallier les biais potentiels de la précédente. Les odds ratio ont varié de 0,5 (protecteur) à… près de 20, démontrant ainsi qu'ils dépendaient avant tout du choix des témoins.
Finalement, une étude associant plusieurs groupes témoins différents[5], et les études de cohortes construites pour trancher définitivement la question, ont retrouvé des odds ratio, et un risque relatif, compris entre 2 et 3 et similaire à celui de… la première étude.

II.2.3.a- Origine des témoins :

Différentes sources sont possibles.

Les contrôles recrutés parmi **les patients hospitalisés** pour une raison différente de la maladie étudiée sont très fréquemment utilisés et présentent de nombreux avantages. Ils sont disponibles, faciles à contacter et les données seront donc peu coûteuses à collecter. Ils sont soumis aux **mêmes biais de recrutement** que les cas hospitalisés dans le même département, et ceci diminue les **biais différentiels de sélection**. Ils sont souvent, en tant que malades, plus enclins à collaborer que des sujets sains tirés au sort dans la population générale. Ils peuvent être plus conscients de leurs antécédents médicaux que des sujets interviewés à froid dans la population générale, ce qui diminuera les **biais de souvenir** (recall bias). Enfin, ils peuvent être examinés et interrogés par le même médecin que celui qui examinera et interrogera les cas, et ceci diminuera les **biais d'information** pouvant provenir d'un interrogatoire ou d'un examen effectué par deux observateurs différents.

Les contrôles recrutés en milieu hospitalier présentent cependant des **inconvénients certains**. Ils ne sont tout d'abord **pas ou peu représentatifs** de la population saine, et l'on sait que les habitudes toxiques, tabac ou alcool par exemple, sont très différentes chez des sujets hospitalisés par rapport à celles observées dans la population générale.
Bien qu'hospitalisés dans le même hôpital ou le même service que les cas, les témoins peuvent induire un **biais de sélection** si l'hôpital est un centre de référence pour le traitement ou le diagnostic de la maladie étudiée, mais non pas des maladies pour lesquelles seront hospitalisés les témoins. Les cas peuvent alors provenir d'une population beaucoup plus large que les témoins, et les deux groupes ne seront plus comparables.

Enfin, les témoins peuvent être hospitalisés pour une affection présentant des **facteurs de risque communs** avec la maladie étudiée : l'association tabac-cancer du poumon ne pourrait être mise en évidence dans un service de pneumologie utilisant des bronchitiques chroniques comme témoins.

Choisir les témoins dans la **population générale** dont proviennent les cas assure le niveau maximum de **comparabilité** entre cas et contrôles. Les témoins peuvent être tirés au sort sur les listes électorales, être contactés par composition de numéros de téléphone obtenus à l'aide de tables de nombres au hasard, ou choisis au hasard sur les registres communaux... Il peut cependant être difficile de contacter des personnes engagées dans la vie active, les personnes

contactables peuvent ne pas être représentatives de la population générale (personnes âgées retraitées, personnes en arrêt-maladie à domicile...). Des sujets sains peuvent ne pas avoir prêté attention à tel facteur de risque dans leurs antécédents, alors que les patients ont pu y réfléchir longuement et **cette disparité de souvenir** introduira un biais dans l'évaluation de l'exposition dans les deux groupes. Enfin, des sujets sains peuvent n'être que très peu motivés pour participer à l'étude et ceci peut introduire un **biais d'information** conséquent.

Des témoins choisis parmi **la famille, les amis ou les voisins** apportent un certain nombre de solutions aux problèmes évoqués. Ils partagent les caractéristiques de la population saine générale, mais peuvent être plus motivés pour participer à l'étude. Ils permettent d'éviter les biais de sélection relatifs au niveau socio-économique, à l'environnement ou aux caractéristiques ethniques. Mais les membres d'une même famille, les amis proches sont plus enclins à partager la même exposition à un facteur de risque donné (tabac par exemple) que les sujets malades et **l'ampleur d'une association** entre le facteur de risque et la maladie peut donc être **sous-estimée**.

II.2.3.b- Nombre de groupes contrôles et nombre de contrôles par cas :

L'utilisation de groupes contrôles multiples peut permettre de pallier les biais provenant du choix d'un groupe contrôle donné. Si l'association entre facteur de risque et maladie est retrouvée que la comparaison soit faite avec un groupe de témoins hospitalisés, un groupe de témoins tirés au sort dans la population générale ou un groupe de témoins constitué de voisins directs, et si l'ampleur de l'association ne diffère pas significativement d'une comparaison à l'autre, il sera probable que l'association corresponde à une réalité et non pas à un artefact lié à la structure de l'étude. Si au contraire l'association diffère considérablement selon le groupe de comparaison utilisé, il se peut que la structure de l'étude soit questionnable et qu'il existe d'importants biais de sélection ou d'information au niveau d'un ou de plusieurs groupes contrôles expliquant les discordances.

Indépendamment du nombre de groupes contrôles, se pose la question du nombre de contrôles par cas.
Lorsque le coût d'obtention de l'information est comparable dans les deux groupes et que le nombre de cas et de contrôles est suffisamment important, le meilleur rapport est de 1/1. L'analyse statistique des données sera en effet plus facile, de même que les calculs de taille d'échantillon.
Lorsque par contre le nombre de cas que l'on peut inclure dans l'étude est faible, soit parce que le coût de la collection de l'information est élevé, soit parce que la maladie est rare, augmenter le nombre de contrôles par cas augmentera la puissance de l'étude et par conséquent les chances de mettre en évidence une association si elle existe.

L'accroissement en puissance est cependant faible lorsque l'on dépasse **4 contrôles pour un témoin**, et il est en général inutile d'aller au-delà de ce rapport.

II.2.4- Méthodes de sélection des cas et des contrôles. Méthodes d'échantillonnage :

Cas et contrôles proviennent d'une population de sujets malades et de sujets sains. Les sujets inclus dans l'étude doivent être représentatifs des populations dont ils proviennent. Ils doivent en constituer un **échantillon non biaisé**, c'est-à-dire que leur sélection, conditionnée par leur

statut vis-vis de la maladie étudiée, ne doit pas être influencée par leur exposition au facteur de risque suspecté ou par tout autre facteur pouvant jouer un rôle dans la physiopathologie de la maladie, sous peine d'induction d'un **biais de sélection** pouvant influencer les résultats (Cf. l'étude sur cancer du poumon et tabagisme conduite dans un service hospitalier pour les cas et les témoins).

II.2.4.a- Echantillonnage au hasard (random sampling) :

Cette procédure représente la situation où tous les sujets, cas ou contrôle, ont une probabilité égale d'être sélectionnés dans leur population d'origine par tirage au sort ou par l'utilisation de tables de nombres au hasard.

II.2.4.b- Echantillonnage systématique :

Cette procédure représente la situation où les sujets sont tous inclus systématiquement dans l'étude pendant une certaine période, et sont par conséquent représentatifs de tous les cas survenus pendant la période considérée.

II.2.4.c- Echantillonnage par stratification :

Les sujets sont sélectionnés au hasard au sein de groupes préalablement définis : population urbaine, population rurale, hommes, femmes... de façon à assurer un recrutement suffisant dans les différents groupes d'intérêt, permettant d'établir des conclusions valides au sein de chacun d'eux.

II.2.4.c- Echantillonnage avec appariement (matching) :

Chaque cas est apparié à un ou plusieurs contrôles **sur la base d'une variable que l'on veut éliminer de la comparaison**. Par exemple, on appariera par sexe ou par âge lorsque l'on étudie une maladie plus fréquente dans un sexe ou dans un groupe d'âge, de sorte que le cas et le contrôle soient comparables en regard de la variable âge ou sexe déterminante pour la survenue ou l'évolution de la maladie.
Le but intuitif de l'appariement est de contrôler un éventuel élément confondant (par exemple le sexe ou l'âge, comme facteurs de risque naturels d'un certain nombre de maladies). Ce but est pleinement réalisé dans les études en cohorte (qui incluent les patients sur la base des facteurs de risque, et en appariant sur un facteur de risque, l'éliminent de la comparaison).
En revanche dans les études cas-témoins, les cas comme les témoins ne doivent pas, sous peine de biais, être sélectionnés sur la base d'un facteur de risque, fût-il le facteur d'appariement, mais uniquement sur la base de l'absence ou de la présence de la maladie.

L'appariement dans une étude cas-témoins permet d'augmenter la puissance de l'étude en fournissant autant de témoins que de cas pour la caractéristique d'appariement (le sexe par exemple). Il ne permet pas le contrôle du facteur confondant correspondant, et peut induire un biais d'appariement[6].

Exemple : La maladie de Horton est une vascularite touchant les personnes âgées de plus de 50 ans, femmes dans 70 % des cas. La lésion anatomo-pathologique laisse supposer que l'athérosclérose peut faire le lit de la maladie. Une étude cas-témoins étudiant les facteurs de risque cardiovasculaire a été construite, avec appariement par âge et par sexe des témoins aux cas : cet appariement permettait de ne pas comparer un groupe de 70 % de

femmes et 30 % d'hommes âgés en moyenne de 75 ans, à un groupe tiré au sort dans la population générale, qui aurait été constitué de 50 % de femmes et 50 % d'hommes, d'un âge moyen de 50 ans. Or, les facteurs de risque cardiovasculaire, et tout particulièrement le tabagisme, sont âge et sexe dépendants. L'analyse de la série totale ne permettait de dégager qu'une tendance à la maladie chez les fumeurs. L'analyse séparée des hommes et des femmes mettait en évidence le tabagisme comme le facteur de risque le plus important chez les femmes de cette génération, alors que le risque disparaissait dilué dans l'omni-tabagisme (70 %) des hommes de cette génération sur le groupe complet de patients et de témoins[7]. *Dans une étude cas-témoins, l'appariement, en rapprochant trop les témoins des cas sur leur facteur d'exposition lié au facteur d'appariement (le tabac dans cette tranche d'âge est très lié au sexe, facteur d'appariement), peut induire un biais d'appariement amenant une diminution artéfactielle de l'odds ratio et son rapprochement de 1.*

II - ANALYSE DES DONNEES

Le calcul d'un risque est facile dans une étude de cohorte, puisque la cohorte a été formée sur la base de l'exposition au facteur de risque et que l'on examine ensuite la proportion de sujets présentant la maladie parmi le groupe des personnes exposées et non exposées.

Si cette proportion est égale à X % parmi les sujets exposés, et Y % parmi les sujets non exposés, le rapport X / Y donne le risque relatif des personnes exposées par rapport aux personnes non exposées (fig. 3).

Dans une étude cas-témoins, basée sur la connaissance du statut des sujets vis-à-vis de la maladie et non pas de l'exposition, il n'est pas possible de déterminer quelle proportion de sujets exposés développera la maladie car, même si tous les cas dans une population donnée étaient enregistrés, la proportion de gens exposés dans la population de référence n'est pas connue. On détermine donc la proportion de sujets présentant l'exposition au facteur de risque parmi le groupe de sujets atteints d'une part, et parmi le groupe de sujets sains d'autre part (fig. 4). Il s'agit de la démarche inverse de celle suivie dans une étude de cohorte.

L'estimation du risque relatif peut cependant se faire sous certaines hypothèses, grâce au calcul du rapport de cotes (odds ratio) (tableau 1).

Cote (odd) d'exposition parmi les cas:

$$\frac{\text{Proportion de cas exposés}}{\text{Proportion de cas non exposés}} = \frac{a / (a+c)}{c / (a+c)} = \frac{a}{c}$$

Cote (odd) d'exposition parmi les témoins:

$$\frac{\text{Proportion de témoins exposés}}{\text{Proportion de témoins non exposés}} = \frac{b / (b+d)}{d / (b+d)} = \frac{b}{d}$$

$$\text{Rapport de cotes} = \frac{\text{Cote d'exposition parmi les cas}}{\text{Cote d'exposition parmi les témoins}} = \frac{a / c}{b / d} = \frac{ad}{bc}$$

Le rapport de cotes exprime le risque relatif lorsque la maladie est rare, et se calcule très simplement par le rapport des produits en croix de la table 2 × 2 (tableau 1).

La transformation logarithmique du rapport de cotes donne les coefficients des différents paramètres des variables dans une équation de régression logistique. Une équation de régression logistique peut donc intégrer de multiples variables représentant de multiples facteurs de risque étudiés dans une étude cas-témoins, et déterminer quels sont leur rôle et leur importance respectifs dans le risque de développer la maladie.

III - INTERPRETATION DES DONNEES

Nous avons vu qu'une attention toute particulière devait être portée aux biais potentiels lors de l'élaboration d'une étude cas-témoins. Ce type d'étude est, en effet, par sa structure, plus sujet à l'existence de multiples biais qu'une étude de cohorte et, a fortiori, une étude randomisée.

Les biais de sélection surviennent lorsque l'inclusion des cas ou des témoins dépend de quelque façon de l'exposition que l'on se propose d'étudier. La sélection des patients ne se fait plus alors uniquement sur la maladie, mais est influencée également par l'exposition au facteur de risque, qui dès lors ne peut plus être étudié sans erreur. Une étude de l'association entre bronchite chronique et tabac où tous les patients seraient recrutés à la consultation anti-tabac mettrait en évidence une association plus forte que l'association réelle (puisque tous les patients sont tabagiques) et échouerait sans doute dans la mise en évidence d'autres facteurs de risque (climatiques, antécédents d'asthme chronique, certaines maladies professionnelles…).

Les biais de détection sont les biais relatifs aux anomalies de diagnostic de la maladie, et peuvent être de nature multiple : la maladie peut ne pas être détectée parce qu'elle évolue longtemps à un stade infra-clinique, parce que les tests diagnostiques dont nous disposons ne sont pas assez sensibles, parce que les critères diagnostiques choisis pour l'étude sont douteux. Au contraire, elle peut être trop facilement détectée à un stade asymptomatique, alors qu'elle n'aurait peut-être jamais fait parler d'elle, et le groupe de patients s'enrichira alors de patients à facteurs de risque et facteurs pronostiques potentiellement différents de ceux des patients symptomatiques (exemple : leucémie lymphoïde chronique de sujet âgé dépistée sur une numération formule sanguine systématique ; myélome de stade 1 dépisté sur une électrophorèse des protéines de routine, etc). Les patients non détectés peuvent être différents des patients détectés, et le facteur de risque isolé peut alors ne pas être généralisable à l'ensemble des patients pour lesquels un jour, la maladie sera diagnostiquée. Si l'on utilise la radiographie pulmonaire annuelle pour le dépistage du cancer du poumon, détecte-t-on préférentiellement les cas d'évolution lente, de meilleur pronostic que les cas pouvant apparaître et devenir symptomatiques en moins de 12 mois ?

Les biais d'observation se subdivisent en biais de souvenir et biais de mauvaise classification:

- Les biais de souvenir sont tout particulièrement fréquents dans les études cas-témoins, car l'exposition au facteur de risque est déterminée *a posteriori*. Le patient ayant réfléchi sur sa maladie est plus à même de se rappeler tel ou tel facteur de risque que le sujet sain du groupe témoin. Si les souvenirs ne sont pas de qualité équivalente dans les deux groupes, la comparaison est obligatoirement faussée et a pour conséquence une valeur faussée, généralement exagérée, du rapport de cotes (odds ratio). Certaines maladies au contraire peuvent entraîner des troubles de la mémoire ou du comportement, et la collecte des données est alors moins bonne dans le groupe malade. Le rapport de cotes en est faussement abaissé.

- Les biais de mauvaise classification surviennent lorsque le statut d'exposition ou de la maladie a été faussement rapporté. Des sujets sains peuvent alors être classés dans le groupe des sujets malades, et réciproquement. Des sujets non exposés peuvent se trouver classés parmi les sujets exposés. Ce "mélange" des groupes aboutit à une dilution des causes et des effets dans chaque groupe étudié. Si cela s'est opéré sans direction préférentielle, on assiste à un affaiblissement du rapport de cotes qui se rapproche de 1.

Si la confusion s'est toujours opérée dans le même sens et si par exemple tous les sujets exposés se trouvent classés par erreur parmi les sujets malades, le rapport de cotes est anormalement augmenté. A l'inverse le rapport de cotes est anormalement abaissé en cas de mauvaise classification unidirectionnelle lorsque les sujets exposés sont préférentiellement classés parmi les sujets témoins. On peut même dans les cas extrêmes arriver à une inversion de l'association réelle !

L'appariement peut lui-même être à l'origine de nouveaux biais dans une étude cas-témoins. Ceci survient lorsque le facteur d'appariement est associé avec le facteur d'exposition que l'on veut étudier, sans qu'il existe de relation de cause à effet entre eux. Il convient alors, dans l'analyse des données, de prendre en compte le facteur d'appariement et de calculer les rapports de cotes de façon séparée dans chaque sous-groupe (par exemple, dans le groupe masculin et dans le groupe féminin si le sexe était le facteur d'appariement).

IV - AVANTAGES ET FAIBLESSES DES ETUDES CAS-TEMOINS

A - Avantages

- Elles sont de réalisation rapide et peu coûteuse comparativement aux études de cohorte.

- Elles sont particulièrement adaptées à l'étude de maladies ayant une longue période de latence.

- Elles sont particulièrement adaptées à l'étude de maladies rares.

- Elles peuvent examiner plusieurs facteurs de risque d'une même maladie.

B - Faiblesses

- Elles sont peu rentables pour l'évaluation de facteurs de risque rares, sauf si le risque est très élevé.

- Elles ne peuvent pas calculer directement l'incidence de la maladie dans les populations exposées et non exposées, sauf si l'ensemble des cas de la population considérée sont enregistrés.

- La relation de cause à effet et la séquence temporelle entre facteur de risque présumé et maladie sont parfois difficiles à établir, car les données concernant l'exposition sont collectées en même temps que les données concernant la maladie.

- Elles sont particulièrement sujettes aux biais (biais de sélection et biais de souvenir essentiellement).

Fig. 1 - Structure d'une étude cas-témoins

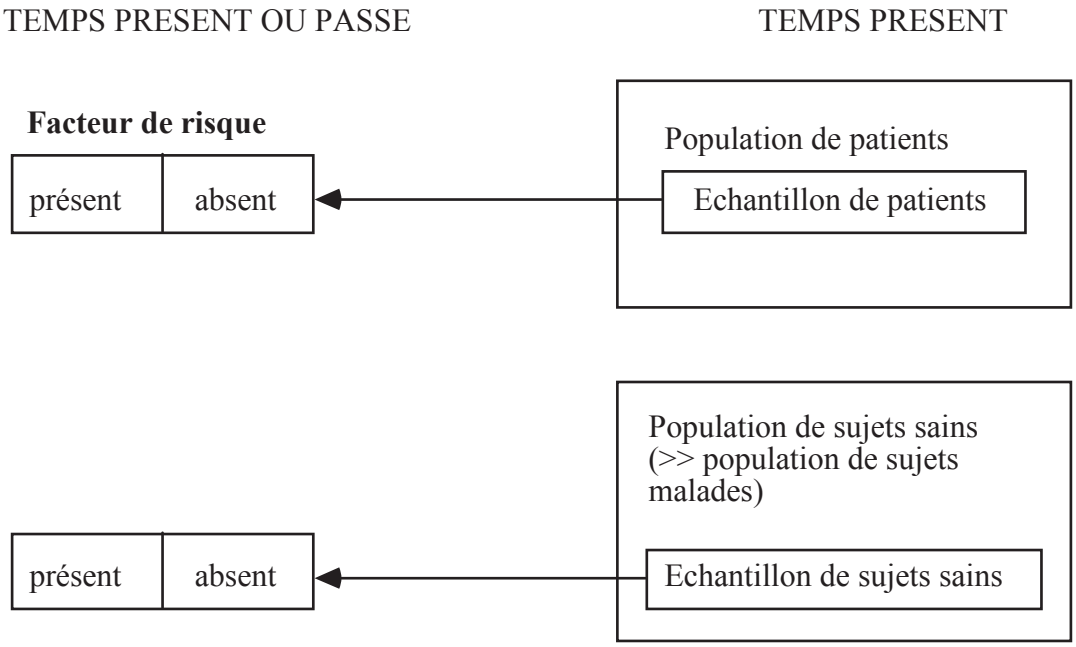
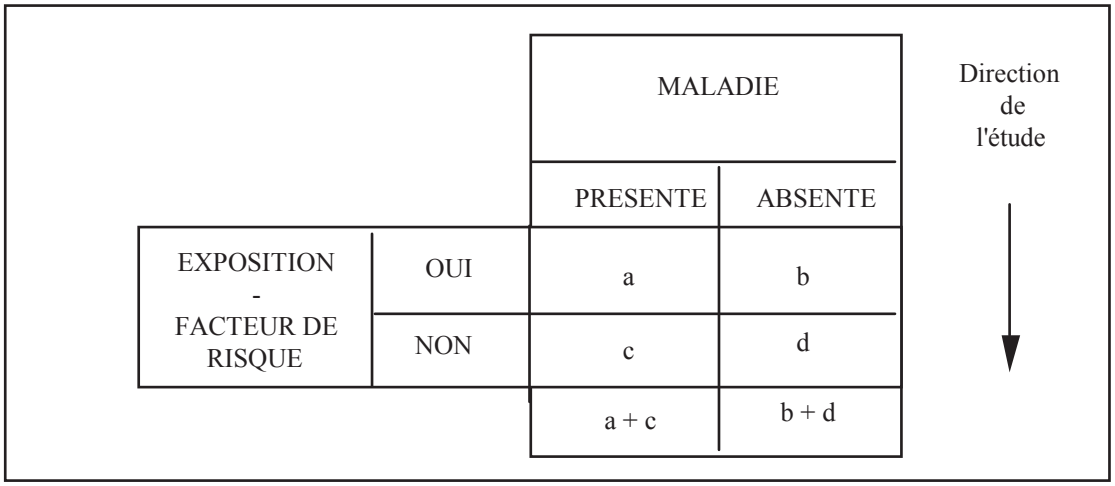


Fig. 2 - Représentation d'une étude cas-témoins par la table 2×2 : (a + c) cas et (b + d) témoins sont définis au début de l'étude



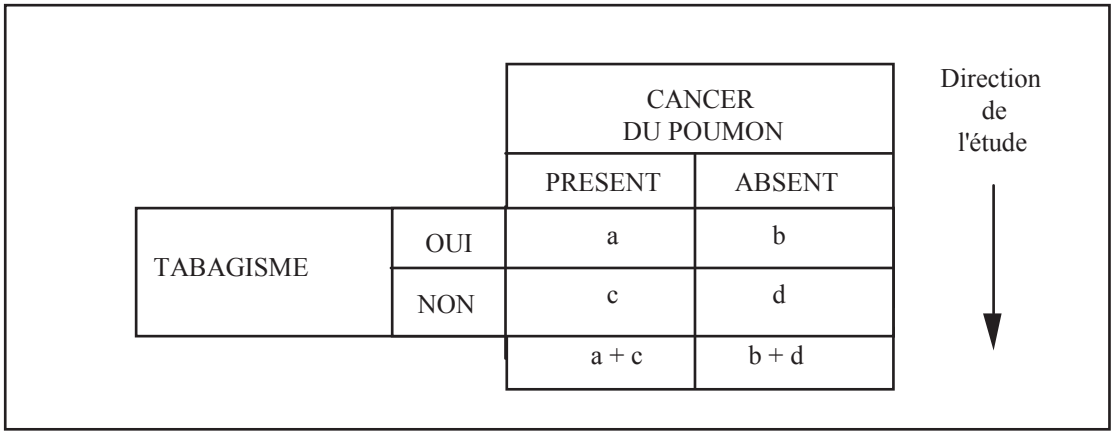


Fig. 3 -

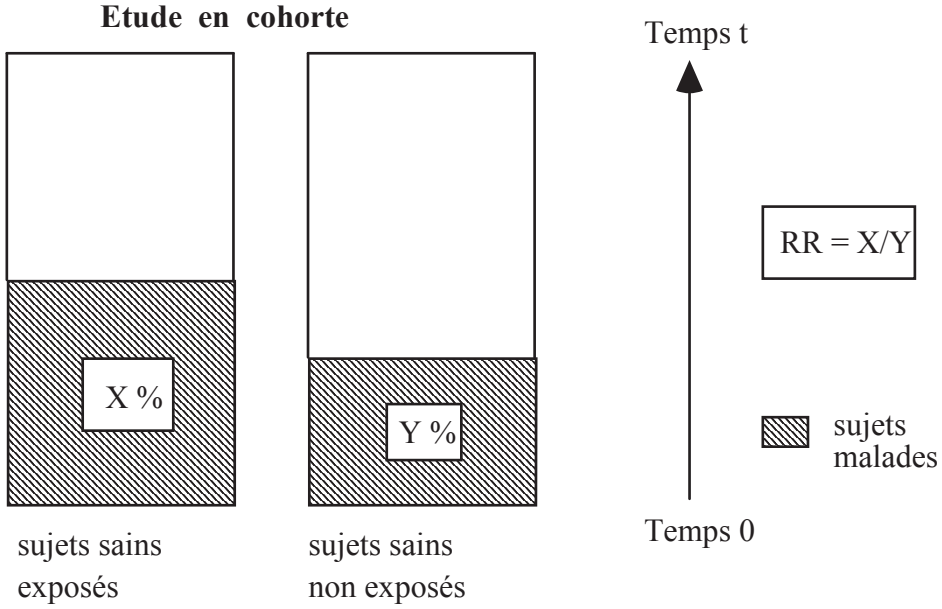

Etude en cohorte
X %
Y %
sujets sains
exposés
sujets sains
non exposés
Temps t
Temps 0
RR = X/Y
sujets
malades

Fig. 4 -

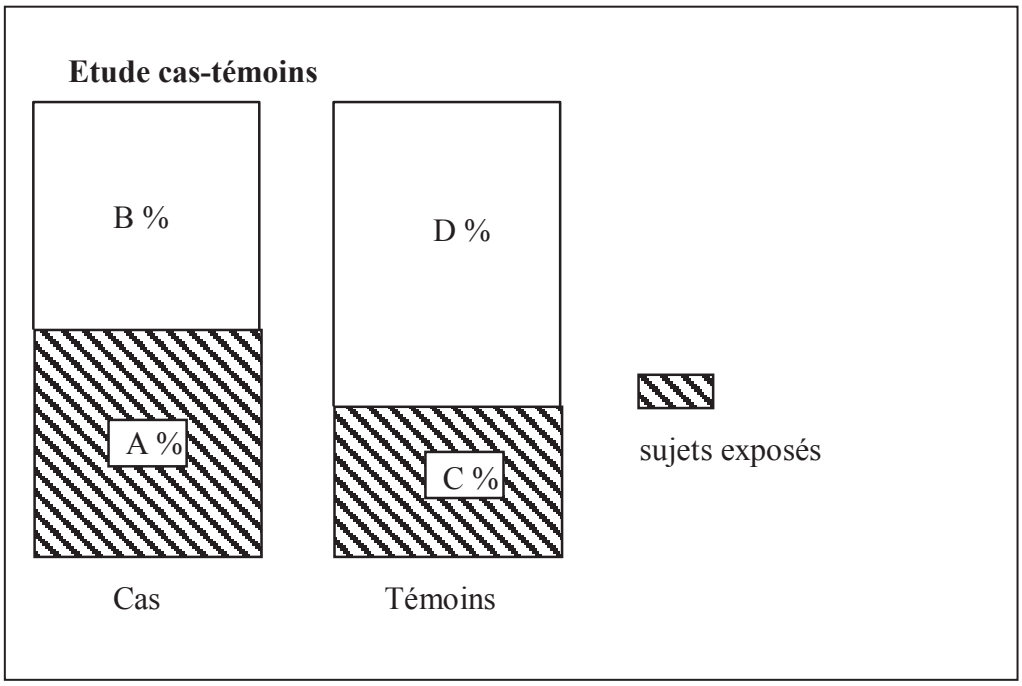


Tableau 1 - Calcul du rapport de cotes dans une étude cas-témoins; lecture verticale de la table 2×2

	Cas	Témoins	
Exposés	a	b	a + b
Non exposés	c	d	c + d
	a + c	b + d	a + b + c + d

Références

[1] - Barrier J, Pion P, Massari R, Peltier P, Rojouan J, Grolleau JY. Approche épidémiologique de la maladie de Horton dans le département de Loire -Atlantique. 110 cas en 10 ans (1970-1979). Rev Med Int 1982;3:13-20.

[2] - Seeff LB, Buskell-Bales Z, Wright EC, Durako SJ, Alter HJ, Iber FL, et al. Long-term mortality after transfusion-associated non-A, non-B hepatitis. NEJM 1992;327:1906-1911.

[3]- Smith DC, Prentice R, Thompson DJ, Herrmann WL. Association of exogenous estrogen and endometrial carcinoma. NEJM 1975;293:1164-1167.

[4] - Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. NEJM 1978;299:1089-1094.

[5] - Hulka BS, Grimson RC, Greenberg BG, Kaufman DG, Fowler WC Jr, Hogue CJ, Berger GS, Pulliam CC. "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen. Am J Epidemiol 1980;112:376-87.

[6] - Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third Edition, 2008. Lippincott, Williams and Wilkins Editions.

[7] - Duhaut P, Pinède L, Demolombe-Rague S, Loire R, Seydoux D, Ninet J, Pasquier J. Giant cell arteritis and cardiovascular risk factors : a multicenter, prospective case-control study. GRACG. Arthritis Rheum 1998;41:1960-1965.

CHAPITRE VI

LES ETUDES DE COHORTE

Pierre Duhaut, Jean Schmidt

L'étude cas-témoins nous a donné la possibilité de comparer des sujets sains et des sujets malades afin d'identifier un éventuel lien entre la maladie et un facteur de risque supposé. Cependant, les contraintes méthodologiques sont lourdes, à la hauteur de la sophistication de l'étude, et les pièges sont nombreux. La liste des biais est longue … Mais ces contraintes sont acceptables en regard de la légèreté relative de la logistique. Ce type d'étude convient donc aux maladies rares.

Malheureusement un renseignement d'importance fait défaut : l'incidence de la maladie. Par rapport à la prévalence, l'incidence a une dimension supplémentaire, car elle prend en compte le temps. Cela intéresse au plus haut point le clinicien. Cette notion de temps permet d'estimer la probabilité d'un sujet de rencontrer l'événement en question dans le futur. Cela est impossible avec une étude cas-témoins puisque l'on part d'un groupe formé artificiellement de sujets présentant l'événement digne d'intérêt. Un nouveau type d'étude est nécessaire.

Comme pour l'étude cas-témoins, ce type d'étude concerne le risque, qui fournit des informations importantes au clinicien - et au patient - en termes de prédiction, de causalité, de diagnostic et de prévention. Il s'agit de l'étude de cohorte, au nom suggestif puisqu'une cohorte représentait chez les Romains le dixième d'une légion, soit 600 hommes, et que ce terme se prête particulièrement bien à l'idée de suivi d'un groupe important d'hommes en marche vers un futur (incertain).

L'étude cas-témoins était le plus souvent rétrospective, l'étude de cohorte est le plus souvent prospective. Toutes deux sont des études d'observation, analytiques, non expérimentales.

Enfin, l'étude de cohorte est la seule qui permette d'appréhender la survie, et les facteurs pronostiques.

Plan du chapitre

B - Etudes de cohorte rétrospective ou historique : les registres

C- Etudes de cohorte de patients : notion de survie et de facteurs pronostiques.

D - Les études cas-témoins insérées dans les études de cohorte

IV - EXPRESSION DES RESULTATS

V - AVANTAGES ET FAIBLESSES DES ETUDES DE COHORTE

A - Avantages

B - Faiblesses et inconvénients

Les études de cohorte représentent la forme la plus rigoureuse des études épidémiologiques non expérimentales. Elles seules permettent d'évaluer l'incidence d'une maladie, d'établir une relation de cause à effet entre facteur de risque et maladie avec le moins de biais possibles, d'évaluer le temps de latence et le risque relatif avec une précision maximale.

Toutefois, la précision, et la validité des informations fournies sont obtenues au prix d'un investissement souvent considérable en temps et en moyens, et les études de cohorte, comme toutes les études épidémiologiques, sont soumises à des travers spécifiques qui peuvent en entacher la validité.

L'incidence indique le pourcentage de nouveaux cas diagnostiqués au cours d'une période de temps dans une population donnée:

$$\text{Incidence} = \frac{\text{Nouveaux cas diagnostiqués durant une période p}}{\text{Population totale considérée}}$$

I - STRUCTURE ET MISE EN PLACE DE L'ETUDE

Par opposition aux études cas-témoins, la sélection des sujets ne se fait plus sur la présence ou l'absence de la maladie, mais sur la présence ou l'absence du facteur de risque supposé. Les sujets sont suivis ensuite au cours du temps, et la survenue d'événements incidents enregistrée (figures 1 et 2).

Les étapes sont les suivantes:

- Définition précise de la question posée et de la population que l'on veut étudier (population générale, population vivant dans un certain secteur géographique, employés d'une entreprise, profession particulière, catégorie sociale spécifique…).

- Si l'ensemble de la population ne peut pas être étudié, sélection d'un échantillon représentatif, au sein de cette population. Cet échantillon, pour être représentatif, doit être tiré au sort, et chaque individu doit avoir une chance égale d'être sélectionné. Il faut donc disposer, comme pour les études de prévalence, d'une liste précise de la population étudiée.

- Définition précise et mesure du facteur de risque supposé.

- Définition précise des événements incidents que l'on veut collecter. Définition des moyens de diagnostic mis en oeuvre.

- Définition de la période de temps pendant laquelle sont suivis les sujets inclus. Cette durée dépend du facteur de risque, du temps de latence et de la maladie étudiés, et peut dans certains cas être très longue (par exemple, maladies cardio-vasculaires). La cohorte de Framingham est ainsi suivie depuis plus de 40 ans.

- Enoncé des mesures prises afin d'éviter la perte de suivi des sujets au cours du temps, principal écueil des études prospectives.

La définition de l'exposition et de la maladie, et par conséquent la reconnaissance des sujets exposés, non exposés, malades et non malades, pose le même type de problèmes que dans les études transversales ou les études cas-témoins.

Sensibilité, spécificité, reproductibilité, valeur prédictive positive et négative des critères diagnostiques doivent être testées et mesurées. Ceci est d'autant plus critique qu'une étude de cohorte est en règle générale longue, coûteuse, lourde à conduire, et qu'il est nécessaire d'optimiser ses chances de succès.

Toutes ces notions, exposées dans les chapitres 'Etude de prévalence', 'Etudes cas-témoins', 'Evaluation d'un test diagnostique' s'appliquent aux premières étapes de la construction d'une étude de cohorte, avec un soin tout particulier compte tenu de la lourdeur de l'étude que l'on va entreprendre : il s'agit des fondations d'une 'cathédrale épidémiologique', que l'on veut solide dans la durée.

Les sujets étant classés sur la base de leur exposition, il faut inclure un grand nombre de personnes à la phase initiale pour avoir, à la fin de l'étude, un nombre suffisant de cas ayant contracté la maladie en cours de suivi pour mettre en évidence un effet, s'il existe.

Les études de cohorte sont donc des études intéressant un échantillon très large de la population. Suivre de façon régulière 100.000 personnes pendant plusieurs années représente une tâche d'une ampleur certaine … !

Exemple : Sir Richard Doll en Grande-Bretagne a débuté juste après la deuxième guerre mondiale une étude de cohorte s'intéressant aux méfaits du tabagisme. Il a inclus 100 000 médecins et les a suivis au long cours… c'est-à-dire, jusqu'à nos jours, ce qui représente un suivi de plus de 50 ans ! Il est le premier à avoir mis en évidence l'augmentation importante du risque de cancer du poumon chez les fumeurs, avec un risque relatif à 6 par rapport aux non-fumeurs. Cette cohorte bien sûr, lui a également permis d'étudier la survenue d'autres pathologies, mais aussi, comme nous le verrons, les complications d'autres facteurs de risque [1, 2].

II - BIAIS POTENTIELS

Malgré leur force apparente, les études de cohorte sont, elles aussi, soumises à la possibilité de biais pouvant en modifier les résultats, présents à chaque étape de l'étude.

A - Biais de sélection

Les sujets inclus dans l'étude sont-ils vraiment représentatifs de la population dont ils sont issus ?

Exemple : Le taux de réponse aux différents questionnaires de suivi envoyés aux 100 000 médecins de la cohorte de Richard Doll a toujours été voisin de 95 %. En 1991, l'auteur s'est intéressé à la consommation d'alcool dans sa cohorte et à ses effets potentiels. Le taux de réponse est immédiatement tombé à 73 % [3]. Lors du questionnaire suivant s'intéressant à nouveau aux cancers, pathologies cardiovasculaires et pulmonaires, le taux de réponse est remonté vers ses valeurs habituelles [4]. Que s'est-il passé ? Qui sont les 27 % n'ayant pas répondu à une question sensible, alors qu'ils répondaient aux questions plus neutres sur le plan social et individuel ? On peut sans doute penser que les non-buveurs ont bien répondu, ainsi que les consommateurs -très- modérés. Les consommateurs invétérés ont probablement jugé plus prudent de ne pas répondre, tout en restant fidèles à la cohorte par ailleurs, comme l'ont montré les questionnaires suivants !

Il existe donc très certainement un biais de sélection, et la cohorte finalement sélectionnée de facto par les non-réponses n'est certainement pas représentative de la population que l'on se proposait d'étudier.

B - Biais de mauvaise classification

Il existe lorsque les sujets n'ont pas été classés dans la catégorie (exposés, non exposés, malades, non malades) qui leur correspond. Il peut porter sur l'exposition ou sur la maladie survenue en cours de suivi.

Exemples :
- Dans l'étude de Richard Doll portant sur la consommation alcoolique, il est probable que les consommateurs -peu- modérés aient… modéré leur réponse, du moins dans l'appréciation quantitative de l'alcool ingéré et que certains gros consommateurs aient tout simplement répondu qu'ils ne buvaient… strictement rien. Ceci induit un biais de mauvaise classification d'exposition, qui retentira sur le calcul du risque relatif de survenue des pathologies compliquant la consommation d'alcool [3].
- Le diagnostic anatomo-pathologique est souvent considéré comme l'étalon or des diagnostics. Pourtant, l'analyse de la lésion peut être difficile, y compris pour différencier pathologie maligne, dysplasique, ou bénigne. La lecture de la biopsie d'artère temporale pour le diagnostic de maladie de Horton est censée être facile. Pourtant, le coefficient kappa de reproductibilité entre deux anatomo-pathologistes expérimentés ne dépasse pas 86 % pour la reconnaissance d'une biopsie positive versus négative, et 70 % pour la reconnaissance des cellules géantes caractéristiques de la lésion et de la maladie, ce qui laisse un pourcentage important de patients potentiellement mal classés [5].

C - Perte de suivi

Comment faire pour suivre l'ensemble de la population initiale sur plusieurs années ? Des personnes déménageront sans laisser d'adresse. D'autres ne répondront plus aux questionnaires ou ne se présenteront plus aux visites de contrôle.

Ces perdus de vue ne sont pas obligatoirement répartis de façon aléatoire parmi les sujets initialement inclus, mais peuvent appartenir plus particulièrement à un sous-groupe (par exemple, patients exposés).

Si tel est le cas, les conclusions de l'étude sont faussées quelles que soient les précautions que l'on a prises aux étapes précédentes. Si le taux de perdus de vue dépasse 20 %, l'étude est obligatoirement critiquable. Un taux de 15 % amène probablement un certain nombre de discussions.

Exemple : 27 % de sujets interrogés au début n'ont pas répondu au questionnaire 'Alcool', et nous pouvons imaginer qu'ils se comptent surtout parmi les "grands buveurs" (hypothèse très raisonnable). Ils sont de fait, 'perdus de vue' pour tout ce qui concerne la mesure de l'incidence des pathologies liés à l'alcool, même s'ils répondront ensuite aux questionnaires portant sur des pathologies différentes (n'oublions pas qu'il s'agit de médecins, connaissant les facteurs de risque et leurs conséquences). Comment interpréter dans ces conditions les résultats d'une étude par ailleurs bien conduite ? Etait-il possible d'éviter les pièges et quels garde-fous aurait-il fallu prévoir avant de lancer cette partie de l'étude ? Les réponses ne sont pas faciles à apporter dès lors que l'on s'intéresse à des facteurs de risque ou des pathologies marquées d'un certain stigma social.

D - Facteurs de confusion (fig. 3)

Comme dans les études cas-témoins ou les études transversales, les facteurs de confusion sont les facteurs associés à la fois au facteur de risque supposé et à la maladie étudiée. Ils causent la maladie, mais ne sont pas apparents dans l'étude et contribuent à attribuer faussement la causalité de la maladie au facteur de risque supposé qui, lui, est mesuré.

> Exemple : dans l'étude de Richard Doll toujours, les cirrhoses hépatiques bien sûr, mais aussi les hépatocarcinomes, les cancers des voies aéro-digestives supérieures, les accidents et traumatismes surviennent plus facilement chez les grands consommateurs d'alcool. Est-ce à dire que l'alcool en tant que molécule est réellement le facteur causal de la maladie ? A moins que ça ne soient la dénutrition, souvent associée à l'alcoolisme, ou les carences multiples secondaires à la dénutrition, ou encore les troubles psycho-pathologiques dont l'alcoolisme n'est que le reflet ou la conséquence, ou enfin le facteur génétique prédisposant à l'alcoolisme (sous réserve de son existence) qui, à des degrés divers, soient à l'origine de la complication observée ?

Si les facteurs de confusion potentiels n'ont pas été prévus et mesurés de pair avec le facteur de risque supposé, comment apprécier leur impact et leur rôle dans la genèse de la maladie ? Comment les contrôler ?

N'existe-t-il pas souvent des facteurs de confusion ignorés, et donc impossible à prévoir ? Dans la figure 3, le facteur B est un facteur de confusion, non mesuré dans l'étude, mais expliquant en fait l'association positive observée entre le facteur A et la maladie X, car il existe une association fortuite entre A et B. L'apparente association entre A et X est donc expliquée par un biais de confusion.

Une solution est d'essayer de contrôler, dans chaque étude, tous les facteurs de confusion prévisibles en les mesurant et en tenant compte de leur présence dans l'analyse des données (stratification de la population étudiée en fonction des facteurs de confusion, intégration de ces facteurs dans une équation de régression logistique ou linéaire). Une autre solution est de répéter les études dans des populations différentes, avec des investigateurs différents. La concordance des résultats d'études épidémiologiques avec les résultats d'études biologiques (plausibilité des résultats) représente enfin un troisième argument en faveur de leur validité.

III - VARIATIONS DE STRUCTURE

A - Etudes de double cohorte

Lorsque le facteur de risque est rare, même l'inclusion d'un échantillon très large de la population peut ne pas être suffisante pour comporter le nombre nécessaire de sujets exposés. Il faut alors avoir recours à une double cohorte (fig. 4) et effectuer une double sélection. Un premier échantillonnage est réalisé au sein de la population exposée, et un second au sein de la population non exposée. Deux cohortes au lieu d'une sont alors suivies de façon parallèle.

Les avantages de cette double cohorte sont évidents : le nombre de sujets exposés est tel qu'il garantit un nombre suffisant de sujets malades si le facteur d'exposition est véritablement un facteur de risque. On augmente ainsi les chances de tirer des conclusions informatives.

Il existe cependant des inconvénients. Le principal est que les sujets exposés et non exposés ne sont plus issus de la même population, et peuvent par conséquent différer par d'autres

facteurs que le seul facteur d'exposition. Ceci accroît le risque de facteurs de confusion, qui, s'ils sont méconnus et non contrôlés, peuvent conduire à des conclusions fausses.

Imaginons une étude de cohorte visant à mettre en évidence le rôle carcinogène de l'amiante, et comparant deux cohortes assemblées séparément. La première est constituée de mineurs travaillant à extraire l'amiante, la seconde d'employés d'une grosse entreprise de construction électrique. Ces deux cohortes sont-elles semblables à l'égard d'autres carcinogènes, tels que le tabac, l'association tabac-alcool, la consommation de graisses animales, … ? A l'opposé, n'y a-t-il pas chez les employés de l'entreprise de construction électrique un certain degré d'exposition à l'amiante par le biais de la manipulation d'isolants, qui risque de sous-estimer le risque relatif ? Quels sont les facteurs de différence entre les deux cohortes non directement visibles, susceptibles de fausser les résultats ?

Pour éviter ce biais et les facteurs de confusion qui en résultent, il reste la possibilité de réaliser une étude de cohorte unique. En pratique cependant, combien de personnes de la population générale faudra-t-il inclure pour avoir suffisamment de personnes exposées ? Cela est-il réalisable et à quel coût ?

Les études en double cohorte posent donc un problème similaire à celui rencontré dans les études cas-témoins: celui de la similarité ou des différences non prévues existant entre les deux groupes, à l'origine de la présence de facteurs de confusion éventuels.

B - Etudes de cohorte rétrospective ou historique : les registres.

Nous avons examiné jusqu'à présent la structure des études de cohorte prospective. Il est possible en fait de réaliser, aussi paradoxal que cela puisse paraître, une étude de cohorte rétrospective.

Il faut pour que la structure de la cohorte soit respectée et que l'exposition soit mesurée avant la survenue de la maladie, que l'exposition ait été déterminée dans le passé de façon précise (et non pas reconstituée dans le présent à partir des souvenirs des sujets) et que l'ensemble des sujets qui y ont été soumis (population d'étude) soient identifiés ou identifiables.

En pratique, il s'agit souvent d'une cohorte déjà formée lors d'une autre étude, pour laquelle les informations initialement recueillies comprenaient les données sur le facteur de risque que l'on veut examiner *a posteriori*. Il peut s'agir également d'une population parfaitement définie à l'origine, soumise à un risque particulier, correctement suivie et pour laquelle tous les événements incidents ont été enregistrés. Les meilleurs exemples de ce type sont représentés par les registres scandinaves, ayant enregistré de manière systématique depuis des décennies les facteurs de risque, les pathologies incidentes, les résultats d'examens complémentaires, et les traitements suivis par l'ensemble de la population des pays considérés (Norvège, Suède, Danemark). Ces registres permettent d'obtenir rapidement la réponse à une question posée au sein de cohortes historiques, pour peu que l'information nécessaire ait été enregistrée initialement.

Exemples:

- Les conséquences à long terme de pathologies gestationnelles (hypertension gravidique, pré-éclampsie, éclampsie) pouvant interférer avec la croissance et la maturation fœtale sont mal connues, essentiellement parce que les temps de latence peuvent être extraordinairement longs pour certaines conséquences : si la mortalité

périnatale ou la prévalence de handicap psychomoteur grave est facile à mesurer, les conséquences plus fines et lointaines en termes d'apprentissage dans l'enfance, l'adolescence ou la vie adulte sont plus difficiles à appréhender et nécessitent des années de suivi. En revanche, il a été possible de les évaluer grâce au Registre Médical Danois des Naissances, au Registre Danois National des Patients, et aux mesures de fonction cognitive effectuées chez les conscrits au moment de leur incorporation. L'étude a mis en évidence, chez des sujets normaux par ailleurs, une légère diminution moyenne du QI chez les jeunes gens ayant été soumis à ces facteurs de risque néonataux, et un pourcentage légèrement plus important, mais significatif, de sujets à QI bas [6].

- La réanimation néo-natale a fait de grands progrès durant les 30 dernières années, et des enfants à score d'Apgar bas à la naissance ont pu être réanimés et rendus à une vie normale. Le score d'Apgar cependant traduit une souffrance importante pendant l'accouchement, et les conséquences à long terme, pour les mêmes raisons que celles citées précédemment, ne sont pas faciles à évaluer. La même équipe s'est intéressée à cette question sur les registres danois, pour arriver à la conclusion sur une cohorte rétrospective avec suivi de 20 ans, qu'un score d'Apgar inférieur à 7 à la naissance multipliait par 4 le risque de handicap neurologique à long terme, et par 1,33 le risque de diminution du quotient intellectuel, tout en soulignant que plus de 90 % des ces enfants à 20 ans allaient bien et présentaient des caractéristiques similaires à celles de la population générale de même âge [7].

Dans la cohorte rétrospective ou historique, la structure même de l'étude de cohorte est conservée : l'exposition a bien été déterminée au préalable, est bien connue, n'est pas soumise à d'éventuels biais de souvenir. Les événements incidents ont ensuite été enregistrés, et le problème de la séquence temporelle exposition-maladie résultante ne se pose pas. La population était bien identifiée au départ, et il est ainsi possible de déterminer une incidence vraie.

Les études de cohorte rétrospective ou historique peuvent être conduites rapidement, à moindre coût, et permettent de cumuler les avantages respectifs des études de cohorte (rigueur, mesure d'une incidence, séquence temporelle établie, …) et les avantages d'un étude cas-témoins rétrospective (résultats rapides, …), en évitant les écueils majeurs de l'un ou l'autre type.

Les études de cohorte rétrospective, cependant, souffrent des mêmes biais que les études de cohorte prospective, et souvent les majorent. La cohorte sur laquelle elles viennent se greffer n'a pas été conçue pour elles :

- les facteurs de confusion éventuels du nouveau facteur de risque examiné n'ont pas obligatoirement été enregistrés, et par conséquent ne peuvent pas toujours être contrôlés ;

- les biais de sélection peuvent être différents pour le facteur de risque de l'étude rétrospective, qui ne bénéficie pas alors des "garde-fous" mis en place pour limiter les biais de sélection dans l'étude prospective initiale;

- les maladies auxquelles on s'intéresse dans l'étude rétrospective peuvent être différentes de celles examinées dans l'étude prospective, et avoir été enregistrées dans les dossiers avec moins de rigueur ou d'exactitude. On accentue ainsi les biais de mauvaise classification. La structure mise en place pour assurer le meilleur suivi possible des patients et retrouver les

personnes perdues de vue peut être dépendante des maladies sur lesquelles portait l'étude prospective, et être moins adaptée au sujet de l'étude rétrospective conçue secondairement.

L'investigateur de l'étude rétrospective n'a pas le même contrôle de la qualité de l'information recueillie que l'investigateur de l'étude prospective. Il travaille sur des données déjà collectées pour un autre dessein, ou sur des données générales colligées sans a priori de leur utilisation future.

C- Les études de cohorte de patients : étude des facteurs pronostiques et de la survie.

Les études de cohorte peuvent partir de sujets sains exposés ou non exposés, et mesurer l'incidence des événements pathologiques consécutifs, mais aussi partir de sujets malades à partir du diagnostic ou de tout autre événement marquant dans l'évolution de la maladie, et mesurer les événements en découlant, de la guérison, à l'amélioration, aux complications naturelles ou iatrogènes, à l'aggravation jusqu'au décès en relation, ou non, avec la maladie causale. Ces études de cohorte de patients sont construites sur le même modèle que les études de cohorte de sujets sains. Elles permettent par contre d'analyser les facteurs pronostiques, qui peuvent être différents des facteurs de risque.

Exemple : l'œstrogénothérapie post-ménopausique est un facteur de risque de cancer de l'endomètre. Par contre, une fois le cancer déclaré, il s'agit plutôt d'un facteur de bon pronostic : les cancers œstrogèno-induits évoluent mieux que les cancers non induits.

Elles permettent de prendre en compte les facteurs confondants, potentiellement multiples, dans l'évolution des patients. Ainsi, certains facteurs de risque de la maladie étudiée peuvent aussi être des facteurs de risque d'événements survenant pendant l'évolution de la maladie.

Exemple : la maladie de Horton est une vascularite touchant les vaisseaux de moyen calibre, et des accidents vasculaires centraux (AVC) ainsi que des infarctus du myocarde (IMy) sur coronarite inflammatoire ont été rapportés durant son évolution. La maladie de Horton cependant est favorisée par le tabagisme, ainsi que les AVC et IMy athéromateux non inflammatoires. Une étude en double cohorte comparant patients et témoins appariés par âge et par sexe tirés au sort dans la population générale a pu mettre en évidence une augmentation du risque d'AVC et d'IMy durant les deux premières années chez les cas, en relation cependant avec le tabagisme et les facteurs de risque cardiovasculaires classiques en analyse multivariée, plutôt qu'avec l'atteinte inflammatoire elle-même. La distinction n'est pas que théorique, car les thérapeutiques de l'atteinte inflammatoire (corticoïdes) et de la prévention des complications de l'athérome sont radicalement différentes [8].

Certaines de ces études de cohorte de patients ne possèdent pas de groupe témoin permettant d'établir une comparaison de survie : elles peuvent alors faire appel à des statistiques obtenues dans la population générale (qui en France sont fournies par l'INSEE), pour comparer la survie des patients avec une survie théorique, attendue dans la population générale. Bien sûr, cette technique de comparaison ne permet pas le contrôle d'éventuels facteurs confondants ou une analyse fine du pronostic, puisqu'il n'y a pas de données personnelles recueillies dans le groupe de comparaison. Elle permet cependant d'analyser si le groupe de patients évolue en gros, différemment ou non de la population de même âge et sexe (quelles que soient les pathologies présentées par cette population générale).

Exemple : la maladie de Horton survient à un âge moyen de 75 ans, âge auquel l'incidence des maladies cardiovasculaires, comme l'IMy ou l'AVC, est particulièrement élevée.

Même sans groupe de comparaison tiré au sort dans la population générale avec contrôle des facteurs confondants, il a été possible de savoir, par la comparaison de la courbe de survie des patients avec celle de la population générale, que la survie des patients était similaire à celle de la population générale, pour les hommes comme pour les femmes, et ce malgré les cas rapportés de complications gravissimes comme les dissections aortiques ou rupture d'anévrysme [9].

Enfin, des techniques particulières de l'analyse de la survie ont été développées tout particulièrement pour les cohortes de patients, notamment l'analyse dite en survie relative. On a longtemps enseigné que le pronostic de la maladie de Hodgkin, qui présente un premier pic d'incidence entre 20 et 30 ans, et un second entre 60 et 70, était plus sévère chez les sujets âgés que chez les sujets jeunes : effectivement, les courbes de survie diffèrent radicalement. Simplement, les courbes de survie de sujets sains jeunes diffèrent elles aussi radicalement de celles de sujets plus âgés… et dire que la différence, dans le cas de la maladie de Hodgkin, était liée à une gravité particulière de la maladie chez les sujets âgés revenait tout simplement à ignorer que *de facto* les sujets plus âgés étaient… plus âgés. L'outil mathématique de l'analyse de la survie relative permet de prendre en compte, dans ce type de comparaison, la mortalité liée à l'âge et de la différencier de la mortalité liée à la maladie surajoutée. Il a ainsi pu être montré, à partir des données du registre du cancer de la Côte d'Or, que le pronostic du cancer du colon chez les sujets âgés, était le même que chez les sujets plus jeunes, et donc que la maladie se conduisait de façon similaire. Là encore, le débat n'est pas que théorique : si la forme du sujet âgé n'est pas plus grave que celle du sujet jeune, il convient certainement de ne pas la traiter de façon plus agressive… et de ne pas accroître le risque de complications iatrogènes [10].

D - Les études cas-témoins insérées dans les études de cohorte

Il est possible, au sein de la population bien définie de la cohorte préalablement assemblée, de réaliser des études cas-témoins de grande qualité à partir des cas reconnus au cours du suivi. Les témoins sont alors sélectionnés au sein de la population de la cohorte restée saine, et les informations sur les facteurs de risque étudiés obtenues à partir des données enregistrées à la phase initiale de l'étude de cohorte. Cette technique a plusieurs avantages :

1-La population de la cohorte sous-jacente est parfaitement définie : par conséquent, les cas et les témoins peuvent être tirés au sort sans biais de sélection, et être parfaitement représentatifs de la cohorte : la validité des résultats en est considérablement améliorée, puisque, dans une étude cas-témoins, les résultats reposent avant tout sur le choix des témoins (Cf. Chapitre V).

2-Les données utilisées dans l'étude cas-témoins nichée dans l'étude de cohorte sont celles de l'étude de cohorte sous-jacente, qui le plus souvent ont été colligées de façon prospective : ceci permet de limiter les biais de souvenir, voire de réaliser l'équivalent d'une étude cas-témoins prospective.

IV - EXPRESSION DES RESULTATS

Les études de cohorte permettent d'obtenir une mesure de l'incidence de la maladie et du risque relatif, c'est-à-dire du compte des nouveaux cas apparus lors du suivi de la cohorte et une mesure de la quantification du risque chez les personnes exposées par rapport aux personnes non exposées. Il est, là encore, plus facile d'exprimer, sous la forme d'une table 2×2, les résultats obtenus au terme de l'étude de suivi (tableau 1).

$$\text{Incidence} = \frac{a + c}{a + b + c + d} \text{ / an}$$

Soit Inc_{exp} l'incidence chez les sujets exposés : $Inc_{exp} = \frac{a}{a + b}$ / an

et Inc_{non} l'incidence chez les sujets non exposés : $Inc_{non} = \frac{c}{c + d}$ / an

Le risque relatif (RR) est le rapport de l'incidence de la maladie chez les sujets exposés à celle chez les sujets non exposés. Il répond à la question : combien de fois les sujets exposés ont-ils plus de chance que les sujets non exposés de contracter la maladie ?

$$RR = \frac{Inc_{exp}}{Inc_{non}} = \frac{a\,(c + d)}{c\,(a + b)}$$

On peut également calculer le nombre de cas, chez les sujets exposés, qui sont réellement dus à l'exposition. On peut en effet raisonnablement postuler que les cas survenus dans la population exposée regroupent les cas qui y seraient survenus spontanément, comme dans la population non exposée, et les cas secondaires à l'exposition. L'incidence "attribuable" à l'exposition est obtenue par la différence entre l'incidence dans la population exposée et l'incidence dans la population non exposée.

$$Inc_{attr} = Inc_{exp} - Inc_{non}$$

On peut calculer à partir du nombre de cas réellement dus à l'exposition, l'excès de risque que courent les sujets exposés par rapport à la population générale (valeur différente du risque relatif, calculé à partir de l'ensemble des cas survenus chez les sujets exposés).

En pratique, il faut tenir compte, dans le calcul de l'incidence, et du risque relatif, des multiples imperfections de la cohorte. Il y a en effet des sujets perdus de vue, potentiellement à risque de contracter la maladie, et des sujets qui, soit parce qu'ils décèdent, soit parce qu'ils contractent la maladie, ne sont plus à risque. La formule brute donnée ci-dessus doit donc être corrigée en fonction du type de cohorte suivie et des modifications qu'elle subit au cours du temps.

V - AVANTAGES ET FAIBLESSES DES ETUDES DE COHORTE

A - Avantages

- Elles établissent la séquence des événements.
- Elles ne sont pas soumises à la plupart des biais affectant les études cas-témoins :
  - biais de survie sélective,
  - biais de mesure rétrospective des facteurs de risque (biais de souvenir),
  - biais d'appariement (facteur d'appariement associé au facteur de risque étudié).

- Elles permettent le calcul de l'incidence, du risque relatif et d'autres variables évaluant le risque encouru par la population exposée et non exposée.

- Elles conviennent particulièrement à l'étude de maladies fréquentes.

- Elles sont les seules à pouvoir analyser la survie des patients en fonction du temps, et à permettre une comparaison de cette survie entre groupes de patients, ou avec la population générale.

B - Faiblesses et inconvénients

- Elles exigent l'inclusion d'un grand nombre de sujets à la phase initiale.

- Elles ne conviennent pas à l'étude de maladies rares, pour lesquelles le nombre de sujets initialement inclus deviendrait rédhibitoire.

- Elles sont longues et très coûteuses.

- Ce ne sont, par conséquent, pas des études exploratoires : les hypothèses testées doivent avoir acquis au préalable leur fondement scientifique par des études plus légères.

- Elles restent soumises à l'existence potentielle de biais de sélection, de mauvaise classification, ou de biais liés aux facteurs de confusion.

- Elles sont exposées à la présence éventuelle de biais liés à la perte de suivi des sujets inclus.

Fig. 1 - Structure d'une étude de cohorte

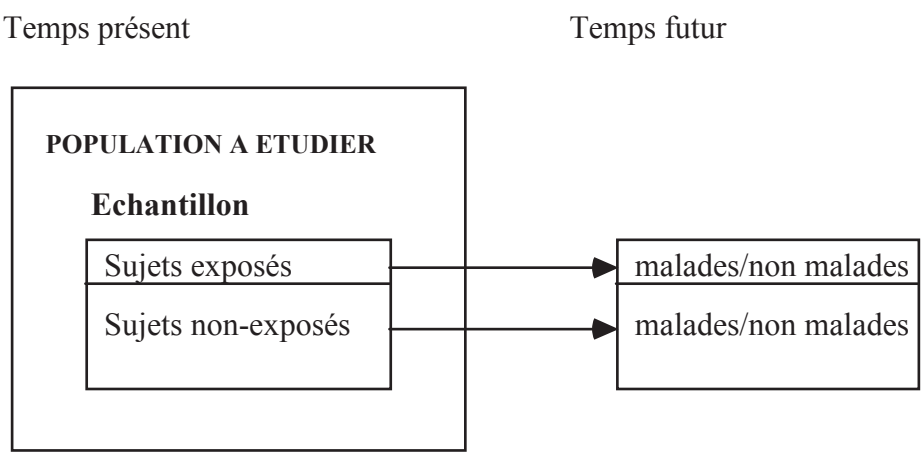
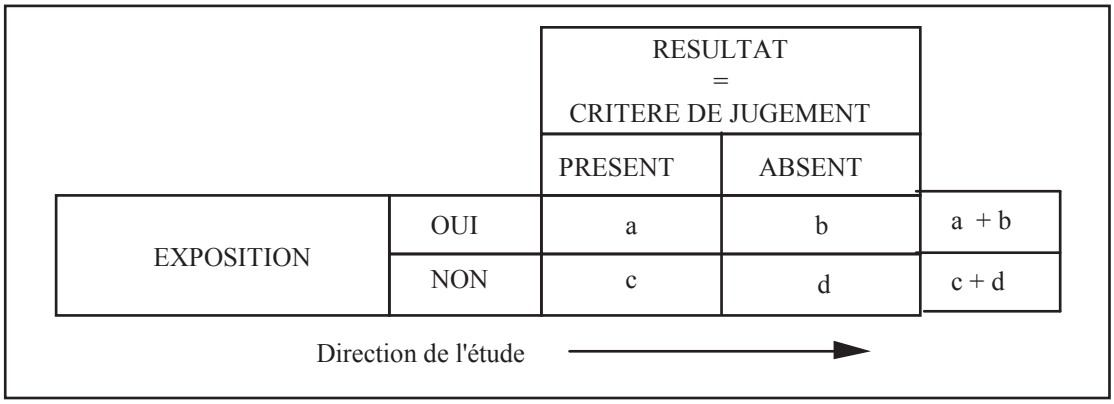


Fig. 2 - Représentation d'une étude de cohorte par la table 2 × 2: au départ de l'étude: (a + b) exposés comparés à (c + d) non exposés



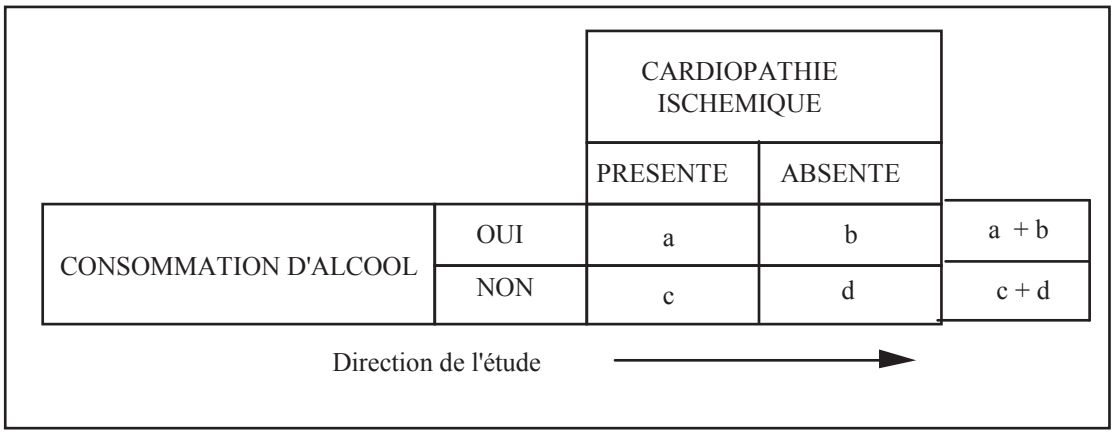


Fig. 3 - Facteur B, facteur de confusion

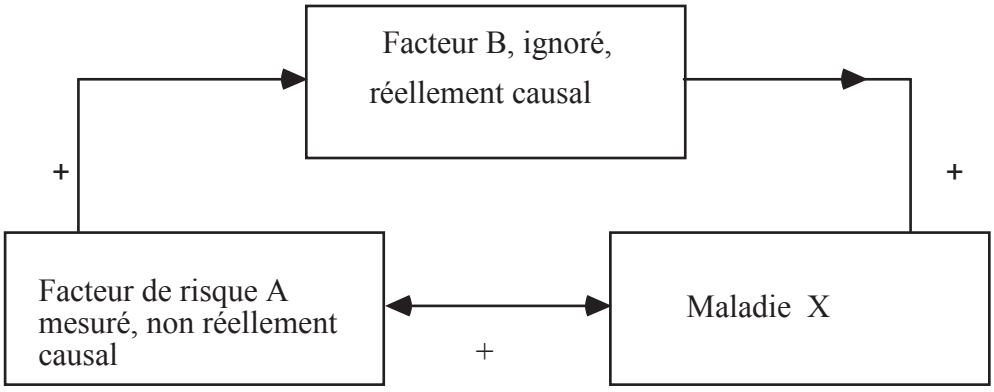


Fig. 4 - Etude de double cohorte

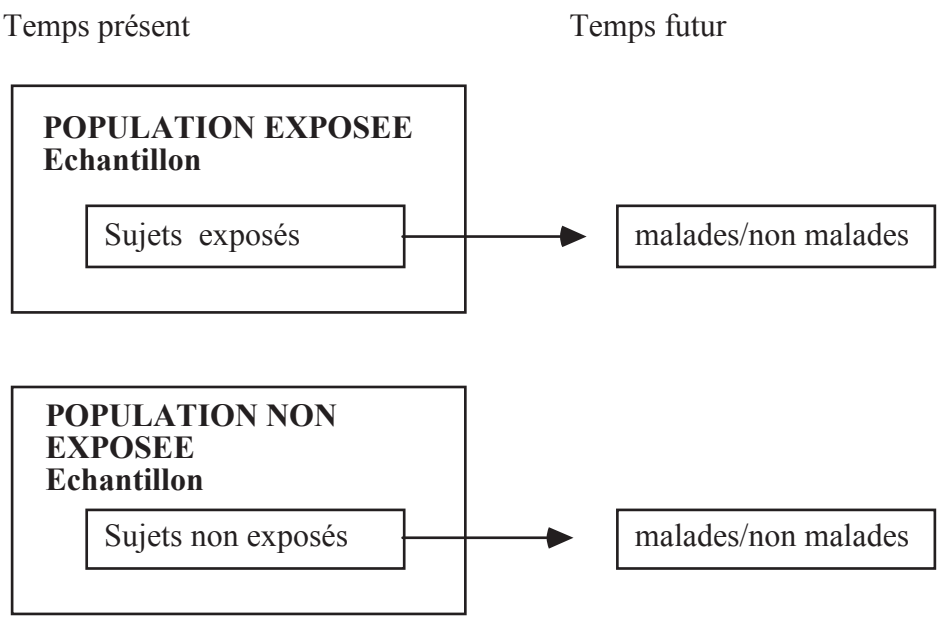


Tableau 1 - Calcul du risque relatif dans une étude de cohorte ; lecture horizontale de la table 2×2

	Malades	Non malades	
Exposés	a	b	a + b
Non exposés	c	d	c + d
	a + c	b + d	a + b + c + d

Références

1- Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. Br Med J 1976;2:1525-1536.
2- Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. Br J Cancer 2005;92:426-429.
3- R. Doll, R. Peto, E. Hall, K. Wheatley, R. Gray. Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. BMJ 1994; 309: 911–918.
4- Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observations on male British doctors. BMJ 1994;309:901-911.
5- Chatelain D, Duhaut P, Loire R, Bosshard S, Pellet H, Piette JC, Sevestre H, Ducroix JP. Diagnostic reproducibility of Giant Cell Arteritis on Temporal Artery Biopsy Specimen : The GRACG Study. 2007 ACR meeting, 10-14 Novembre 2007, Boston, USA.
6- Ehrenstein V, Rothman KJ, Pedersen L, Hatch EE, Sørensen HT. Pregnancy-associated hypertensive disorders and adult cognitive function among Danish conscripts. Am J Epidemiol 2009;170:1025-31.
7- Ehrenstein V, Pedersen L, Grijota M, Nielsen GL, Rothman KJ, Sørensen HT. Association of Apgar score at five minutes with long-term neurologic disability and cognitive function in a prevalence study of Danish conscripts. BMC Pregnancy Childbirth 2009;9:14.
8- Le Page L, Duhaut P, Seydoux D, Bosshard S, Ecochard R, Abbas F, Pétigny V, Cevallos R, Smail A, Salle V, Chatelain D, Loire R, Pellet H, Piette JC, Ducroix JP et l'ensemble des participants de l'étude GRACG (groupe de recherche sur l'artérite à cellules géantes). Incidence des événements cardiovasculaires au cours de la Maladie de Horton : Résultats préliminaires de la double cohorte prospective GRACG. Rev Med Interne 2006;27:98-105.
9- Matteson EL, Gold KN, Bloch DA, Hunder GG. Long-term survival of patients with giant cell arteritis in the American College of Rheumatology giant cell arteritis classification criteria cohort. Am J Med 1996;100:193-196.
10- Michiels C, Boutron MC, Chatelain N, Quipourt V, Roy P, Faivre J. Prognostic factors in colorectal adenocarcinoma of Dukes stage B. Study of a series of population. Gastroenterol Clin Biol 1994;18:456-461.

CHAPITRE VII

LES ESSAIS CLINIQUES

François Delahaye

Une fois que la nature de la maladie a été décrite et qu'on est capable de prédire son évolution naturelle, la question suivante est : « que peut-on faire pour le patient ? »

Face à un patient présentant une maladie donnée, le médecin prescrit un traitement : médicaments, exercice, chirurgie, régime…

Mais il y a aussi beaucoup d'autres façons d'intervenir pour améliorer la santé : dépistage, prévention, éducation, à l'échelle de l'individu comme à l'échelle de la population.

Quelle que soit la nature de l'intervention proposée, le principe est de chercher à savoir si celle-ci fait plus de bien que de mal au patient ou à la communauté. La technique d'évaluation est la même quelle que soit l'intervention : il s'agit de l'essai clinique, qui est un cas particulier d'étude de cohorte.

Dans un essai clinique, les conditions de l'étude — sélection des groupes qui bénéficient de l'intervention, nature de l'intervention, gestion du suivi — sont sous le contrôle de l'investigateur. Il s'agit donc d'une étude expérimentale, analogue dans le principe à celles qui sont conduites en laboratoire. Son principal avantage, par rapport à une étude d'observation, est de présenter un meilleur niveau de preuve lors de la mise en évidence d'une association entre deux facteurs ou d'une différence entre deux groupes.

Plan du chapitre

C - Randomisation

D - Administration des différentes interventions

E - Surveillance des patients

F - Mesure des événements ou critères de jugement dans les différents groupes et comparaison

G - Analyse des résultats

II - TYPES PARTICULIERS D'ESSAIS RANDOMISÉS EN INSU

A - Randomisation après période de rodage

B - Plan factoriel

C - Paires d'individus

D - Pré-randomisation

E - Randomisation de groupes

III - AUTRES PLANS D'ÉTUDE

A - Essais cliniques non randomisés

B - Essais cliniques sans insu

C - Essais cliniques en séries chronologiques

D - Plan d'étude en permutations croisées

E - Expérimentation naturelle

IV - CONCLUSION

Les essais cliniques sont des études de cohorte dans lesquelles l'investigateur manipule le facteur étudié, par exemple une intervention thérapeutique, et observe l'effet sur le critère de jugement. Celui-ci appartient à l'une des cinq catégories suivantes : la mort, la maladie, le handicap, l'inconfort et l'insatisfaction (ou leur inverse).

Le principal avantage de l'expérimentation sur l'étude d'observation est la puissance de l'inférence causale qu'elle permet (niveau de preuve très élevé).

La randomisation — répartition au hasard des sujets dans les différents groupes — est la meilleure solution pour contrôler l'influence des facteurs de confusion, qu'ils soient connus ou inconnus, en répartissant également ces facteurs dans les différents groupes, de sorte que leur effet sur le critère de jugement s'annule et qu'ainsi, si une différence est observée entre les deux groupes, elle est exclusivement due à l'effet du facteur étudié (figure 1).

L'insu permet d'éviter les co-interventions.

Les plus fréquents des essais cliniques sont les essais médicamenteux, qui constituent une forme particulière des essais thérapeutiques, eux-mêmes classés dans la catégorie des essais d'intervention (à côté de la thérapeutique, les interventions peuvent être aussi de dépistage, de prévention ou d'éducation ; l'essai se fait sur le même modèle). Nous parlerons surtout dans ce chapitre des essais thérapeutiques.

L'essai thérapeutique est essentiel pour le développement de nouveaux traitements. La première étape du développement utilise l'expérimentation *in vitro* et sur des animaux, et permet de préciser la pharmacologie et la toxicologie du produit. Le développement comporte ensuite quatre phases chez l'homme :

- les études de phase I visent à préciser la sûreté et la tolérance ; elles sont faites chez un petit nombre de sujets ;
- les études de phase II précisent l'efficacité optimale du traitement ;
- les études de phase III établissent l'efficacité du traitement, le plus souvent grâce à des essais thérapeutiques comparatifs, idéalement randomisés ;
- les études de phase IV, après la commercialisation, visent à établir les éventuelles nouvelles indications et les effets indésirables non décelés pendant les étapes précédentes.

I - ESSAI THÉRAPEUTIQUE RANDOMISÉ EN INSU

L'élaboration d'un essai thérapeutique randomisé en aveugle, ou mieux dit en insu (ETRI), comporte sept étapes : la constitution de la cohorte d'étude, la réalisation des mesures de base, la

randomisation, l'administration de la thérapeutique (le facteur étudié), la surveillance des patients, la mesure et la comparaison des événements dans les différents groupes (le ou les critères de jugement), et enfin l'analyse des résultats.

A - Constitution de la cohorte d'étude

Lors de cette première étape, on précise les caractéristiques de la population et le mode d'échantillonnage.

1 - Critères d'inclusion

Ils définissent les principales caractéristiques des populations potentielles et accessibles (figure 1 en haut). Les caractéristiques cliniques et démographiques (âge, sexe, race) permettent de définir la population potentielle, celle à laquelle on peut généraliser les résultats de l'étude. Les critères géographiques et temporels permettent de définir la population accessible, la partie de la population potentielle qui est disponible pour l'étude. La définition de ces critères passe par l'acceptation d'un compromis entre buts scientifiques et contraintes pratiques.

2 - Critères d'exclusion

Ils sont définis pour éliminer, parmi les sujets éligibles, ceux qui risquent d'interférer avec la qualité des données ou l'interprétation des résultats : alcooliques, patients ayant des problèmes psychiatriques, sujets susceptibles de déménager…

Les critères d'exclusion permettent d'améliorer la faisabilité d'une étude, mais il faut les utiliser avec modération car une meilleure homogénéité de la population étudiée se fait au détriment de la « généralisabilité » des résultats.

Certains critères d'exclusion sont imposés par l'éthique, ou par le désir d'un sujet de ne pas participer.

3 - Population

Il y a deux grandes sortes de populations accessibles :

- les sujets issus d'échantillons définis à l'hôpital sont peu chers et faciles à recruter, mais les facteurs de sélection peuvent avoir un effet important, notamment sur la possibilité de généralisation des résultats à une large population ;
- les sujets recrutés à domicile constituent un échantillon représentatif d'une région spécifique ; ces échantillons sont particulièrement utiles pour guider la pratique clinique ou de santé publique dans la communauté ; mais il existe deux grands désavantages : la difficulté de réalisation et le coût élevé.

On peut parfois éviter d'avoir recours à un échantillonnage, et donc en éviter les biais, lorsque la population accessible ainsi définie l'est réellement dans son ensemble — ce fut le cas en 1976 pour l'épidémie de légionellose où tous les cas ont été identifiés. C'est la meilleure approche. Habituellement cependant, la population accessible est trop grande, il faut alors sélectionner un groupe plus petit, c'est l'échantillonnage.

4 - Échantillonnage

L'échantillonnage probabiliste utilise le tirage au hasard, afin de garantir que chaque unité de la population a une probabilité spécifiée de sélection. C'est l'approche la plus rigoureuse.

Il peut s'agir :

- d'un tirage au sort simple, par exemple grâce à une table de nombres au hasard générée par un ordinateur (tableau 1) ;
- d'un tirage au sort stratifié (la population est divisée en sous-groupes selon des caractéristiques telles que sexe ou âge, et le tirage au sort est fait dans chacun de ces groupes) ;
- ou d'un tirage au sort de groupes (on tire au hasard des groupes naturels d'individus, par exemple des équipes de footballeurs).

L'échantillonnage non probabiliste est beaucoup plus facile que l'échantillonnage probabiliste, mais moins rigoureux, puisqu'il n'utilise pas le hasard.

B - Mesures de base

Certaines mesures permettent de caractériser les sujets inclus dans l'étude : le nom et le prénom, l'adresse et le numéro d'hospitalisation, mais aussi des caractéristiques démographiques et cliniques, tels l'âge, le sexe et le diagnostic. Ces mesures sont des informations importantes car elles permettent aussi de comparer la composition des groupes qui constituent l'essai. Habituellement, le premier tableau du rapport final d'un ETRI compare les diverses caractéristiques de base des sujets des différents groupes. Le but est de vérifier que les différences ne sont pas plus grandes que celles qu'on aurait pu observer par le simple fait du hasard. Si tel était le cas, cela suggérerait une erreur technique dans le processus de randomisation. Cela ferait en outre courir le risque, en cas de différence observée entre les deux groupes pour le critère de jugement, que cette différence ne soit pas due à l'intervention thérapeutique, mais à la ou les caractéristiques dont on observe un déséquilibre entre les deux groupes.

Il est souvent utile de mesurer le critère de jugement au début de l'étude comme à la fin. Lorsque le critère est une variable dichotomique (voir chapitre 16), il est important de démontrer qu'il n'est pas présent au début. Lorsqu'il s'agit d'une variable continue, on peut utiliser la différence entre les

deux groupes dans le degré de changement de la variable au cours de l'étude. Cette approche permet de contrôler les différences initiales, et peut conférer à l'étude une puissance plus grande que la simple comparaison des valeurs à la fin de l'étude.

Il faut mesurer les divers prédicteurs connus de l'événement étudié (critère de jugement), c'est-à-dire les facteurs connus comme pouvant avoir une influence sur le critère de jugement, indépendamment du facteur étudié. Cela permet l'ajustement statistique des résultats, qui réduit les effets d'une mauvaise distribution inopinée des variables prédictives entre les deux groupes. L'efficacité de l'étude en est ainsi augmentée. Cela permet aussi à l'investigateur d'examiner ces autres prédicteurs dans une autre question de recherche.

Enfin, il ne faut pas mesurer trop de variables, car cela augmente le coût et la complexité.

C - Randomisation

L'allocation aléatoire établit les bases du test de la significativité statistique de différences entre les groupes quant à l'événement d'intérêt. Elle permet que l'âge, le sexe, et d'autres caractéristiques de base, connues ou inconnues, qui pourraient « confondre » une association observée soient réparties également (sauf variation aléatoire) entre les groupes randomisés. Ainsi, le résultat observé entre les groupes à la fin de l'essai peut-il être attribué à l'effet de l'intervention, puisque l'effet que pourraient avoir des facteurs de confusion — facteurs qui modifient les effets du facteur étudié sur le critère de jugement, du fait de leur lien à la fois avec le facteur étudié et avec le critère de jugement — est également réparti dans les divers groupes.

Les effets d'une mauvaise répartition par le simple fait du hasard (en moyenne 1 caractéristique de base sur 20 est répartie différemment entre les groupes au risque 0,05) sont pris en compte dans les tests statistiques permettant de calculer la probabilité que le hasard soit responsable de la différence observée entre les groupes pour l'événement étudié.

Parce que la randomisation est une des pierres angulaires d'un bon ETRI, il est important qu'elle soit bien faite. Les deux éléments les plus importants sont qu'une procédure de vraie allocation au hasard soit élaborée, et que le processus de randomisation soit inaltérable, de telle façon que des biais, intentionnels ou non, ne puissent influencer le processus d'allocation.

Habituellement, le patient subit les investigations de base, est considéré éligible pour l'inclusion, et donne son consentement éclairé. C'est ensuite qu'il est randomisé, par l'application, manuelle ou automatique, d'un algorithme à un ensemble de nombres au hasard, et son allocation dans l'un des groupes est irréversible.

D - Administration des différentes interventions

La randomisation protège de l'influence des facteurs de confusion présents au moment de l'allocation dans les différents groupes. En revanche, elle n'a aucun effet sur ceux qui apparaissent pendant la surveillance.

Chaque fois que possible, l'investigateur doit concevoir le mode d'administration de l'intervention de telle façon que ni les sujets, ni quiconque ayant des contacts avec eux, ne connaissent le traitement reçu (tableau 2). Le terme simple insu est utilisé lorsque le patient ne sait pas quel produit (intervention testée ou placebo) il prend ; il s'agit de double insu lorsque ni le patient ni le médecin ne savent ce que le patient prend.

Dans une **étude ouverte** (sans insu), l'investigateur peut prêter une attention particulière au patient lorsqu'il sait que celui-ci reçoit le traitement testé. Cette attitude différente peut représenter une véritable intervention (co-intervention). Les co-interventions peuvent aussi affecter le groupe témoin (par exemple, les sujets découvrant qu'ils sont sous placebo demandent d'autres traitements). Ces co-interventions peuvent être la véritable raison d'une différence de fréquence de l'événement étudié entre les groupes. Une solution partielle au problème des interventions non planifiées est de spécifier et de standardiser l'intervention.

Une stratégie beaucoup plus efficace est de faire **l'étude en double insu**, c'est-à-dire de cacher la nature du traitement assigné, à la fois au sujet et à l'investigateur. Quand le double insu est techniquement bon, toute intervention non planifiée doit affecter également les deux groupes (avec l'exception, comme pour la randomisation, d'une mauvaise distribution par le fait du hasard), et ne peut altérer la comparaison de l'événement entre groupes.

Les contraintes logistiques peuvent être lourdes. Il faut préparer des capsules identiques (forme, taille, couleur, goût…), et développer des systèmes indéréglables d'étiquetage et de distribution.

Il peut être nécessaire de mettre en place un mécanisme de levée de l'insu fonctionnant 24 heures sur 24 quand la situation exige que l'on soit capable de savoir très rapidement quelle drogue le sujet prend.

Une autre difficulté majeure est de faire en sorte que ni les sujets ni les investigateurs ne puissent deviner le traitement assigné.

De nombreuses interventions thérapeutiques ne peuvent pas être réalisées à l'insu du médecin ou du patient (par exemple, intervention chirurgicale).

Il est important de choisir une intervention susceptible d'être généralisée à la pratique médicale quotidienne. Le choix du bon traitement peut être particulièrement difficile dans les études

nécessitant une surveillance de plusieurs années, car un traitement courant au début de l'étude peut être démodé à sa fin.

Les meilleurs groupes témoins sont ceux qui ne reçoivent pas de traitement actif, mais un placebo identique dans la forme, la couleur, le goût… au médicament étudié. Cette stratégie compense l'éventuel effet placebo de la thérapeutique testée, de telle façon qu'une différence entre les groupes d'étude peut être effectivement attribuée à un effet biologique.

Une autre possibilité est la comparaison d'un traitement avec un autre traitement considéré comme efficace. Si aucune différence n'est mise en évidence, le risque est de conclure que les deux traitements sont équivalents. En réalité, la méthodologie des tests d'équivalence est différente.

E - Surveillance des patients

Un patient est **observant (« compliant »)** lorsqu'il suit les consignes données par le corps médical sur le traitement instauré ou l'intervention proposée. On parle aussi d'adhérence au traitement. L'observance du patient (par exemple, venir aux consultations programmées, et le faire à la date prévue, prendre le produit prescrit…), doit être bonne.

L'effet de l'intervention, et donc la puissance de l'étude, sont diminués lorsque les sujets ne sont pas observants. L'investigateur doit essayer de choisir une intervention aisément tolérée, il faut privilégier la prise unique.

L'observance doit être mesurée, par exemple par autoquestionnaire, comptage de comprimés ou analyse de métabolites urinaires.

La non observance suggère une volonté délibérée de la part du patient de ne pas tenir compte des conseils et prescriptions. Mais d'autres facteurs peuvent aussi intervenir : le patient peut mal comprendre quel médicament il doit prendre, et à quelles doses, il peut se trouver à court de médicament, ne pas avoir d'argent pour aller à la pharmacie ou confondre les différents médicaments…

F - Mesure des événements ou critères de jugement dans les différents groupes et comparaison

Dans le **choix du type de critère de jugement**, l'investigateur doit souvent mettre en balance des considérations opposées.

Souvent les événements choisis comme critère de jugement d'une étude ne sont pas les véritables événements, mais des événements de substitution du véritable phénomène intéressant, ce qui limite les inférences possibles (par exemple, dans l'essai d'un médicament fibrinolytique dans l'infarctus du myocarde, on peut utiliser comme critère de jugement de l'efficacité du produit la fraction d'éjection ventriculaire gauche ou la perméabilité coronaire plutôt que la mortalité).

La **mesure du critère de jugement** doit être exacte. Les variables continues ont l'avantage par rapport aux variables dichotomiques d'augmenter la puissance de l'étude, autorisant ainsi à recruter un nombre plus petit de patients. Si on ne peut éviter une variable dichotomique, la puissance dépend plus du nombre d'événements que du nombre total de sujets.

Il est souvent souhaitable d'avoir **plusieurs variables** mesurant des aspects différents (par exemple, dans l'étude de l'efficacité d'un médicament dans la prévention secondaire après infarctus du myocarde, les événements étudiés peuvent être la mortalité et la récidive d'infarctus).

L'investigateur doit également prévoir de **mesurer les effets secondaires** liés à l'intervention, du symptôme relativement mineur à la complication grave, voire la mort. Évaluer si les bénéfices d'une intervention dépassent ses risques est le but principal de la plupart des ETRI. Malheureusement, les effets secondaires rares sont habituellement impossibles à détecter, quelle que soit la taille de l'étude, et ne peuvent être découverts que par une étude cas-témoins après que le produit a été très largement utilisé dans la population.

Le critère de jugement doit pouvoir être mesuré sans connaître l'assignation du sujet à tel ou tel groupe.

L'utilisation du double insu est particulièrement importante quand la mesure du critère de jugement nécessite une intervention subjective de la part de l'observateur. La connaissance par l'observateur de l'appartenance du patient au groupe traité ou au groupe témoin peut modifier son évaluation du critère de jugement. Le double insu évite que le biais de constatation soit plus important dans un groupe que dans l'autre.

Le terme **triple insu** est parfois utilisé pour bien montrer que le traitement pris est inconnu de trois personnes : le patient, la personne administrant le traitement, et la personne mesurant le critère de jugement.

Les stratégies pour obtenir un taux de réponse élevé sont les mêmes que celles utilisées dans les études de cohorte : exclusion des sujets pour lesquels la surveillance paraît difficile (alcooliques, sujets psychiatriques…) ; information claire du sujet sur l'importance d'une surveillance correcte, élimination de ceux qui trouvent cette surveillance difficile ; enregistrement des coordonnées d'un ou deux proches qui sauront toujours où le sujet se trouve, et de celles du médecin traitant ; contacts téléphoniques réguliers avec les patients.

Avoir une surveillance pour 100% ou presque des sujets peut être primordial lorsque l'événement est rare et qu'il constitue une cause possible de perte de vue.

G - Analyse des résultats

Lorsque le critère de jugement évalué est dichotomique, on compare les proportions d'événements dans les groupes grâce au **test de $\chi 2$.**

Quand la variable événement est continue, un **test t** peut être utilisé, ou un **test non paramétrique** lorsque la variable n'est pas distribuée normalement.

Les méthodes prenant le temps en compte sont utiles lorsqu'il y a des différences de durée de surveillance entre les participants, et l'**analyse de régression selon le modèle de Cox** peut être utilisée pour ajuster sur les distributions inégales des facteurs de confusion de base (ce qui augmente la puissance).

Trois problèmes doivent être considérés au moment de l'élaboration d'une étude :

- **primauté de l'analyse selon l'intention de traiter**, la seule valable méthodologiquement : les résultats sont analysés — les patients restant dans le groupe auquel ils ont été assignés, même s'ils ont changé de groupe pendant l'essai (par exemple, un sujet randomisé dans le groupe chirurgical mais qui n'est finalement pas opéré, ou l'inverse, un sujet randomisé dans le groupe traitement médical, mais qui est opéré dans les suites) ;
- **rôle ancillaire des analyses de sous-groupes** quand elles n'ont pas été prévues *a priori* dans le protocole avant le début de l'essai : les études de sous-groupes faites *a posteriori* ne sont là que pour générer des hypothèses, qui devront être étudiées dans un nouvel essai ;
- **opportunité de fixer des règles d'arrêt prématuré** en prévoyant des analyses intermédiaires (avec des tests statistiques particuliers) : il faut pouvoir arrêter un essai prématurément si l'intervention s'avère efficace plus rapidement que prévu (pour faire bénéficier l'ensemble de la population concernée de cette avancée thérapeutique) ou au contraire si elle s'avère délétère (pour ne pas y exposer plus les patients qui la reçoivent).

II - TYPES PARTICULIERS D'ESSAIS RANDOMISÉS EN INSU

A - Randomisation après période de rodage

Elle est utile pour augmenter la proportion de sujets observants. Après l'identification de la cohorte d'étude et l'obtention du consentement, tous les sujets sont mis sous placebo. Plus tard, ceux qui ont été observants sont randomisés. L'exclusion des sujets non observants avant la randomisation augmente la puissance de l'étude et permet une meilleure estimation des effets de l'intervention.

Une variante de ce plan d'étude est l'utilisation, pendant la période de rodage, du produit actif. Ici la réponse d'une variable intermédiaire (c'est-à-dire qui se situe entre l'intervention et l'événement) peut être utilisée comme critère pour la randomisation (par exemple, pour l'étude de l'effet d'un

anti-arythmique sur la mortalité, on sélectionne les patients chez lesquels, pendant la période de rodage, l'anti-arythmique a entraîné la disparition du trouble du rythme).

Il est important dans le rapport d'un essai après rodage de préciser les éventuelles différences dans les caractéristiques de base entre les patients randomisés et les patients non randomisés.

B - Plan factoriel

Il permet de répondre à plusieurs questions de recherche dans un seul essai.

Un exemple est l'étude de l'influence de l'aspirine sur l'infarctus du myocarde et de celle du bêta-carotène sur le cancer (figure 2). Les sujets ont été randomisés en 4 groupes, et chacune des deux hypothèses a pu être testée en comparant deux moitiés de la cohorte d'étude :

- tous ceux prenant de l'aspirine sont comparés à tous ceux sous placebo d'aspirine (sans se préoccuper du fait que la moitié de chaque groupe reçoit du bêta-carotène) ;
- tous ceux prenant du bêta-carotène sont comparés à tous ceux sous placebo de bêta-carotène (sans se préoccuper du fait que la moitié de chaque groupe reçoit de l'aspirine).

L'investigateur a deux études complètes pour le prix d'une (il n'y a pas besoin de plus de sujets que le nombre nécessaire pour un essai ne testant qu'une hypothèse).

Le plan factoriel est un plan d'étude extrêmement efficace. La principale limitation est le problème des interactions entre les deux relations cause-effet en cours d'étude.

C - Paires d'individus

Une stratégie pour répartir de façon égale dans les deux groupes les facteurs de confusion de base est de sélectionner des paires d'individus, qu'on apparie selon des facteurs importants, tels que sexe et âge, puis qu'on randomise, chaque sujet d'une paire pour chacun des groupes de traitement.

Une version particulièrement séduisante de cette approche est possible lorsque les circonstances permettent un contraste des effets du traitement et du témoin dans deux parties d'un même individu au même moment. Par exemple, chez des sujets ayant une rétinopathie diabétique, chaque sujet a, au hasard, un œil assigné au traitement (photocoagulation), l'autre œil servant de témoin.

D - Pré-randomisation

Elle suppose la randomisation avant d'obtenir le consentement éclairé, demandé ensuite en utilisant des formulaires différents pour les deux groupes.

Cette approche peut augmenter le taux d'inclusion par la suppression de la barrière psychologique qu'entraîne parfois l'incertitude de l'assignation à un groupe de traitement, mais la puissance est

diminuée au prorata de la proportion de sujets qui refusent d'être inclus, mais qui doivent être analysés de façon à satisfaire à la règle de l'analyse selon l'intention de traiter.

Par ailleurs, la mesure du critère de jugement chez des sujets qui ont refusé de participer pose un problème éthique.

E - Randomisation de groupes

Au lieu de randomiser des individus, un investigateur peut choisir de randomiser des groupes naturels de sujets, des usines, des villes…

Parmi les avantages d'une telle conception, il y a notamment le fait d'éviter que les sujets qui reçoivent une intervention transmissible, tel un conseil diététique, puissent discuter de cette intervention avec des connaissances appartenant à la même sorte de population, mais affectées à l'autre groupe.

Mais l'estimation de la taille de l'étude et l'analyse sont plus difficiles.

III - AUTRES PLANS D'ÉTUDE

A - Essais cliniques non randomisés

Ils sont beaucoup moins satisfaisants que les ETRI pour contrôler l'influence des facteurs de confusion. Des méthodes d'analyse permettent d'ajuster pour les facteurs de base inégalement répartis, mais cela ne règle pas le problème des facteurs de confusion inconnus ou non mesurés.

B - Essais cliniques sans insu

Ils sont eux aussi moins satisfaisants que les ETRI, du fait du risque de confusion par des co-interventions, et de biais de constatation de l'événement affectant un groupe plus que l'autre.

Quand les circonstances n'autorisent pas le double insu, le simple insu est habituellement possible (le patient ne sait pas ce qu'il prend). Ce plan d'étude ne protège cependant pas des co-interventions, et il devrait rarement être nécessaire : des interventions qui peuvent être cachées aux patients peuvent habituellement l'être aussi aux investigateurs.

Une forme plus fréquente d'insu partiel est le processus de constatation de l'événement (critère de jugement) en insu, dans une étude ouverte. De telles études peuvent fournir des conclusions très utiles, mais ces conclusions sont habituellement moins solides que celles des études en double insu.

C - Essais cliniques en séries chronologiques

Ils peuvent être utiles pour certains types de questions. Chaque sujet est son propre témoin pendant les périodes séquentielles de traitement et de contrôle. Cela signifie que les caractéristiques personnelles telles qu'âge, sexe, et facteurs génétiques, facteurs de confusion potentiels, ne sont pas

réparties également, mais tout bonnement éliminées. Cela signifie également que l'étude nécessite deux fois moins de sujets, puisque chaque sujet fournit à la fois les observations témoins et expérimentales.

Ce plan d'étude n'est utile que dans certaines circonstances : études dont l'événement répond rapidement et de façon réversible à l'intervention ou expérimentations à long terme ne pouvant être randomisées. Les plus grands inconvénients sont le problème des facteurs de confusion selon le temps, et celui de l**'effet de report** (influence résiduelle de l'intervention sur l'événement après que cette intervention a été arrêtée).

D - Plan d'étude en permutations croisées

L'influence de covariables dépendantes du temps peut être contrôlée par un plan d'étude en permutations croisées, dans lequel la moitié des participants est randomisée à recevoir le placebo d'abord, puis le traitement, l'autre moitié faisant l'inverse (figure 3).

Cette approche, ou un équivalent tel que le carré latin lorsqu'il y a plus de deux groupes, a des avantages substantiels : contrôle des facteurs de confusion, doublement effectif de l'échantillon.

Cependant, les inconvénients sont souvent encore plus grands : doublement de la durée de l'étude, complexité accrue de l'analyse et de l'interprétation.

Les études en permutations croisées ne sont un bon choix que quand les sujets sont difficiles à recruter, quand il y a de bonnes raisons de penser que l'effet de report n'est pas un problème, ou quand cet effet de report constitue par lui-même une partie de la question de recherche.

E - Expérimentation naturelle

L'investigateur analyse une situation dans laquelle quelqu'un d'autre applique une intervention. Les expérimentations naturelles ressemblent en fait autant à des études d'observation qu'à des expérimentations puisque l'intervention n'est pas manipulée par l'investigateur, et le contrôle des facteurs de confusion est plutôt limité, sauf si l'expérimentation comprend une randomisation.

IV - CONCLUSION

L'essai clinique randomisé en double insu constitue la référence en matière de protocole de recherche clinique, fournissant le niveau de preuve le plus élevé quand il s'agit de mettre en évidence une relation entre deux facteurs.

La randomisation est l'intervention fondamentale puisqu'elle supprime le risque d'erreur lié aux facteurs de confusion, et le processus d'insu élimine le risque de biais lié aux co-interventions.

Les résultats d'un essai sont jugés par rapport à deux grandes questions :

- le traitement est-il efficace dans des circonstances idéales ? On juge l'efficacité du traitement chez les patients qui le reçoivent, et qui sont pleinement coopérants, c'est-à-dire observants ;
- le traitement est-il efficace dans des conditions ordinaires ? On juge l'efficacité du traitement chez les patients à qui l'on a proposé ce traitement, et qui sont libres de l'accepter ou de le refuser ; ce sont des patients dont l'observance peut ne pas être bonne ; c'est la différence entre approche expérimentale (circonstances idéales) et approche pragmatique (circonstances ordinaires ou intention de traiter) ; cette dernière approche offre des résultats plus facilement généralisables.

Il est évident qu'il n'est pas possible de répondre à toutes les questions de recherche clinique par un essai, pour des raisons éthiques, méthodologiques ou budgétaires. Ce sont alors les études d'observation analytiques de type cohorte ou cas-témoins qui sont utilisées. Mais leur validité peut être jugée par l'écart qui existe entre leur protocole et celui théorique d'un essai clinique que l'on aurait voulu pouvoir réaliser pour répondre à la question.

Dans la pratique médicale, un essai clinique randomisé peut perturber les relations médecin-patient. La possibilité pour le patient d'appartenir au groupe témoin, l'allocation du traitement au hasard ou la gestion en insu peuvent rendre inconfortable la relation médecin-patient habituelle, hors expérimentation, dont le seul objectif est le soin apporté au patient.

On a pu reprocher à l'essai clinique de mettre certains patients dans une situation dans laquelle ils ne peuvent pas bénéficier du meilleur traitement possible.

Si effectivement il existe un bon niveau de preuve pour affirmer la supériorité d'un traitement, ne pas l'offrir à l'ensemble des patients n'est pas éthique.

Mais si réellement on n'a pas la preuve de cette supériorité, alors il est légitime d'offrir au patient ce traitement aussi bien que son alternative.

On pourrait même dire qu'il n'est pas éthique de proposer à des patients des traitements qui n'ont pas été évalués de façon rigoureuse et dont on ne connaît pas l'efficacité.

D'autre part, bien que l'essai randomisé soit coûteux et difficile à mener, l'alternative à l'essai, c'est-à-dire l'administration d'un traitement sans information solide sur son efficacité, est probablement bien plus coûteuse. Finalement, un essai bien conçu et bien mené peut faire économiser de l'argent.

Il faut donc considérer que le principe de l'expérimentation en clinique est juste. Il s'agit seulement, mais c'est fondamental, d'offrir des garanties pour qu'un patient ne puisse pas participer à un essai contre sa volonté, spécialement dans cette relation médecin-patient où le médecin peut avoir un

pouvoir considérable. Proposer à un patient de participer à un essai de manière telle qu'il peut refuser, et, s'il accepte, avec la garantie que ses droits seront respectés, telles sont les contraintes légitimes imposées par les comités d'éthique devant lesquels tout protocole d'essai thérapeutique doit être soumis pour approbation. Ces comités veillent au respect des principes fondamentaux de l'éthique en recherche biomédicale : principe de l'intérêt et du bénéfice de la recherche, principe de l'innocuité de la recherche, principe du respect de la personne, et principe de justice.

L'information du patient est fondamentale. Il est difficile d'être sûr que le fait que le patient a signé le formulaire de consentement signifie réellement qu'il a compris toutes les informations concernant la recherche. Cela veut dire que la capacité de l'investigateur à communiquer pleinement et honnêtement avec le patient est primordiale.

Références

Cucherat M. Méthodologie et interprétation des essais cliniques. Éditions Flammarion Médecine-Sciences, 2004

Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. Éditions Springer-Verlag, 1998

Tableau 1 - Table de nombres au hasard

09801 20131 47650 20546 79800

01638 79004 13891 00746 26571

05441 02614 89720 18096 10974

58001 07467 19853 10074 32052

01985 49872 30106 24198 10023

14941 10123 45678 91019 51032

57489 32002 47921 00164 59758

74431 01320 48372 85967 45116

50206 12497 65773 12131 41516

17181 92021 22232 42526 27282

15424 70461 61241 21234 37989

15200 76746 59116 01246 42749

75975 46013 01654 97978 67240

10404 25704 01310 42795 79573

20275 12707 58067 84150 05178

01314 69746 02040 71781 65405

Tableau 2 - Les quatre niveaux d'insu

Le processus d'insu peut intervenir à quatre niveaux dans un essai clinique :

- celui qui est chargé de la répartition des patients dans chaque groupe ne sait pas comment sont répartis les patients déjà inclus dans l'étude afin de ne pas risquer de modifier sa façon d'inclure les patients suivants dans l'étude ;
- les patients ne savent pas dans quel groupe de traitement ils se trouvent ; ainsi, il y a moins de risque pour qu'ils changent leur observance ou qu'ils décrivent leur état en fonction de leur appartenance à l'un ou l'autre groupe ;
- les médecins chargés de la surveillance des patients qui participent à l'étude ne savent pas à quel groupe leurs patients appartiennent ; ainsi, leurs soins ne risquent pas d'être modifiés, même inconsciemment ;
- lorsque le chercheur évalue le ou les critères de jugement, il ne sait pas à quel groupe appartient le patient ; ainsi, la mesure du critère de jugement ne risque pas d'être modifiée, même inconsciemment.

Le terme « simple insu » fait référence au patient seul, le terme « double insu » au patient et au chercheur.

Figure 1 - Schéma général d'un essai thérapeutique

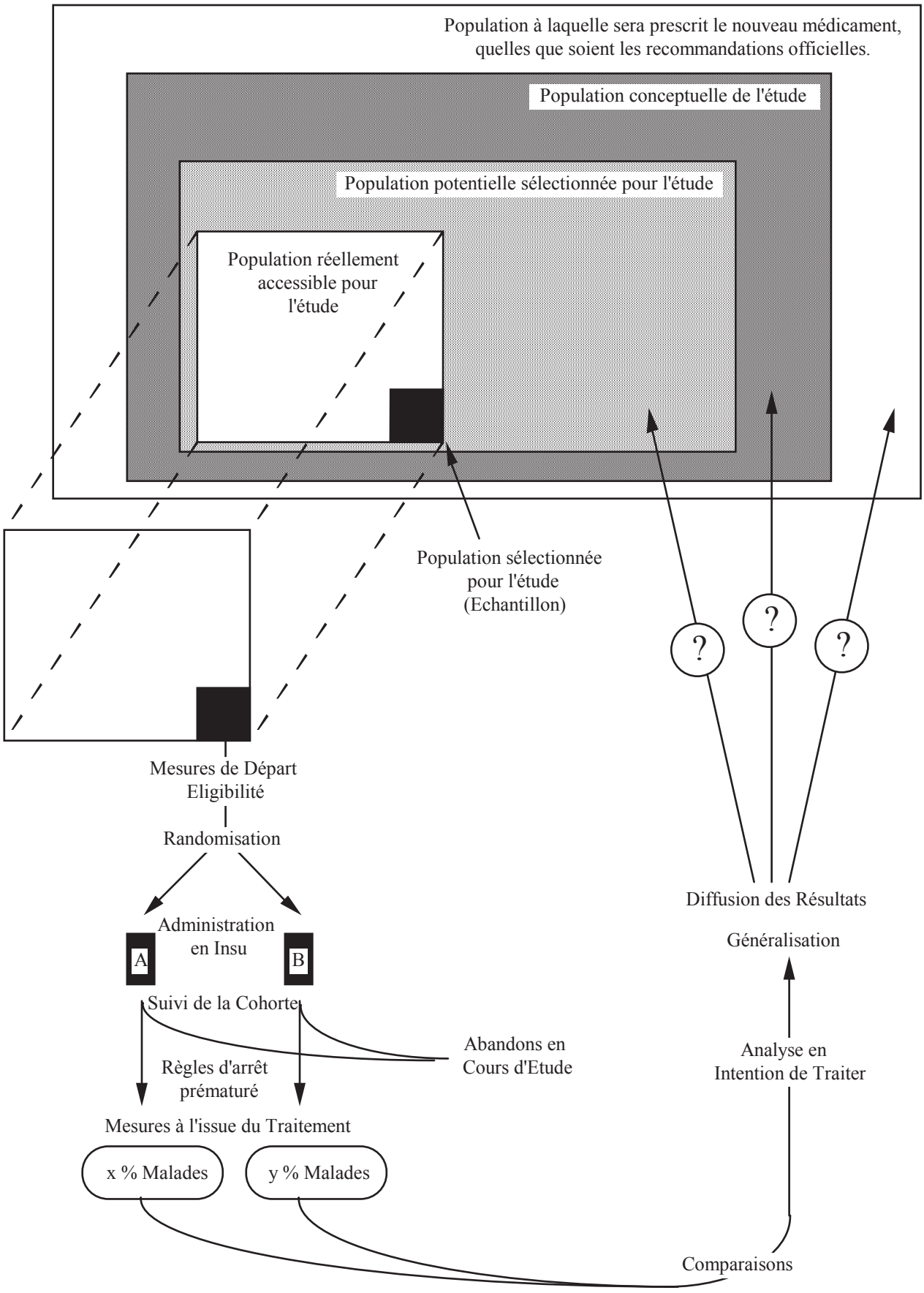


Figure 2 - Essai thérapeutique randomisé en insu par plan factoriel

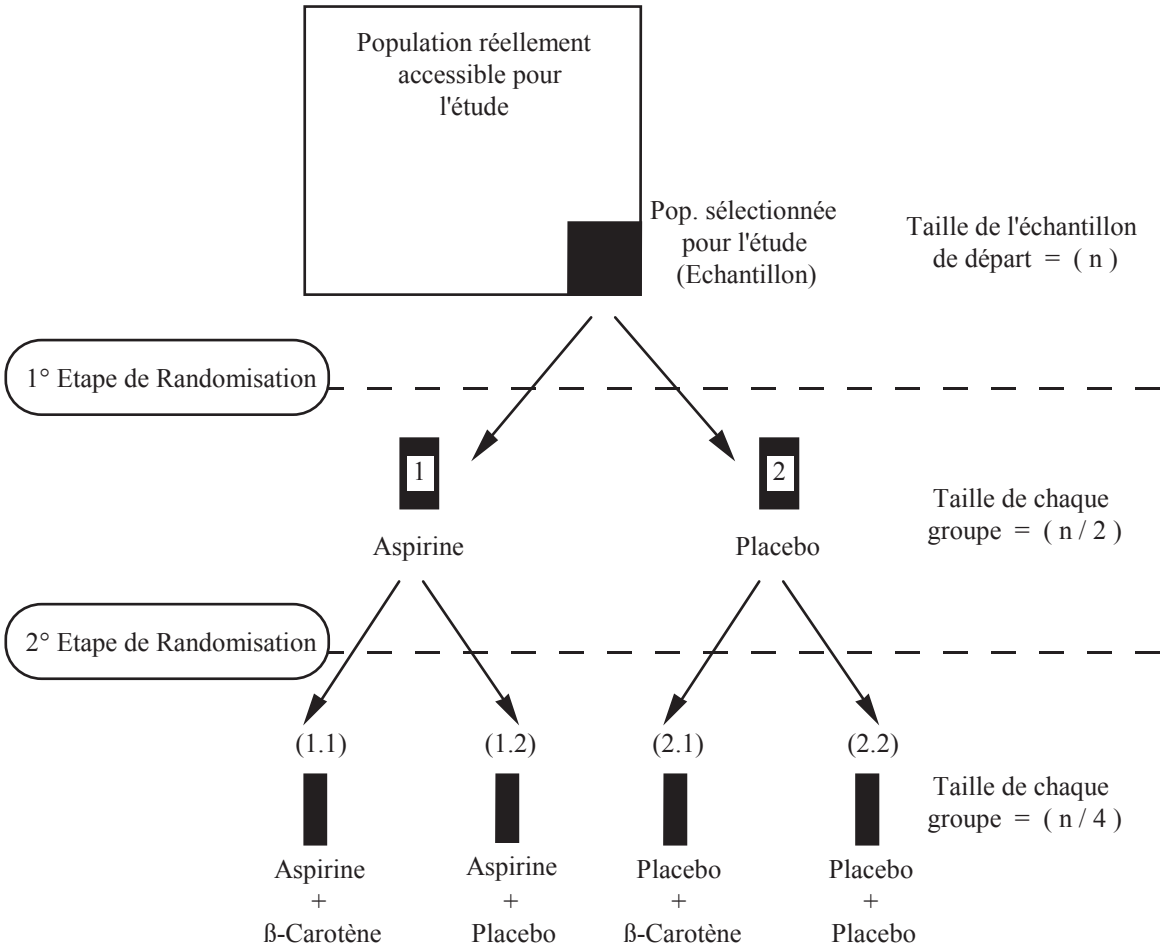
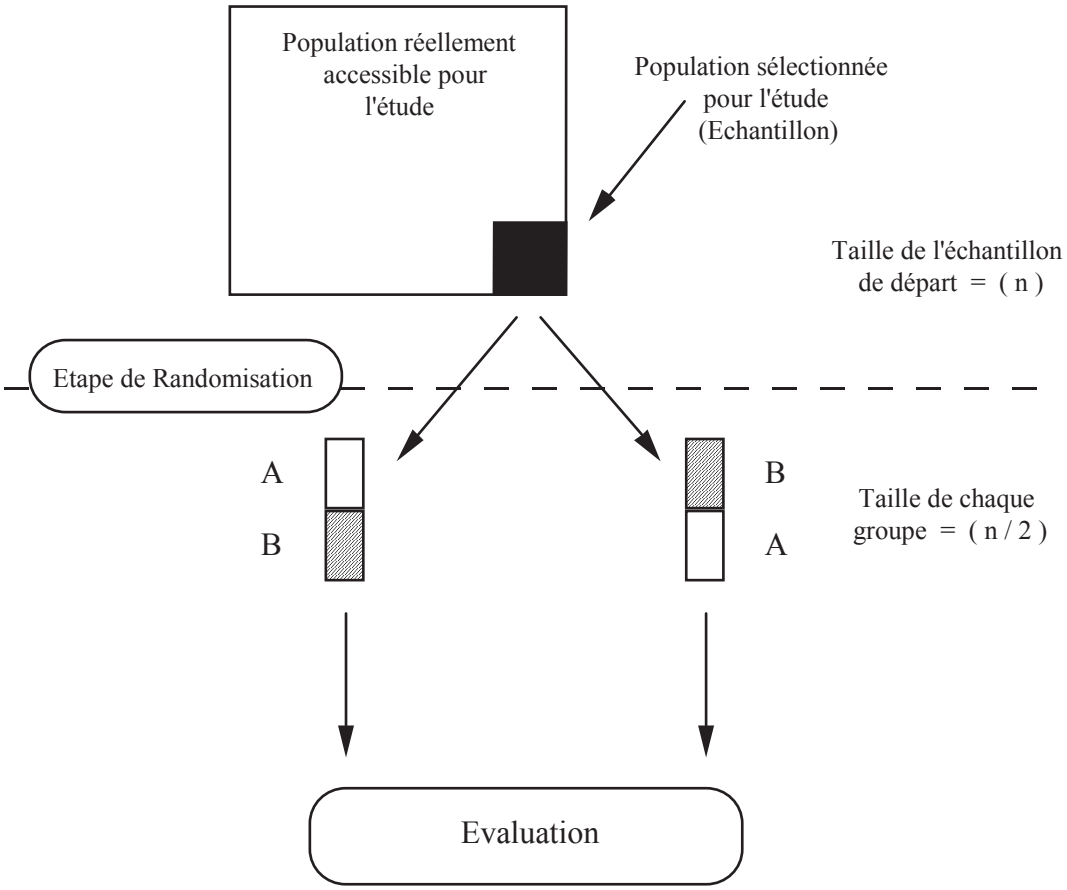


Figure 3 - Plan d'étude en permutations croisées



CHAPITRE VIII

LA MÉTA-ANALYSE

François Delahaye

La méta-analyse est un type d'étude particulier qui tranche avec les précédents : le chercheur n'a pas collecté lui-même les données utiles à l'étude, il n'a pas été en contact direct avec les sujets étudiés (ou leur dossier).

Méta en grec signifie « à travers ». La méta-analyse est donc littéralement une analyse d'analyse. Le chercheur collecte les études dont les données sont susceptibles d'être combinées.

La méta-analyse a fait son apparition dans la littérature médicale dans les années soixante-dix. Cette méthode, de plus en plus utilisée, constitue un moyen de contourner des difficultés logistiques insurmontables lorsqu'un très grand nombre de sujets est nécessaire pour mettre en évidence un effet.

La méta-analyse est donc un type d'étude à part entière qui nécessite, plus que les autres encore, une recherche bibliographique exhaustive.

Plan du chapitre

La meilleure revue critique et synthèse possible de l'information disponible est essentielle pour tous ceux qui doivent prendre des décisions, que ce soit face à un patient, pour établir une stratégie commune pour des groupes de patients similaires, ou pour formuler des hypothèses de recherche en médecine, en épidémiologie, ou en politique et administration de santé.

Inventée par les chercheurs des sciences de l'éducation et de psychologie au début des années 1970, la méta-analyse, terme créé par Glass en 1976, est une évaluation qualitative et quantitative de l'information médicale, et sa synthèse et son intégration structurée. « L'analyse des analyses » (méta- signifie « ce qui dépasse, englobe »), est « l'analyse statistique d'un grand nombre de données provenant de plusieurs analyses, afin d'en intégrer les résultats ».

Les mots « méta-analyse » et « revue systématique » sont souvent utilisés de façon synonyme alors qu'ils n'ont pas tout à fait le même sens. La revue systématique utilise une procédure structurée (par exemple pour la recherche de la littérature). La méta-analyse est une technique statistique de combinaison des résultats.

En médecine, les premières méta-analyses furent publiées au milieu des années 1970. La technique laissa sceptique jusqu'au milieu des années 1980, puis prit de l'ampleur avec l'équipe de Peto, à Oxford. En 1993, Iain Chalmers, un épidémiologiste d'Oxford, a fondé la Collaboration Cochrane (du nom d'Archie Cochrane, chercheur qui contribua grandement au développement de l'épidémiologie), une organisation internationale sans but lucratif, dont le but est de produire, disséminer et actualiser des méta-analyses dans le domaine médical (www.cochrane.org). Il y a désormais des standards de conduite et de rapport d'une méta-analyse (Cochrane, PRISMA : Preferred Reporting Items for Systematic reviews and Meta-Analyses).

La méta-analyse correspond à toute méthode systématique qui utilise des techniques statistiques pour combiner des données venant d'études indépendantes afin d'obtenir une estimation de l'effet global d'une variable sur un événement défini. On peut ainsi faire une méta-analyse d'études descriptives, d'études d'intervention, ou d'études validant des outils cliniques, par exemple des méthodes diagnostiques, mais le plus souvent les méta-analyses portent sur des essais thérapeutiques.

La méta-analyse permet :

- de trancher lors de conclusions discordantes ;
- d'augmenter la puissance pour les événements majeurs et les analyses de sous-groupes ;
- de rétrécir les bornes de la taille de l'effet (augmenter la précision) ;
- de répondre à des questions nouvelles.

Six raisons peuvent conduire à réaliser une méta-analyse :

- obtenir des estimations plus stables de l'effet d'un traitement ;
- aider à interpréter la « généralisabilité » des résultats ;
- conduire des analyses sur des sous-groupes ;
- aider aux demandes d'autorisations de mise sur le marché ;
- aider à la planification d'essais cliniques ;
- contrebalancer l'excès d'enthousiasme qui accompagne souvent l'introduction de nouvelles drogues.

I - MÉTA-ANALYSE QUALITATIVE OU QUANTITATIVE

Deux grandes approches sont possibles :

A - L'approche qualitative

Elle consiste à accorder une importance différente aux diverses études selon leur qualité méthodologique. Les études sont revues selon un ensemble de critères permettant de juger de la validité scientifique et des possibilités d'application clinique des résultats.

Le but d'une telle méta-analyse est de tirer des conclusions des études jugées méthodologiquement supérieures.

L'approche qualitative comporte des étapes fondamentales :

- formulation de la question ;
- recherche des études ;
- définition des critères de jugement de la crédibilité scientifique des études ;
- application de ces critères à chaque étude ;
- analyse de la relation entre la crédibilité scientifique d'une étude et ses conclusions.

Prenons l'exemple du BCG et de la prévention de la tuberculose. Le BCG est utilisé largement pour prévenir la tuberculose depuis plus de 70 ans, mais son efficacité est controversée. Cela, en partie au moins, du fait de résultats discordants des différents essais.

Clemens et coll. décrivent d'abord comment l'essai clinique idéal devrait être conduit, puis analysent la littérature disponible et comparent les méthodes de ces essais avec leurs résultats (tableau 1). Selon Clemens, « une démonstration adéquate de la détection non biaisée de la tuberculose n'était disponible que pour les trois essais rapportant une efficacité de 75% et plus ; dans la plupart des essais rapportant une efficacité basse, les intervalles de confiance étaient larges, ne pouvant exclure une efficacité grande, mais dans tous les essais rapportant une efficacité grande, les intervalles de confiance étaient étroits, excluant une efficacité basse ». La conclusion des auteurs est que le BCG peut apporter une protection, et que des biais ou une puissance statistique insuffisante peuvent avoir contribué à la discordance des résultats.

B - L'approche quantitative

Elle consiste en un résumé quantitatif des résultats des différentes études, de façon à créer une seule étude, large, à la puissance statistique plus grande.

L'approche quantitative doit elle aussi suivre plusieurs étapes, mais le plus souvent, on combine les deux approches, qualitative et quantitative, et les étapes en sont :

- formulation de la question ;
- recherche bibliographique ;
- élaboration de critères précisant les attributs (conditions cliniques, traitements, événements) qui vont être groupés et comparés ;
- classification et codification des études retenues ;
- définition des critères de jugement de la crédibilité scientifique des études ;
- étude de la qualité des études ;
- analyse statistique des données ;
- formulation des résultats ;
- analyses de sensibilité ;
- analyse de la relation entre la crédibilité scientifique d'une étude et ses conclusions ;

- interprétation des résultats et conclusions.

II - LES SEPT ÉTAPES DE LA MÉTA-ANALYSE (figure 1)

A - Objectifs

Bien entendu, l'objectif de la méta-analyse doit toujours être clairement précisé. Un protocole doit toujours être rédigé, et ce avant l'exécution de la méta-analyse. Précis et rigoureux, il doit en particulier préciser les hypothèses, et toutes les procédures utilisées, notamment celles de la recherche de la littérature, les critères de sélection des essais, la définition des événements, la technique d'analyse de l'hétérogénéité, et les méthodes statistiques.

B - Recherche de la littérature

Cette tâche colossale est fondamentale ! De l'exhaustivité de la recherche dépend en effet la qualité de la méta-analyse.

La recherche doit faire appel à plusieurs méthodes simultanément. La collecte est faite :

- bien sûr par repérage grâce aux banques bibliographiques, manuel (Index Medicus, Excerpta Medica…) ou automatisé (Medline, Cancerlit, Pascaline…), et par consultation des comptes rendus de congrès et des bibliographies des articles et livres sur le sujet ;
- mais aussi grâce aux discussions avec les collègues et les experts, à la consultation des laboratoires pharmaceutiques et des organismes gouvernementaux finançant des essais.

Alors qu'on pourrait les croire infaillibles, les recherches automatisées ne sont pas parfaites. Toutes les méthodes de recherche citées plus haut doivent être utilisées, et pas seulement les banques de données bibliographiques.

Un gros écueil de la méta-analyse, mais il est commun à toutes les sortes de revues de la littérature, est représenté par le biais de publication, qui correspond à la soumission préférentielle et à l'acceptation préférentielle des études montrant des résultats positifs. Il n'existe pas de solution parfaite encore à ce biais de publication. Certains ont proposé de calculer le nombre d'études qu'il faudrait pour changer les conclusions de la méta-analyse. Une autre réponse est de tenir à jour des registres de tous les essais en cours. Ainsi, on sait le devenir de tous les essais, y compris ceux qui ont été interrompus et ceux dont les résultats ont été négatifs. On utilise souvent une représentation sous forme d'une figure en entonnoir inversé (« funnel plot ») (figure 2).

Le choix des critères d'inclusion et d'exclusion des études dans la méta-analyse peut être fondé sur diverses variables : le plan d'étude (on limite souvent une méta-analyse aux essais randomisés), la taille de l'étude (on peut exiger un nombre minimum de sujets par groupe), la population d'étude, le type de groupes traités et de groupes témoins (une certaine dose…), l'événement étudié… Les critères, qui dépendent des objectifs de la méta-analyse, doivent être listés dans le protocole, avec la raison de leur choix.

Faut-il mettre toutes les études ? Cela augmente la représentativité des conclusions, mais diminue la validité statistique de la synthèse en incluant les études moins rigoureuses. Cette décision dépend de l'objectif de la méta-analyse, ainsi par exemple est-on beaucoup plus sévère dans la sélection des essais pour une méta-analyse faisant partie d'un dossier d'autorisation de mise sur le marché que pour une méta-analyse exploratoire.

La décision que les études sont suffisamment similaires pour que leurs résultats puissent être agrégés est subjective, et il est difficile de développer des critères universels garantissant la sélection appropriée des études.

Le rapport doit contenir la liste des études incluses et celle des études exclues, afin que le lecteur puisse savoir sur quoi repose la méta-analyse, et connaître les études réfutées, ainsi que la raison de l'exclusion.

C - Extraction des données de chaque étude

Il y a trois grands types de données utilisées dans les méta-analyses.

1 - Les données individuelles

Un des premiers exemples en est donné par Canner.

Canner analyse les six essais les plus importants de l'efficacité de l'aspirine dans la prévention secondaire de la mortalité après infarctus du myocarde. Aucun de ces essais ne met en évidence d'effet statistiquement significatif de l'aspirine. La méta-analyse de cinq de ces essais permet d'objectiver un effet bénéfique de l'aspirine ($p = 0,014$).

L'ajout d'AMIS (Aspirin Myocardial Infarction Study) change tout : dans cet essai, le plus gros et de loin, l'aspirine a un effet défavorable. Lorsque les six essais sont regroupés, le rapport de cotes, favorable à l'aspirine, passe de 0,76 (cinq essais) à 0,90 (six essais), et la signification statistique disparaît. Le test d'hétérogénéité est à la limite de la significativité.

Depuis, il y a eu de nombreuses méta-analyses sur données individuelles. Citons par exemple, en cardiologie, les travaux de la Prospective studies collaboration, de l'Antiplatelet Trialists' Collaboration, de l'Antithrombotic Trialists' Collaboration ou du groupe INDANA.

Comment expliquer cette hétérogénéité ?

Il n'y a pas de différences évidentes dans le plan d'étude, les doses utilisées, ou le temps écoulé entre l'infarctus et la thérapeutique. En revanche, il existe une mauvaise répartition des caractéristiques de base. Ayant accès aux données individuelles de trois des essais, Canner peut ajuster (régression linéaire multiple) selon ces caractéristiques de base. Le test d'hétérogénéité devient beaucoup moins significatif (0,22), et le test d'association (six essais) devient significatif (0,04), en faveur d'un effet favorable de l'aspirine.

Cette technique, beaucoup plus laborieuse, et qui nécessite l'acceptation des investigateurs de prêter leurs données, mais qui permet de répondre à beaucoup plus de questions, a connu un essor important ces dernières années.

2 - Les résumés tirés des publications

C'est actuellement à partir de ces publications que sont faites la grande majorité des méta-analyses. Mais souvent les données nécessaires manquent, et on conserve dans l'analyse les biais qui ont pu survenir par inclusion inadéquate de certains sujets randomisés, d'où l'intérêt de demander les données à l'expérimentateur.

3 - Les résumés par sous-groupes (notamment par sexe ou par âge, dans la même publication ou dans des publications ultérieures)

Ce processus d'extraction des données est long, ennuyeux, sujet à erreur, donc à biais. Pour s'en prémunir le plus possible, des bordereaux d'extraction des données doivent avoir été conçus, et ils doivent être remplis si possible par plusieurs personnes, avec réunions de consensus pour régler les désaccords.

Il y a différentes sortes d'événements intéressants :

- variable continue (tension artérielle, score de qualité de la vie) ;
- variable binaire (mortalité, complications) ;
- variable ordinale (stade tumoral) ;
- variable liée au temps (survie sans maladie).

Les analyses portent surtout sur des variables binaires. Quatre mesures de l'effet du traitement sont souvent utilisées : si Pc est la proportion d'événements dans le groupe contrôle et Pt la proportion d'événements dans le groupe traité, on peut mesurer l'effet sous forme de :

- différence absolue : Pc - Pt
- risque relatif : $\frac{Pt}{Pc}$ et rapport de cotes (odds ratio)
- réduction relative du risque : $\frac{Pc - Pt}{Pc}$
- nombre de sujets à traiter : $\frac{1}{Pc - Pt}$

Deux indices sont particulièrement parlants :

- la réduction relative du risque est la différence de risque entre les deux groupes, rapportée au risque dans le groupe témoin ; si la mortalité est de 10% dans le groupe témoin et de 5% dans le groupe traité, la réduction relative du risque est de 50% ; cependant, la présentation isolée de la réduction relative du risque est fallacieuse : entendre que la réduction relative du risque d'événement grâce à une intervention est de 50%, c'est souvent, inconsciemment, penser que l'intervention évite un événement sur deux ; il faut aussi un risque absolu : la réduction relative du risque est de 50% lorsque le risque absolu passe de 80% à 40%, une réduction cliniquement très importante ; la réduction relative du risque est aussi de 50% lorsque le risque absolu passe de 2 / 1 milliard à 1 / 1 milliard, une réduction cliniquement… infinitésimale ! ;
- le nombre de sujets à traiter pendant une certaine période de temps pour éviter un événement est l'inverse de la différence absolue de risque ; si la mortalité est de 10% dans le groupe témoin et de 5% dans le groupe traité, le nombre de sujets à traiter est de 20.

Un autre indice, très utilisé en psychologie, est la taille de l'effet, la différence entre la moyenne dans le groupe traité et celle dans le groupe témoin, divisée par l'écart-type dans le groupe témoin.

D - Évaluation de la qualité de chaque étude

Ce processus étant particulièrement subjectif, protocole et bordereaux d'évaluation de la qualité des études sont obligatoires, avec lecture par au moins deux personnes et réunions de règlement des désaccords, après préparation des articles en enlevant toute identification.

À l'issue de l'examen, un score de qualité est donné à chaque étude. Ce score peut être utilisé :

- soit comme seuil, d'inclusion ou d'exclusion d'une étude ;
- soit pour donner un poids respectif à chaque étude ;
- soit pour comparer le résultat de l'étude et son score de qualité.

E - Regroupement des données

Cette étape est celle qui distingue le plus la méta-analyse des autres techniques de synthèse de l'information. Mais avant de réaliser le regroupement proprement dit, il faut d'abord vérifier l'absence d'hétérogénéité.

1 - L'homogénéité

Une hypothèse sous-jacente lors de la combinaison de plusieurs études est que les différences entre les résultats des études sont dues au hasard seul, et donc que tous les résultats sont homogènes. Mais cette hypothèse doit être discutée. Si les variations ne sont pas dues au hasard seul, les techniques de regroupement des données sont plus compliquées, et éventuellement déraisonnables.

Une première étape de l'analyse de l'hétérogénéité consiste dans une étude graphique (figure 3). Il existe bien entendu des techniques statistiques plus formelles pour tester l'homogénéité, en particulier un $\chi 2$ de Mantel-Haenszel ou des techniques de régression. Mais leur puissance est limitée, et l'analyse graphique associée est particulièrement utile.

La non uniformité peut être due à une certaine caractéristique, par exemple la dose utilisée. Un nouveau graphique séparant les essais en différents groupes selon la dose, permet de retrouver une homogénéité au sein de chaque groupe (figure 4).

2 - Les méthodes statistiques

Elles peuvent être regroupées en quatre catégories :

- méthodes combinant les p ;
- méthodes combinant les valeurs de tests statistiques (z, t) : les plus anciennes et les plus simples, elles sont comme les précédentes très limitées ;
- méthodes fondées sur un modèle :
    - pour un événement binaire, modèle binomial : logarithme du rapport de cotes, différence des taux, méthode de Mantel-Haenszel, de Peto ou de Cochran ; ces méthodes ont plusieurs avantages : les événements sont comparés à l'intérieur de chaque essai, ce qui augmente la précision du résultat global ; la différence pour chaque taux d'événement est pondérée par sa variance, les essais ayant les événements les plus stables (généralement les essais de plus grande taille) sont ceux qui ont le plus d'influence ; l'agrégation des « Observés – Espérés » fournit une estimation globale en plus du test statistique ;
    - pour une variable quantitative : analyse de variance ;
- méthodes de modélisation (régression linéaire multiple, régression logistique).

Il est conseillé d'utiliser plusieurs techniques (les résultats sont « robustes »), de choisir des techniques qui donnent un poids à chaque essai, autorisant une définition raisonnable du modèle sous-jacent et permettant de tester l'hétérogénéité.

F - Analyses de sensibilité

Il faut se demander quelle est la sensibilité des résultats de la méta-analyse à la façon de faire cette méta-analyse. En d'autres termes, on peut faire des analyses de sensibilité. Par exemple, les résultats sont-ils différents si on inclut les essais randomisés et non randomisés au lieu de n'inclure que les essais randomisés ?

Les études peuvent être regroupées selon des caractéristiques des différents groupes de patients ou du plan d'étude (allocation aléatoire ou non aléatoire, dose du médicament actif…), de façon à déterminer l'influence de ces caractéristiques sur les résultats de la méta-analyse.

Les analyses de sous-groupes au sein d'un essai clinique posent plusieurs problèmes : comparaisons multiples, interprétation erronée des différences, interactions. Un problème supplémentaire pour les méta-analyses est représenté par l'hétérogénéité des caractéristiques des diverses études faisant l'objet de la méta-analyse. Si la puissance des tests est augmentée par l'augmentation du nombre de sujets grâce au regroupement, il faut cependant rester très prudent dans la réalisation, et l'interprétation, des analyses de sensibilité.

G - Présentation des résultats

Après que les objectifs et les méthodes, notamment les techniques statistiques et les procédures de contrôle de qualité, ont été précisés, les résultats sont présentés sous forme de tables, mais aussi très souvent grâce à des figures représentant pour chaque essai l'estimation de l'effet et son intervalle de confiance, puis les mêmes données pour le total (« forest plot ») (figure 5).

Finalement on conclut, en discutant les résultats de la méta-analyse selon le choix des études, leur qualité, l'homogénéité. La qualité et les limitations de la méta-analyse doivent aussi être discutées, et la portée des résultats évaluée.

III - ÉVALUATION DE LA QUALITÉ DE LA MÉTA-ANALYSE

Le lecteur d'une méta-analyse doit se poser plusieurs questions avant d'adopter les conclusions du travail :

- l'objectif est-il clairement précisé ?
- y-a-t-il évidence d'un protocole de travail ?
- les techniques de recherche de la littérature sont-elles précisées ? Le problème d'un biais de publication est-il envisagé ?
- les critères d'inclusion et d'exclusion sont-ils spécifiés, les articles inclus et exclus sont-ils listés, les raisons de l'exclusion sont-elles données ?
- les traitements sont-ils suffisamment similaires pour permettre de regrouper les résultats ? De même pour les groupes témoins ?
- les tests d'homogénéité, graphiques et statistiques, sont-ils présentés ?
- la technique statistique de regroupement des données est-elle correcte ?
- des analyses de sensibilité ont-elles été faites ?
- des conclusions quant à l'efficacité du traitement et pour de futures recherches sont-elles tirées ?

IV - AVANTAGES ET LIMITES DE LA MÉTA-ANALYSE

On assiste actuellement à une floraison de méta-analyses, et plusieurs équipes en ont précisé la méthodologie.

Cette technique a en effet plusieurs avantages :

- elle permet d'estimer l'importance d'un effet ;
- elle augmente la puissance statistique ;
- elle augmente la « généralisabilité » ;
- elle oblige à la rigueur dans les méthodes, la lecture, le recueil des données ;
- elle diminue la part du subjectif.

Cependant de nombreux auteurs la critiquent :

- elle ignore la qualité des études : on a vu que des outils existent, mais il est vrai que des améliorations sont possibles ;
- il est illogique de combiner les résultats d'études utilisant des patients différents, des techniques de mesure différentes, et réalisées à des moments différents : le méta-analyste doit présenter un résumé des caractéristiques pour chaque étude, mettre en exergue les différences, tester l'hétérogénéité, et discuter ses résultats et la « généralisabilité » en fonction de ces divers éléments ;
- il y a un biais de publication potentiel : mais cela est le lot de toute synthèse de l'information ;
- la validité de la méta-analyse dépend du degré de complétude et de précision de l'information rapportée dans les différentes études : là encore, cela ne lui pas spécifique ;
- la validité des techniques statistiques de la méta-analyse doit être établie.

Malgré les critiques, les limitations, cette technique de la méta-analyse, lorsqu'elle est prudente et bien réalisée, apporte des informations supplémentaires, permettant d'améliorer la qualité de notre réponse à une question particulière.

Références

Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H. Considerations for the meta-analysis of randomized clinical trials. Summary of a panel discussion. Controlled Clin Trials 1989;10:254-81

Canner PL. An overview of six clinical trials of aspirin in coronary heart disease. Stat Med 1987;6:255-63

Clemens JD, Chuong JJH, Feinstein AR. The BCG controversy. A methodological and statistical reappraisal. JAMA 1983;249:2362-9

Colton T, Freedman LS, Johnson AL, eds. Proceedings of the workshop on methodologic issues in overviews of randomized clinical trials. Stat Med 1987;6:217-410

L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med 1987;107:224 -33

Leucht S, Kissling W, Davis JM. How to read and understand and use systematic reviews and meta-analyses. Acta Psychiatr Scand 2009;119:443-50

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65-94

Moher D, Tricco AC. Issues related to the conduct of systematic reviews: a focus on the nutrition field. Am J Clin Nutr 2008;88:1191-9

Tseng TY, Dahm P, Poolman RW, Preminger GM, Canales BJ, Montori VM. How to use a systematic literature review and meta-analysis. J Urol 2008;180:1249-56

Série d'articles sur la méta-analyse

Mulrow CD, Cook DJ, Davidoff F. Systematic reviews: Critical links in the great chain of evidence. Ann Intern Med 1997;126:389-91

Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: Synthesis of best evidence for clinical decisions. Ann Intern Med 1997;126:376-80

Hunt DL, MacKibbon KA. Locating and appraising systematic reviews. Ann Intern Med 1997;126:532-8

MacQuay HJ, Moore. Using numerical results from systematic reviews in clinical practice. Ann Intern Med 1997;126:712-20

Badgett RG, Okeefe M, Henderson MC. Using systematic reviews in clinical education. Ann Intern Med 1997;126:886-91

Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. Ann Intern Med 1997;127:37-42

Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. Ann Intern Med 1997;127:210-6

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997;127:380-7

Livres

Cucherat M. Méta-analyse des essais thérapeutiques. Éditions Masson, 1997

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Éditions Wiley, 2009

Tableau 1 - Protection contre les biais et justesse de la précision statistique dans 8 essais majeurs du BCG

Essai	Protection adéquate contre les biais de				Précision statistique adéquate	Efficacité protectrice observée
	susceptibilité	surveillance	méthode diagnostique	interprétation		
Indiens d'Amérique	Oui	Oui	Oui	Oui	Oui	80%
Angleterre	Oui	Oui	Oui	Oui	Oui	76%
Chicago	Probable	Oui	Oui	Oui	Oui	75%
Porto Rico	Oui	Non	Non	Non	Non	29%
Madanapalle	Équivoque	Non	Non	Probable	Non	20%
Georgia-Alabama	Oui	Non	Non	Équivoque	Non	6%
Chingleput	Probable	Non	Non	Oui	Oui	-32%
Georgia	Oui	Non	Non	Non	Non	-56%

Figure 1 - Étapes de la méta-analyse

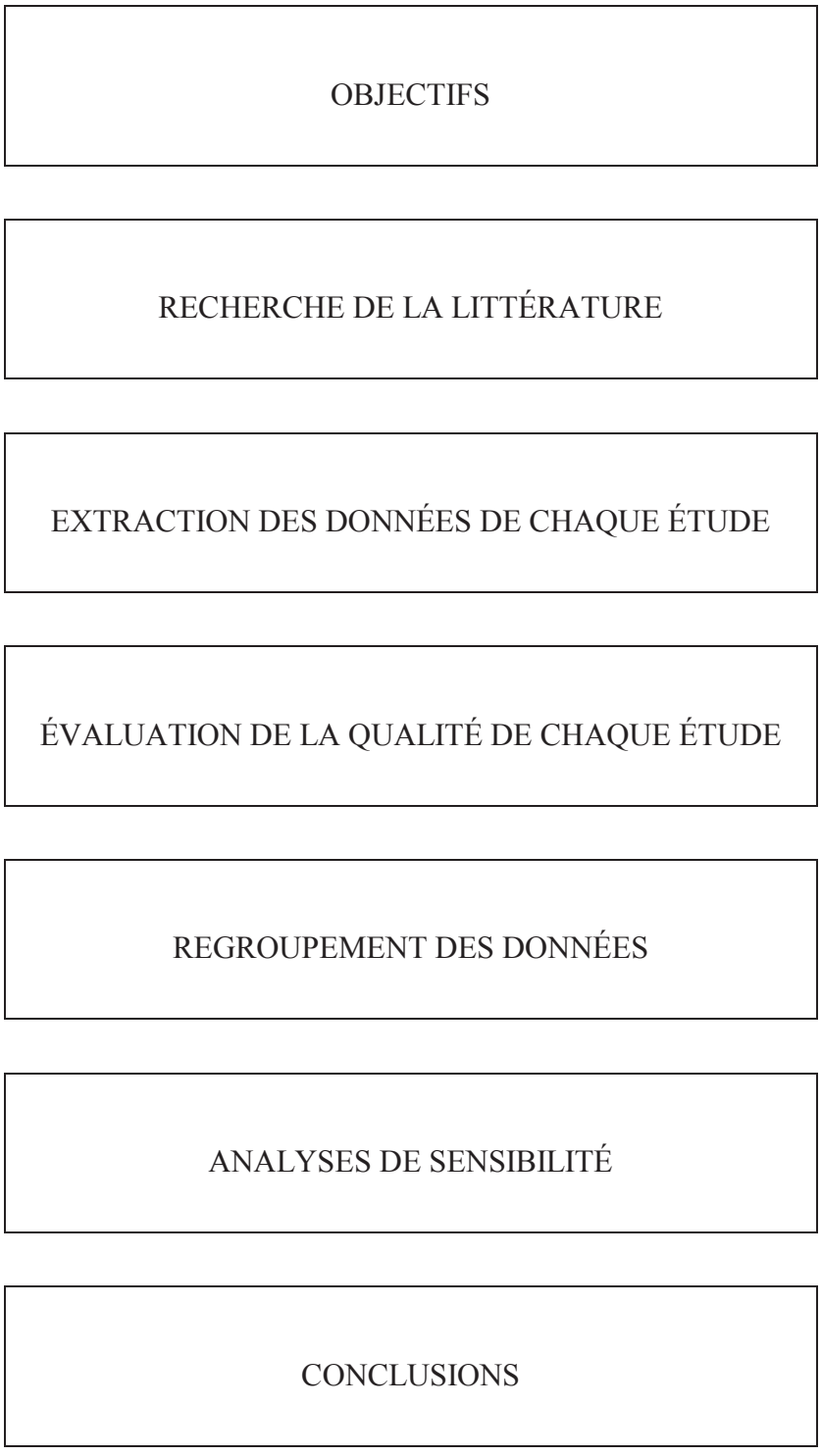


Figure 2 - Figure en entonnoir inversé (« funnel plot ») à la recherche d'un biais de publication

Chaque essai est présenté sous forme d'un point à l'intersection de la taille de l'effet et du nombre de sujets de l'essai. Une allure symétrique suggère qu'il ne manque pas d'essai, une allure asymétrique évoque un biais de publication.

Figure 3 - Analyse graphique de l'hétérogénéité

Les carrés représentent pour chaque essai le taux d'événements dans le groupe témoin et dans le groupe traité. Dans la figure 3A, les carrés sont répartis de façon assez homogène ; dans la figure 3B, la répartition des carrés est hétérogène : les carrés sont au-dessous de la diagonale quand les taux d'événements sont bas et au-dessus de la diagonale quand les taux d'événements sont hauts.

Figure 4 - Analyse graphique de l'hétérogénéité

Les carrés représentent pour chaque essai le taux d'événements dans le groupe témoin et dans le groupe traité. Dans la figure 3B, l'hétérogénéité est due à la dose. L'homogénéité apparaît lorsqu'on distingue les essais selon la dose, faible ou forte.

Figure 5 - Représentation graphique des résultats de la méta-analyse

L'indice utilisé est le rapport de cotes (odds ratio), représenté par le gros carré, la ligne horizontale figurant l'intervalle de confiance à 95%, avec ses bornes (petits carrés). Chaque ligne représente les résultats d'un essai, la dernière figurant les résultats de la méta-analyse.

CHAPITRE IX

LES ETUDES DE STRATEGIES DIAGNOSTIQUES

Muriel Rabilloud, René Ecochard, Gilles Landrivon

Les examens cliniques ou paracliniques sont utilisés:

1 - Soit pour savoir si le patient est porteur d'une affection (par exemple, l'échographie pour la lithiase vésiculaire), afin, si le test est positif, d'envisager un traitement (ici, la chirurgie).

2 - Soit pour quantifier la valeur d'un paramètre (par exemple, la digitalinémie au cours d'un traitement digitalique) afin d'adapter une thérapeutique (ici, la posologie du tonicardiaque).

3 - Soit pour visualiser des structures normales ou pathologiques (par exemple, le réseau vasculaire avant une intervention chirurgicale intracrânienne, ou une lithiase du cholédoque au cours d'un cathétérisme rétrograde de la papille).

4 - Soit enfin pour déterminer l'ampleur d'une atteinte, afin d'établir un pronostic sans qu'il en découle de décision thérapeutique.

La plus grande partie de ce chapitre concerne la situation 1. Les situations 2 et 3 sont traitées dans le chapitre XIV sur les instruments de mesure. La situation 4 se rapproche de la situation 1 (notions de sensibilité et de spécificité, …) mais s'en écarte par l'absence de décision à l'issue du résultat.

Plan du chapitre

1 – Les valeurs prédictives et le théorème de Bayes

2 – Les probabilités pré et post test et les ratios de vraisemblance

III - QUELQUES ASPECTS PARTICULIERS AUX ETUDES D'EVALUATION D'UN TEST DIAGNOSTIQUE

A – Quelques biais spécifiques aux études d'évaluation des tests diagnostiques

1 – Biais de vérification

2 – Biais lié à l'utilisation d'un gold standard imparfait

B – L'intervalle de confiance et le calcul du nombre de sujets nécessaire

1 – Intervalle de confiance de la sensibilité et de la spécificité

2 – Calcul du nombre de sujets nécessaire

IV - CONCLUSION

I - LE TEST DIAGNOSTIQUE REPLACE DANS SON CONTEXTE

Dans ce paragraphe, nous allons planter le décor, en introduisant les termes de probabilité pré-test, sensibilité, spécificité, probabilité post-test, seuil de traitement et utilité.

Tout examen trouve sa place dans une histoire, histoire qui commence par un symptôme, ou un examen de dépistage réalisé en l'absence de signe d'appel.

En amont du test diagnostique il y a un contexte clinique (âge, sexe, antécédents, symptômes déjà présents, éventuellement résultats d'autres examens) qui permet d'établir une probabilité d'existence de la maladie (avant de réaliser le test diagnostique étudié). Cette probabilité est appelée **probabilité pré-test**.

Un résultat positif du test va changer votre avis sur l'état du patient, la probabilité d'existence de la maladie étant plus élevée dans ce cas, moins élevée devant un résultat négatif. La probabilité que le patient soit porteur de la maladie connaissant le résultat du test est appelée **probabilité post-test**.

Parfois, la probabilité de la maladie sachant que le test est positif est de 100 % (égale à 1). C'est le cas si le test n'a aucun faux positif, par exemple l'histologie avec présence de tissu néoplasique sur une biopsie. Cet examen a une **spécificité** de 100 % (probabilité que le test soit négatif en l'absence de cancer).

Parfois, la probabilité de la maladie sachant que le test est négatif est nulle (égale à 0). Le test a complètement éliminé l'hypothèse d'atteinte par l'affection recherchée car il n'y a aucun faux négatif. C'est le cas par exemple du scanner abdominal pour le diagnostic de masse kystique rénale, ce dernier examen ayant une **sensibilité** de 100 % (probabilité que le test soit négatif en présence de la maladie).

Le plus souvent, la probabilité post-test est différente de 0 ou de 100%. Elle dépend du contexte clinique (probabilité pré-test) et de la qualité du test. Si le test est très spécifique et qu'il revient positif, l'information apportée par le test pour affirmer la présence de la maladie est importante et entraîne une augmentation importante de la probabilité d'avoir la maladie.

Il se peut que le test, qu'il soit positif ou négatif, ne change pas la décision thérapeutique. Quelle serait alors son **utilité** dans la situation 1 définie dans l'introduction du chapitre (savoir si le patient est porteur d'une affection, afin d'envisager un traitement si le test est positif) ?

En effet, pour chaque action thérapeutique ou diagnostique invasive, le clinicien a un seuil de probabilité en dessous duquel il s'abstient. On ne réalise pas une biopsie mammaire pour des images banales de mammographie en l'absence d'anomalie à la palpation. Ce seuil est le plus souvent appelé **seuil de traitement**. Il dépend du bénéfice global que l'on attend de l'intervention. Chaque geste thérapeutique ou diagnostique invasif a un bénéfice global, résultat de l'équilibre entre l'amélioration potentielle de l'état de santé et les effets secondaires possibles.

La mise sous anticoagulant en cas de suspicion d'embolie pulmonaire est une décision dans laquelle interviennent le risque d'hémorragie et le bénéfice apporté par le traitement. Le seuil de traitement est bas dans le cas de l'embolie pulmonaire. Dès qu'il y a une probabilité de 10 à 20 % d'embolie pulmonaire (voire moins) la décision est prise d'immobiliser le patient et de le décoaguler en attendant la réalisation des examens complémentaires. Ce seuil bas est dû d'une

part à la gravité des complications évitées par le traitement et d'autre part à la relative sécurité d'un traitement anticoagulant bien conduit.

Au contraire, lorsque le geste est lourd de conséquences, telle une gastrectomie totale ou une amputation, on attend un niveau élevé de probabilité d'existence de la maladie (par exemple, gastrite de Ménétrier, tumeur osseuse maligne, …) pour intervenir. Si l'incertitude demeure, on préfère, en effet, différer le geste et reprendre les examens complémentaires.

Si la probabilité de maladie dépasse largement le seuil de traitement, ce dernier est entrepris sans réaliser d'autres examens paracliniques, qui n'ont, en effet, aucune chance de faire changer l'indication du traitement. Un résultat négatif serait étiqueté faux négatif, le reste du contexte étant trop accusateur pour que cet examen puisse faire changer le diagnostic. Face à un cordon sous-cutané rouge et douloureux, la phlébite superficielle est affirmée et le traitement est mis en place. Une phlébographie normale aurait l'intérêt d'éliminer une atteinte profonde mais ne remettrait pas en cause l'atteinte superficielle qui serait donc traitée. Le test concerné (ici, la phlébographie) a donc comme intérêt de montrer l'extension de l'atteinte, non de faire évoluer la probabilité de phlébite superficielle. En effet, négative ou positive, elle n'a pas la possibilité d'infléchir le diagnostic suffisamment pour remettre en cause le traitement. Sa sensibilité est insuffisante pour qu'un résultat négatif abaisse la probabilité post-test en dessous du seuil de traitement.

Il est de coutume (et exact) de dire qu'un test diagnostique ne doit être réalisé que s'il a une chance de faire passer la probabilité de maladie de "l'autre côté" du seuil de traitement, c'est-à-dire de changer la décision. Devant un bilan hormonal évocateur d'insuffisance surrénale, on ne réalise pas de test à l'eau. Celui-ci n'a en effet aucune chance de faire changer de diagnostic car il a trop de faux négatifs. A l'inverse, dans le suivi d'un patient en neurologie, la percussion du tendon d'Achille sera utilisée, la perte du réflexe à ce niveau ayant une spécificité suffisante pour que sa positivité déclenche un bilan complémentaire.

II – L'EVALUATION D'UN TEST DIAGNOSTIQUE

A – Les études d'évaluation d'un test diagnostique

1 – Les différentes phases d'évaluation

Comme pour l'évaluation de l'efficacité d'un nouveau médicament, il est possible de définir 3 phases dans l'évaluation d'un nouveau test diagnostique.

La première phase, appelée aussi **phase exploratoire**, correspond à la phase précoce d'évaluation d'un nouveau test. L'objectif est de savoir si ce test peut avoir un intérêt diagnostique. Il s'agit par exemple de vérifier qu'un nouveau biomarqueur a une valeur en moyenne plus élevée chez les malades que chez les non-malades et qu'il fait mieux que le simple hasard pour prédire l'existence de la maladie. A ce stade, les études réalisées doivent permettre d'obtenir une réponse rapide pour décider de poursuivre l'évaluation ou de passer à autre chose.

La deuxième phase, appelée aussi **phase de challenge**, a pour objectif de mesurer les performances diagnostiques d'un nouveau test dans différents sous-groupes de malades et de non-malades. Les performances diagnostiques d'un test sont quantifiées par sa sensibilité et sa spécificité pour les tests ayant une réponse dichotomique, ou par les sensibilités et spécificités associées aux différents seuils de positivité pour un test ayant une réponse

ordinale ou continue. On dit classiquement que la sensibilité et la spécificité sont les qualités intrinsèques d'un test car elles ne font pas intervenir la prévalence de la maladie. La sensibilité est estimée chez les malades et la spécificité chez les non-malades. En revanche, elles dépendent souvent des caractéristiques des malades ou des non-malades. Par exemple, la sensibilité de la mammographie pour le diagnostic de cancer du sein dépend de la taille de la tumeur. Elle est plus faible dans une population de femmes dépistées que dans une population de femmes venant en consultation spécialisée à un stade plus avancé de la maladie. Au cours de cette phase d'évaluation, le nouveau test peut également être comparé aux autres tests existants.

La troisième phase, appelée aussi **phase clinique,** a pour objectif de mesurer les performances diagnostiques d'un nouveau test et de le comparer aux autres tests dans la population ciblée. Cela implique que l'étude porte sur un échantillon représentatif de la population dans laquelle le test va être utilisé. Pour les tests nécessitant une interprétation par un lecteur tels que les examens d'imagerie médicale, il est également nécessaire de réaliser l'étude avec un échantillon représentatif des médecins qui seront amenés à lire l'examen. C'est également à cette phase précédant l'utilisation du test en pratique clinique, que le seuil de positivité et l'impact du choix du seuil sur la sensibilité et sur la spécificité sont étudiés pour les tests avec une réponse ordinale ou continue.

2 – Les différents types d'étude

Les principaux types d'étude que l'on retrouve dans le domaine de l'évaluation des tests diagnostiques sont des études de type cas-témoins, des études de type cohorte et des essais cliniques randomisés.

Les études de type cas-témoins

Ces études sont appelées études de type cas-témoins car lorsque les sujets entrent dans l'étude, leur statut malade ou non malade est connu. Elles reposent sur la constitution d'un échantillon de sujets dont on sait qu'ils ont la maladie et de façon indépendante d'un échantillon de sujets dont on sait qu'ils n'ont pas la maladie. Le test à évaluer est ensuite mesuré dans le groupe des sujets malades et dans le groupe des sujets non malades. Ce type d'étude est utilisé à la phase exploratoire de l'évaluation d'un nouveau test. Les sujets inclus dans l'échantillon de malades sont souvent à un stade assez avancé de la maladie, alors que les sujets inclus dans l'échantillon de non-malades sont souvent des sujets sains qui n'ont aucune pathologie pouvant mimer la maladie que l'on cherche à diagnostiquer. Cela aboutit souvent à une surestimation des performances diagnostiques du test à évaluer. Ce type d'étude est également utilisé à la phase de challenge car il permet de constituer des groupes de sujets malades à différents stades de la maladie et des groupes de sujets non malades avec des caractéristiques différentes par exemple en termes d'âge ou de comorbidités.

Les études de type cohorte

Ces études sont appelées études de type cohorte car lorsque les sujets sont inclus dans l'étude, leur statut malade ou non-malade n'est pas connu. Un échantillon représentatif de la population dans laquelle le test va être utilisé est constitué. Les sujets inclus dans l'étude ont tous le test à évaluer et leur statut malade ou non-malade est déterminé de façon indépendante du résultat du test. La détermination du statut malade ou non-malade nécessite de disposer d'un test de référence parfait appelé *gold standard* (règle d'or). L'étude CASS [1] est un exemple d'étude de type cohorte. Dans cette étude, un échantillon de 1465 hommes pour

lesquels il existe une suspicion de coronaropathie a été constitué. L'objectif de l'étude était d'évaluer les performances de l'épreuve d'effort et de la douleur thoracique recherchée à l'interrogatoire pour faire le diagnostic de coronaropathie. Tous les sujets inclus dans l'étude ont eu, outre les 2 tests à évaluer, une coronarographie permettant de les classer dans le groupe des sujets ayant une coronaropathie ou dans le groupe de sujets n'ayant pas de coronaropathie. Ce type d'étude est surtout utilisé à la phase clinique de l'évaluation d'un test. A ce stade de l'évaluation, il est recommandé de privilégier les études multicentriques pour augmenter la représentativité de l'échantillon étudié.

Les essais cliniques randomisés

Lorsqu'il n'existe pas de *gold standard* parfait, l'évaluation d'un nouveau test peut se faire par un essai clinique randomisé avec un bras correspondant à la stratégie diagnostique et thérapeutique habituelle et un bras qui intègre le nouveau test dans la stratégie diagnostique et thérapeutique. Dans ce type d'étude, le critère de résultat est un critère clinique. Le test sera jugé performant si le résultat clinique est significativement meilleur dans le bras incluant le nouveau test. Ce type d'étude permet également d'évaluer l'impact de l'introduction du test dans la stratégie diagnostique et thérapeutique.

B – L'évaluation des performances d'un test diagnostique

1 – Sensibilité et spécificité

La sensibilité et la spécificité d'un test sont des probabilités conditionnelles. La sensibilité est la probabilité que le test soit positif (en faveur de la maladie) sachant que le sujet est malade. Il s'agit de la capacité du test à identifier les malades. La spécificité est la probabilité que le test soit négatif (pas en faveur de la maladie) sachant que le sujet n'a pas la maladie. Il s'agit de la capacité du test à identifier les non-malades.

Leur estimation peut être obtenue à partir des résultats d'une étude de type cas-témoins ou de type cohorte présentés sous forme d'un tableau 2×2 (Tableau 1). Il y a quatre résultats possibles en fonction du résultat du test et du statut vis-à-vis de la maladie. Le résultat du test est positif et le sujet est malade, il s'agit d'un vrai positif (VP). Le résultat du test est négatif et le sujet est malade, il s'agit d'un faux négatif (FN). Le résultat du test est négatif et le sujet est non-malade, il s'agit d'un vrai négatif (VN). Le résultat du test est positif et le sujet est non-malade, il s'agit d'un faux positif (FP).

La sensibilité est estimée chez les malades par la proportion de tests positifs : $\frac{VP}{VP+FN}$

La spécificité est estimée chez les non-malades par la proportion de tests négatifs : $\frac{VN}{VN+FP}$

Il s'agit des valeurs les plus probables de sensibilité et spécificité du test compte tenu des données observées (estimations du maximum de vraisemblance). Elles sont obtenues par une lecture verticale du tableau 2×2.

Les résultats de l'étude CASS ont permis d'estimer la sensibilité et la spécificité de la douleur thoracique pour faire le diagnostic de coronaropathie dans une population de sujets à risque (Tableau 2).

La sensibilité de la douleur thoracique était estimée à : $\frac{969}{1023} = 94{,}7\%$

La douleur thoracique était présente (positive) chez environ 95% des patients porteurs d'une coronaropathie.

La spécificité de la douleur thoracique était estimée à : $\frac{197}{442} = 44{,}6\%$

La douleur thoracique était absente (négative) chez environ 45% des sujets n'ayant pas de coronaropathie.

2 – La courbe ROC et l'aire sous la courbe ROC

On peut distinguer **trois types de réponse pour les tests diagnostiques**. La réponse peut être **dichotomique** comme pour la douleur thoracique. La réponse peut être **ordinale** ou **quantitative continue**. Un exemple de test avec une réponse ordinale est le score BIRADS développé par le collège américain de radiologie. Il s'agit d'un score à 5 niveaux qui permet de classer les mammographies en fonction du degré de suspicion de cancer. Les marqueurs biologiques tels que les PSA pour le diagnostic de cancer de la prostate, sont des exemples de tests avec une réponse quantitative continue.

A la phase précoce de l'évaluation des tests avec une réponse dichotomique la performance est mesurée par la sensibilité et la spécificité. Avec une réponse ordinale ou quantitative continue, il n'est pas possible de résumer la performance diagnostique par l'estimation d'une sensibilité et d'une spécificité. Il existe autant de sensibilités et de spécificités que de seuils de positivité possibles. La courbe ROC (de l'anglais Receiver Operator Characteristic) permet de représenter la relation entre la probabilité que le test soit positif chez les malades (sensibilité) et la probabilité que le test soit positif chez les non-malades (1-spécificité).

L'étude de Hall FM et al. [2] portait sur 400 femmes ayant eu une biopsie du sein pour suspicion de cancer à la mammographie et une palpation normale. Parmi ces femmes, 119 avaient un cancer du sein. Les auteurs ont relu les mammographies sans avoir connaissance du résultat de la biopsie et les ont classées selon le degré de suspicion de cancer (Tableau 3). Selon le seuil de positivité choisi pour classer les mammographies comme positives, la sensibilité et la spécificité évoluent en sens inverse. Si seules les mammographies avec un haut degré de suspicion de cancer sont classées comme positives, la sensibilité est faible car il y a beaucoup de faux négatifs. Au contraire, la spécificité est élevée car il y a peu de faux positifs. Plus le seuil de positivité choisi est bas, meilleure est la sensibilité et moins bonne est la spécificité. A partir des données observées (Figure 1), il est possible d'estimer la sensibilité (Se) et la spécificité (Sp) de la mammographie pour chaque seuil dans une population de femmes ayant une suspicion de cancer.

- Mammographie considérée comme positive pour les suspicions hautes de cancer et comme négative pour les suspicions moyennes, légères et minimes:

(a) $\mathrm{Se} = \frac{47}{119} = 0{,}39 \qquad 1 - \mathrm{Sp} = 1 - \left(\frac{281 - 6}{281}\right) = 1 - \frac{275}{281} = 1 - 0{,}979 \approx 0{,}02$

- Mammographie considérée comme positive pour les suspicions hautes ou moyennes

(b) $\mathrm{Se} = \frac{104}{119} = 0{,}87$ $\qquad 1\text{-}\mathrm{Sp} = 1 - \left(\frac{281-(6+117)}{281}\right) = 1\text{-}\frac{158}{281} = 1\text{-}0{,}56 = 0{,}44$

- Mammographie considérée comme positive pour les suspicions hautes, moyennes ou légères

(c) $\mathrm{Se} = \frac{113}{119} = 0{,}95$ $\qquad 1\text{-}\mathrm{Sp} = 1\text{-}\left(\frac{281\text{-}(6+117+37)}{281}\right) = 1\text{-}\frac{121}{281} = 1-0{,}43 = 0{,}57$

Si toutes les mammographies sont considérées comme positives, toutes les femmes ayant un cancer sont bien classées (Se=1), mais toutes les femmes n'ayant pas de cancer sont faussement positives (Sp=0). A l'autre extrême si toutes les mammographies sont considérées comme négatives, toutes les femmes n'ayant pas de cancer sont bien classées (Sp=1), mais toutes les femmes ayant un cancer sont faussement négatives (Se=0). A partir de chaque couple (sensibilité, 1- spécificité) estimé pour les différents seuils observés, il est possible en reliant les points de construire la courbe ROC empirique (Figure 1) permettant de représenter la performance diagnostique globale de la mammographie. Si l'on considère que la mesure du degré de suspicion de cancer est un continuum entre la normalité et le cancer certain, la courbe en pointillé représente la courbe ROC de la mesure latente quantitative continue d'où est issue la réponse ordinale observée.

Plus la courbe ROC se rapproche de l'angle supérieur gauche correspondant à une sensibilité de 1 et une spécificité de 1, meilleure est la performance globale du test. Au maximum un test quantitatif dont la courbe ROC passe par le point de sensibilité 1 et spécificité 1, est un *gold standard* parfait. Dans ce cas, les distributions des valeurs chez les malades et les non-malades ne se recouvrent pas et tous les sujets sont bien classés. Cet idéal rarement atteint est symbolisé par le soleil sur la figure 1.

Un test diagnostique dont la courbe ROC est sur la diagonale, est un test pour lequel la probabilité d'avoir une réponse positive chez les malades est égale à la probabilité d'avoir une réponse positive chez les non-malades quel que soit le seuil de positivité. Il ne fait pas mieux que le hasard. La pièce de monnaie symbolise la situation que l'on aurait en jetant une pièce et en décidant que le test est positif chaque fois que l'on tombe sur face et négatif chaque fois que l'on tombe sur pile (Se=0,5 et 1-Sp=0,5).

K symbolise les aptitudes diagnostiques du docteur Knock cher à Jules Romain. Le médecin affirmant que "tout bien portant est un malade qui s'ignore" a une sensibilité parfaite mais inquiète inutilement tous les bien-portants. Sa spécificité est nulle.

La performance diagnostique globale du test est d'autant meilleure que la courbe ROC s'éloigne de la diagonale. Elle se quantifie par l'estimation de l'aire sous la courbe. Un test dont la courbe ROC est sur la diagonale et qui n'a donc pas d'intérêt diagnostique, a une aire sous la courbe de 0,5. Elle peut s'interpréter comme la probabilité qu'un sujet malade ait une valeur du test supérieure à celle d'un sujet non malade, lorsqu'une valeur élevée du test est en faveur de la maladie. Le test est d'autant meilleur pour discriminer les malades des non-malades que son aire sous la courbe se rapproche de 1.

Une méthode non paramétrique d'estimation de l'aire sous la courbe consiste à calculer pour toutes les paires (malade, non malade), la proportion de paires pour lesquelles la valeur du test chez le sujet malade est supérieure à la valeur du test chez le sujet non malade, lorsqu'une valeur élevée du test est en faveur de la maladie. Il s'agit de la statistique de Mann et Whitney. L'aire sous la courbe ROC de la mammographie pour faire le diagnostic de cancer du sein dans l'étude de Hall FM et al. est estimée à 0,81 avec un intervalle de confiance à 95% compris entre 0,76 et 0,85. La mammographie fait significativement mieux que le hasard car la borne inférieure de l'intervalle de confiance est supérieure à 0,5.

3 - Choix du seuil de positivité d'un test

A la phase clinique de l'évaluation d'un test diagnostique ordinal ou continu, la détermination d'un seuil de positivité est nécessaire. Le seuil de positivité optimal est celui qui maximise l'utilité dans la population dans laquelle le test est utilisé. **L'utilité** est définie comme une mesure de l'état de santé ou de la préférence pour un état de santé ; il s'agit par exemple de l'espérance de vie pondérée par la qualité de vie. L'utilité moyenne dans la population dépend de l'utilité de chacune des situations (sujet malade et traité, sujet malade et non traité, sujet non malade et non traité, sujet non malade et traité) et de la fréquence de chacune de ces situations.

L'utilité moyenne pour le seuil c, notée U(c) s'écrit :

$$U(c) = Se \times p \times U_{VP} + (1 - Se) \times p \times U_{FN} + Sp \times (1 - p) \times U_{VN} + (1 - Sp) \times (1 - p) \times U_{FP}$$

Se = sensibilité du test

Sp = spécificité du test

p = prévalence de la maladie ou probabilité pré-test

U_{VP}, U_{FN}, U_{VN}, U_{FP} sont les utilités associées aux quatre situations : sujet malade et traité (vrai positif), sujet malade et non traité (faux négatif), sujet non malade et non traité (vrai négatif), sujet non malade et traité (faux positif).

L'utilité moyenne, U(c), peut être réécrite en fonction du bénéfice net en termes d'utilité à traiter à raison un sujet malade et du coût net en termes d'utilité à traiter à tort un sujet non malade

Les méthodes permettant d'estimer le seuil qui maximise l'utilité moyenne dépassent le cadre de cet ouvrage et ne sont donc pas présentées. Le lecteur intéressé pourra trouver ces méthodes dans les références données en fin de chapitre.

C - Evolution de la probabilité d'avoir la maladie à l'issue du test

1 – Les valeurs prédictives et le théorème de Bayes

L'estimation de la sensibilité et de la spécificité permet d'évaluer les performances diagnostiques d'un test, mais pour le clinicien qui va utiliser le test, ce qui compte ce sont les valeurs prédictives positive et négative.

La valeur prédictive positive (VPP) est la probabilité que le sujet ait la maladie sachant qu'il a un test positif. La valeur prédictive négative (VPN) est la probabilité que le sujet n'ait pas la

maladie sachant qu'il a un test négatif. Ces valeurs prédictives dépendent de la sensibilité et de la spécificité du test, mais également de la prévalence de la maladie ou probabilité pré-test. Dans une étude de type cohorte il est possible d'estimer les valeurs prédictives directement à partir du tableau 2×2 par une lecture horizontale du tableau. Reprenons les résultats de l'étude CASS présentés dans le tableau 2. L'échantillon constitué pour l'étude est *a priori* représentatif de la population des sujets adressés pour une coronarographie en raison d'une suspicion de coronaropathie. La prévalence de la maladie dans cette population peut être estimée à partir des données de l'étude à 70 % (1023/1465).

La valeur prédictive positive de la douleur thoracique est estimée à : $\frac{969}{1214} = 80\%$

La valeur prédictive négative de la douleur thoracique est estimée à : $\frac{197}{251} = 78\%$

En revanche les études de type cas-témoins ne permettent pas d'estimer directement les valeurs prédictives, car elles ne reposent pas sur l'inclusion d'un échantillon représentatif d'une population, mais sur l'inclusion indépendante d'un échantillon de malades et d'un échantillon de non-malades dont les effectifs sont fixés par l'investigateur. Ayant une estimation de la sensibilité et de la spécificité du test à partir d'une étude de type cas-témoins, et par ailleurs une estimation de la prévalence de la maladie dans la population d'intérêt, il est possible d'estimer les valeurs prédictives positive et négative du test en utilisant le théorème de Bayes.

Le théorème de Bayes permet de façon générale d'inverser les probabilités conditionnelles et de passer par exemple de la probabilité que le test soit positif sachant que le sujet est malade (sensibilité) à la probabilité que le sujet ait la maladie sachant que le test est positif (VPP).

$$\begin{aligned}
\text{VPP} = \text{P}(\text{M/Test}+) &= \frac{\text{P}(\text{M et Test}+)}{\text{P}(\text{Test}+)} = \frac{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})}{\text{P}(\text{Test}+\text{ et M})+\text{P}(\text{Test}+\text{ et NM})} \\
&= \frac{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})}{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})+\text{P}(\text{Test}+/\text{NM})\times\text{P}(\text{NM})} \\
&= \frac{\text{Se}\times\text{Prévalence}}{\text{Se}\times\text{Prévalence}+(1-\text{Sp})\times(1-\text{Prévalence})}
\end{aligned}$$

Le théorème de Bayes permet également de passer de la probabilité que le test soit négatif sachant que le sujet est non-malade à la probabilité que le sujet soit non malade sachant que le test est négatif.

$$\begin{aligned}
\text{VPN} = \text{P}(\text{NM/Test}-) &= \frac{\text{P}(\text{NM et Test}-)}{\text{P}(\text{Test}-)} = \frac{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})}{\text{P}(\text{Test}-\text{ et NM})+\text{P}(\text{Test}-\text{ et M})} \\
&= \frac{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})}{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})+\text{P}(\text{Test}-/\text{M})\times\text{P}(\text{M})} \\
&= \frac{\text{Sp}\times(1-\text{Prévalence})}{\text{Sp}\times(1-\text{Prévalence})+(1-\text{Se})\times\text{Pr}\textit{évalence}}
\end{aligned}$$

Se = sensibilité, Sp = spécificité, M = malade, NM = non-malade

Pour illustrer le fait que les valeurs prédictives dépendent beaucoup de la prévalence, nous allons prendre l'exemple de l'utilisation de la mammographie pour faire le diagnostic de cancer du sein en situation de dépistage ou en consultation spécialisée. D'après les résultats de l'étude de Hall FM et al, et en considérant comme positives les mammographies pour lesquelles il y a une haute suspicion de cancer, la sensibilité est estimée à 39% et la spécificité à 98%.

Pour une prévalence de 4 pour mille dans la population des femmes dépistées entre 50 et 65 ans, la valeur prédictive positive est de : $\text{VPP} = \frac{0,39 \times 0,004}{0,39 \times 0,004 + (1 - 0,98) \times (1 - 0,004)} = 7,3\%$

Pour une prévalence de 30% dans la population venant en consultation spécialisée, la valeur prédictive est de : $\text{VPP} = \frac{0,39 \times 0,3}{0,39 \times 0,3 + (1 - 0,98) \times (1 - 0,3)} = 89\%$

L'information apportée par le test est la même dans les 2 cas, mais la probabilité pré-test de la maladie est très différente. La valeur prédictive positive est meilleure dans la population où la proportion de malades est plus importante. En revanche la valeur prédictive négative est meilleure dans la population où la proportion de non-malades est plus importante. Elle est estimée à 99,8% dans la population des femmes dépistées et à 78,9% dans la population des femmes qui viennent en consultation spécialisée.

2 – Les probabilités pré et post-test et les ratios de vraisemblance

L'information apportée par le test dépend de sa sensibilité et de sa spécificité et peut être quantifiée par les ratios de vraisemblance. On distingue le ratio de vraisemblance positif qui correspond à l'information apportée par le test lorsque le test est positif, et le ratio de vraisemblance négatif qui correspond à l'information apportée par le test lorsque le test est négatif.

Le ratio de vraisemblance positif d'un test (*positive likelihood* ratio en anglais, LR+) est le rapport de la vraisemblance d'un résultat positif chez les malades sur la vraisemblance d'un résultat positif chez les non-malades : $\text{RV+} = \frac{\text{P}(\text{Test} + /\text{M})}{\text{P}(\text{Test} + /\text{NM})} = \frac{Se}{1 - Sp}$

Un test qui ne fait pas mieux que le hasard pour discriminer les malades des non-malades, est un test pour lequel la vraisemblance d'un résultat positif chez les malades est égale à la vraisemblance d'un résultat positif chez les non-malades. Cette situation correspond à un RV+ égal à 1. Plus le ratio de vraisemblance positif est supérieur à 1, plus l'information apportée par un résultat positif du test est importante.

Le RV+ permet de passer de la probabilité pré-test à la probabilité post-test lorsque le test est positif. Il multiplie l'Odds pré-test de la maladie. Reprenons l'exemple de la mammographie avec un seuil de positivité correspondant à une haute suspicion de cancer.

$$\text{RV+} = \frac{0,39}{1 - 0,98} = 19,5$$

L'Odds de cancer du sein en situation de dépistage est égal à :

$$\text{Odds pré test} = \frac{\text{prévalence}}{1 - \text{prévalence}} = \frac{0.004}{1 - 0.004} \approx 0{,}004$$

L'Odds post-test lorsque la mammographie est positive :

$$\text{Odds post test} = \text{Odds pré test} \times \text{RV} + = 0{,}004 \times 19{,}5 = 0{,}078$$

La probabilité post-test est égale à : $\frac{\text{Odds post test}}{1 + \text{Odds post test}} = 7{,}2\%$

On retrouve la valeur prédictive positive ou probabilité d'avoir la maladie sachant que le test est positif. Il s'agit d'une autre façon d'appliquer le théorème de Bayes.

Le ratio de vraisemblance négatif (*negative likelihood* ratio en anglais, LR-) est le rapport de la vraisemblance d'un résultat négatif chez les malades sur la vraisemblance d'un résultat négatif chez les non-malades : $\text{RV} - = \frac{\text{P}(\text{Test} - / \text{M})}{\text{P}(\text{Test} - / \text{NM})} = \frac{1 - Se}{Sp}$

Plus le ratio de vraisemblance négatif se rapproche de 0, plus l'information apportée par un résultat négatif du test est importante.

Le RV- de la mammographie est égal à : $\text{RV} - = \frac{1 - 0{,}39}{0{,}98} = 0{,}62$

Si la mammographie est négative l'Odds de la maladie est divisé par 1,6.

$$\text{Odds post test} = \text{Odds pré test} \times \text{RV} - = 0{,}004 \times 0{,}62 \approx 0{,}0025$$

La probabilité post-test est égale à : $\frac{\text{Odds post test}}{1 + \text{Odds post test}} \approx 2{,}5$ pour mille

La probabilité post-test correspond à la probabilité d'avoir la maladie sachant que le test est négatif. Il s'agit de 1 moins la valeur prédictive négative.

Le RV+ dépend surtout de la spécificité du test. Meilleure est la spécificité du test, meilleur est le test pour affirmer la présence de la maladie lorsqu'il est positif. Le RV- dépend surtout de la sensibilité. Meilleure est la sensibilité, meilleur est le test pour éliminer la maladie lorsqu'il est négatif. Prenons l'exemple de 3 tests : la gazométrie dans le sang artériel pour faire le diagnostic d'embolie pulmonaire, la culture de liquide pleural pour faire le diagnostic de tuberculose et le scanner pour faire le diagnostic de masse rénale kystique (tableau 4).

La gazométrie est sensible mais peu spécifique pour le diagnostic d'embolie pulmonaire. Ce test permet de réduire de façon importante la probabilité d'avoir la maladie s'il est négatif en divisant l'Odds pré-test par 10. En revanche il multiplie l'Odds pré-test seulement par 2 lorsqu'il est positif.

A l'inverse la culture de liquide pleural est très spécifique pour le diagnostic de tuberculose mais très peu sensible. Ce test permet d'augmenter de façon importante la probabilité d'avoir la maladie s'il est positif en multipliant l'Odds pré-test par 24. En revanche, il divise l'Odds pré-test seulement par 1,3 s'il est négatif.

Le scanner est un test qui a à la fois une sensibilité de 100% et une spécificité élevée. C'est un test qui n'a pas de faux négatifs. Il permet d'éliminer la maladie lorsqu'il est négatif. Un test positif multiplie par 50 l'Odds pré-test.

III - QUELQUES ASPECTS PARTICULIERS AUX ETUDES D'EVALUATION D'UN TEST DIAGNOSTIQUE

A – Quelques biais spécifiques aux études d'évaluation des tests diagnostiques

1 – Biais de vérification

Dans les études d'évaluation des stratégies diagnostiques, il y a un risque d'obtenir des estimations biaisées chaque fois que le statut malade/non-malade n'est pas mesuré de façon indépendante du test à évaluer ou l'inverse. Par exemple le biais d'incorporation survient lorsque la détermination du statut malade, non-malade repose au moins en partie sur le résultat du test à évaluer. Cela entraîne une surestimation de la sensibilité et de la spécificité.

Dans cette catégorie de biais, on trouve le biais de vérification qui survient lorsque la probabilité d'avoir le *gold standard* dépend du résultat du test à évaluer. Cette situation se présente classiquement lorsque le *gold standard* est invasif ou coûteux et ne peut pas être réalisé chez tout le monde. Dans ce cas, il est plus souvent réalisé chez les sujets qui ont un test positif que chez ceux qui ont un test négatif.

Une étude mise en place pour évaluer les performances diagnostiques de l'électrocardiogramme d'effort a porté sur 414 sujets à risque de coronaropathie. Tous les sujets ont eu un électrocardiogramme d'effort. Tous les sujets ayant un électrocardiogramme d'effort positif ont eu une coronarographie. Pour les sujets ayant un électrocardiogramme d'effort négatif, seuls 40 % pris au hasard ont eu une coronarographie. Les résultats sont présentés dans le tableau 5. La sensibilité et la spécificité estimées chez les sujets qui ont eu une coronarographie sont respectivement de : $Se = \frac{92}{92+46} = 67\%$ et $Sp = \frac{72}{72+27} = 73\%$.

La probabilité d'avoir une coronarographie étant plus élevée chez les sujets ayant un test positif que chez ceux ayant un test négatif, il y a une surreprésentation des tests positifs. La sensibilité du test est surestimée et la spécificité sous-estimée. La probabilité d'avoir une coronarographie ne dépendant que du résultat du test, il est possible d'obtenir les estimations non biaisées de la sensibilité et de la spécificité en utilisant le théorème de Bayes. A partir des résultats présentés dans le tableau 5, nous avons une estimation de :

- la probabilité que le test soit positif dans la population des sujets à risque de coronaropathie : $\frac{119}{119+295} = 28{,}7\%$

- la probabilité d'avoir la maladie sachant le test positif : $\frac{92}{92+27} = 77{,}3\%$

- la probabilité de ne pas avoir la maladie sachant le test négatif : $\frac{72}{46+72} = 61\%$

Estimation de la sensibilité :

$$Se = P(Test+ / M) = \frac{\mathrm{P(M/Test+)} \times \mathrm{P(Test+)}}{\mathrm{P(M/Test+)} \times \mathrm{P(Test+)} + \mathrm{P(M/Test-)} \times \mathrm{P(Test-)}}$$
$$= \frac{0{,}773 \times 0{,}287}{0{,}773 \times 0{,}287 + (1-0{,}61) \times (1-0{,}287)} = 44\%$$

Estimation de la spécificité :

$$Sp = P(Test- / NM) = \frac{\mathrm{P(NM/Test-)} \times \mathrm{P(Test-)}}{\mathrm{P(NM/Test-)} \times \mathrm{P(Test-)} + \mathrm{P(NM/Test+)} \times \mathrm{P(Test+)}}$$
$$= \frac{0{,}61 \times (1-0{,}287)}{0{,}61 \times (1-0{,}287) + (1-0{,}773) \times 0{,}287} = 87\%$$

Lorsqu'il existe un *gold standard* mais qu'il ne peut pas être utilisé chez tous les sujets inclus dans l'étude, il est possible d'estimer les performances du test à évaluer en utilisant le *gold standard* sur un échantillon de sujets positifs et un échantillon de sujets négatifs pris au hasard.

2 – Biais lié à l'utilisation d'un *gold standard* imparfait

Il est très fréquent que le test utilisé comme référence ne soit pas parfait. Si l'on estime la sensibilité et la spécificité du test à évaluer en faisant comme si le test de référence était parfait, ces estimations sont biaisées. En particulier, il est impossible de montrer la supériorité du nouveau test par rapport au test de référence.

Prenons l'exemple d'un nouveau test parfait, dont on évalue les performances par rapport à un test de référence qui a une sensibilité de 90 % et une spécificité de 90 %. Dans une étude portant sur 100 sujets malades et 100 sujets non malades, 10 sujets malades seront classés comme négatifs par le test de référence et 10 sujets non-malades seront classés comme positifs par le test de référence (tableau 6). La sensibilité et la spécificité du nouveau test seront sous-estimées à 90 %.

Dans le cas où les 2 tests sont indépendants conditionnellement au statut vis-à-vis de la maladie, un défaut de sensibilité du test de référence entraîne une sous-estimation de la spécificité du nouveau test. A l'inverse un défaut de spécificité du test de référence entraîne une sous-estimation de la sensibilité du nouveau test.

Il est possible d'estimer les performances diagnostiques d'un test dans la situation où le test de référence n'est pas parfait. Le statut malade, non-malade des sujets inclus dans l'étude n'est pas directement observé, c'est une variable latente. Les résultats positif ou négatif du test à évaluer et du test de référence apportent de l'information sur le statut des sujets.

Dans une étude portant sur un échantillon d'une population dans laquelle les sujets inclus ont eu le test à évaluer et le test de référence, il y a 5 paramètres à estimer : la sensibilité et la spécificité du nouveau test, la sensibilité et la spécificité du test de référence et la prévalence de la maladie. Le tableau 2×2 présentant les résultats croisés des 2 tests permet d'estimer 3 paramètres. Si l'on connaît la sensibilité et la spécificité du test de référence, alors il est possible d'estimer la sensibilité et la spécificité du nouveau test et la prévalence de la maladie. Les données observées apportent 3 degrés de liberté.

Si aucun des paramètres n'est connu avec certitude, alors il est nécessaire d'augmenter l'information apportée par les données. Hui et Walter [3] ont proposé d'utiliser des échantillons de sujets provenant de 2 populations ayant des prévalences de la maladie très différentes. Ils ont pris l'exemple de l'évaluation d'un nouveau test cutané pour faire le diagnostic de tuberculose, le test de Tine. Le test de référence est le test cutané de Mantoux. Ils ont repris les données de 2 études : l'une dans laquelle les 2 tests ont été appliqués à un échantillon provenant de la population d'un district scolaire ayant une faible prévalence de la maladie, et l'autre dans laquelle les 2 tests ont été appliqués à un échantillon d'une population d'un sanatorium ayant une prévalence élevée de la maladie.

Sous l'hypothèse d'indépendance conditionnelle des 2 tests et de performances diagnostiques identiques dans les 2 populations, il y a 6 paramètres à estimer : la sensibilité et la spécificité de chacun des tests et la prévalence de la maladie dans chacune des populations. Les tableaux 2×2 présentant les résultats croisés des 2 tests dans chacune des populations nous apportent chacun 3 degrés de liberté. Le nombre de degrés de liberté est de 6, l'information apportée par les données est donc suffisante pour estimer tous les paramètres. Cette approche peut être généralisée à plus de 2 tests ou plus de 2 populations.

La présentation des méthodes d'estimation dépasse le cadre de cet ouvrage. Le lecteur intéressé pourra trouver ces méthodes dans les références données en fin de chapitre.

B – L'intervalle de confiance et le calcul du nombre de sujets nécessaire

1 – Intervalle de confiance de la sensibilité et de la spécificité

La sensibilité et la spécificité sont estimées respectivement par la proportion de résultats positifs chez les malades et la proportion de résultats négatifs chez les non-malades. Leur variance et erreur standard estimées correspondent à la variance et à l'erreur standard d'une proportion.

Pour la sensibilité : $\text{Variance} = \dfrac{\text{Se}\times(1-\text{Se})}{\text{M}}$ $\qquad$ $\text{Erreur standard} = \sqrt{\dfrac{\text{Se}\times(1-\text{Se})}{\text{M}}}$

M = effectif de malades

Pour la spécificité : $\text{Variance} = \dfrac{\text{Sp}\times(1-\text{Sp})}{\text{NM}}$ $\qquad$ $\text{Erreur standard} = \sqrt{\dfrac{\text{Sp}\times(1-\text{Sp})}{\text{NM}}}$

NM = effectif de non-malades

Si les effectifs de malades et non-malades sont suffisamment grands et si la sensibilité et la spécificité ne sont pas trop proches de 100 %, l'intervalle de confiance peut alors être construit en utilisant la méthode basée sur l'approximation de la distribution binomiale par une distribution de Gauss.

Intervalle de confiance à 95% de la sensibilité ou spécificité estimées : $\text{P} \pm 1{,}96 \times \sqrt{\dfrac{\text{P}\times(1-\text{P})}{\text{N}}}$

P correspond à la sensibilité ou à la spécificité estimées

N correspond à l'effectif de malades ou non-malades

Cette méthode de construction de l'intervalle de confiance fondée sur l'approximation gaussienne est en général applicable quand NP>= 5 et N(1-P)>=5.

Quand les effectifs sont trop petits ou les estimations trop proches de 100 %, il convient de construire l'intervalle de confiance exact fondé sur la loi binomiale.

2 – Calcul du nombre de sujets nécessaire

Quand la sensibilité et la spécificité attendues ne sont pas trop proches de 100 %, la méthode de calcul du nombre de sujets nécessaire pour estimer une sensibilité et une spécificité est la même que pour estimer une proportion de malades dans une population (prévalence de la maladie).

Si l'étude mise en place est une étude de type cas-témoins, le nombre de malades à inclure pour estimer la sensibilité et le nombre de non-malades à inclure pour estimer la spécificité sont déterminés séparément. Il convient alors de fixer la sensibilité et la spécificité attendues, la largeur souhaitée de l'intervalle de confiance et sa probabilité de couverture qui est en général de 95 %.

Si l'étude mise en place est une étude de type cohorte, il est nécessaire de tenir compte de la prévalence de la maladie dans la population d'où sera tiré l'échantillon de l'étude. Dans la plupart des cas, la prévalence de la maladie est inférieure à 50 %. La stratégie à suivre est alors la suivante. On calcule le nombre de malades à inclure pour estimer la sensibilité puis on calcule le nombre de sujets à inclure pour avoir le nombre de malades nécessaire, compte tenu de la prévalence.

$$N_{Total} = \frac{M}{Prévalence}$$

N_{Total} est l'effectif total de sujets à inclure dans l'étude

M est l'effectif de malades

Si la prévalence de la maladie est supérieure à 50 %. La même stratégie est appliquée mais c'est le nombre de non-malades à inclure pour estimer la spécificité qui est calculé en premier.

Dans le cas où la sensibilité et la spécificité attendue sont proches de 100 %, il convient d'utiliser une méthode exacte de calcul du nombre de sujets reposant sur la distribution binomiale.

IV – CONCLUSION

L'objectif de ce chapitre est de donner au lecteur les outils méthodologiques nécessaires pour la mise en place d'une étude visant à estimer les capacités diagnostiques d'un nouveau test. L'accent a été mis sur les études visant à estimer la sensibilité et la spécificité, ce qui correspond à la phase précoce de l'évaluation d'un nouveau test. Les deux ouvrages qui sont donnés en référence permettront aux lecteurs qui le désirent d'aller plus loin [4 ; 5]

Références

1. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tritani F, Chaitman BR, Fisher LD. Correlations among history of angina, ST-segment response and prevalence of coronary artery disease in the coronary artery surgery study (CASS). N Engl J Med 1979; 301: 230-5.

2. Hall FM, Storella JM, Silverstone DZ, Wyshak G. Non palpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography. Radiology 1988; 167: 353-8.

3. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-71.

4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Ed Oxford University Press, New York, 2003.

5. Zhou XH, Obuchowsky NA, McClish DK. Statistical methods in diagnostic medicine. Ed John Wiley & Sons, New York, 2002.

Tableau 1 - Les quatre situations possibles selon le résultat du test diagnostique et le statut malade ou non malade

	Maladie présente	Maladie absente	
Test positif	Vrai Positif (VP)	Faux Positif (FP)	VP+FP
Test négatif	Faux Négatif (FN)	Vrai Négatif (VN)	FN+VN
	VP+FN	FP+VN	N

Tableau 2 – Existence de douleurs thoraciques en fonction de la présence ou non d'une coronaropathie chez des sujets à risque (étude CASS)

	Coronaropathie présente	Coronaropathie absente	
Douleur thoracique	969	245	1214
Pas de douleur thoracique	54	197	251
	1023	442	1465

Tableau 3 – Classement de 119 femmes ayant un cancer du sein et de 281 femmes n'ayant pas de cancer du sein selon le degré de suspicion de cancer à la mammographie

Résultat mammographie Degré de suspicion de cancer	Cancer du sein	Pas de cancer du sein
Haute	47	6
Moyenne	57	117
Légère	9	37
Minime	6	121
Total	119	281

Tableau 4 – Sensibilité, spécificité et ratios de vraisemblance positif et négatif de trois tests différents

Test	Sensibilité	Spécificité	RV+	RV-
Gazométrie pour diagnostic d'embolie pulmonaire	0,95	0,5	1,9	1/10 = 0,1
Culture de liquide pleural pour le diagnostic de tuberculose	0,24	0,99	24	1/1,3 = 0,77
Scanner pour le diagnostic de masse kystique rénale	1	0,98	50	0

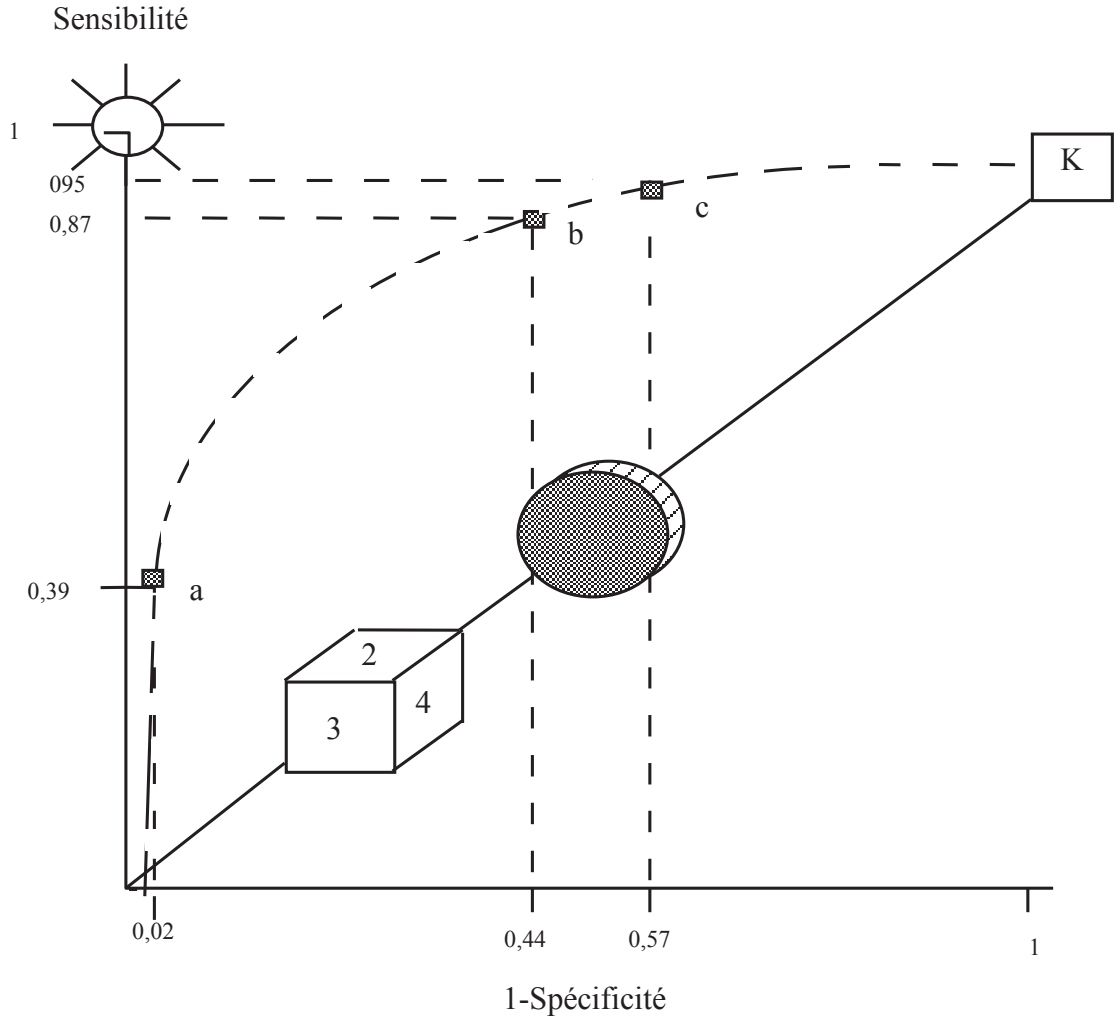
Tableau 5 – Résultats de l'électrocardiogramme (ECG) d'effort chez 414 sujets à risque de coronaropathie

	Résultat de la coronarographie			
	Maladie coronarienne	Pas de maladie coronarienne	Pas de coronarographie	Total
ECG d'effort positif	92	27	0	119
ECG d'effort négatif	46	72	177	295

Tableau 6 – Résultats d'une étude visant à évaluer les performances d'un nouveau test parfait par rapport à un test de référence qui a une sensibilité de 90% et une spécificité de 90%. L'étude porte sur un échantillon de 200 sujets comprenant 100 malades et 100 non-malades

	Test de référence positif	Test de référence négatif
Test à évaluer positif	90 VP	10 FP
Test à évaluer négatif	10 FN	90 VN
	100 = 90 VP + 10 FN	100 = 90 VN + 10 FP

Figure 1 – Courbe ROC de la mammographie pour le diagnostic de cancer du sein chez des femmes ayant une biopsie positive (étude de Hall FM et al)



CHAPITRE X

LES ETUDES ECONOMIQUES

Isabelle Jaisson-Hot, Catherine Buron, Jean-Paul Moatti, Cyrille Colin

Actuellement, les cliniciens ont le sentiment d'être confrontés à une double pression: l'une émanant des pouvoirs publics qui souhaiteraient les voir devenir les acteurs d'une politique de maîtrise des dépenses de santé, l'autre provenant des patients qui demanderaient à bénéficier sans limite des innovations technologiques médicales, souvent coûteuses.

Face à ces contraintes, il devient de plus en plus difficile de préserver ce fait déontologique fondamental que la décision médicale doit rester un contrat individuel entre le clinicien et le patient. Parce que les dépenses de santé augmentent près de deux fois plus vite que la richesse nationale, il est clair que les politiques de maîtrise des dépenses de santé vont continuer à peser encore longtemps sur le corps médical.

Dans l'hypothèse où le problème de l'équilibre à long terme des comptes de la Sécurité Sociale ne se poserait plus, la question de l'utilisation efficiente des ressources dans le système de santé n'en demeurerait pas moins, dès lors que les ressources financières disponibles ne sont jamais infinies. Toute utilisation non efficiente dans le secteur de la santé sacrifie la possibilité de produire plus de bien-être pour la collectivité, soit dans ce secteur lui-même, soit dans le reste de l'économie. Les mesures de maîtrise des dépenses de santé engagées ne font qu'accélérer la prise de conscience que toute décision médicale a des conséquences concrètes en termes d'allocation de ressources et implique un choix (implicite) de ne plus pouvoir utiliser les ressources ainsi consommées à d'autres fins.

La responsabilisation des cliniciens, qui engagent l'utilisation des ressources collectives, dans une politique de maîtrise des dépenses de santé, nécessite la connaissance des outils d'évaluation économique applicables à la pratique médicale.

Dans ce chapitre, nous nous efforcerons de mettre en perspective les analyses économiques appliquées aux stratégies de santé par rapport à ces débats sur le fonctionnement d'ensemble du système de santé.

PLAN DU CHAPITRE

Dans nombre de situations, le souci d'optimiser l'allocation des ressources qui constitue la finalité première de l'analyse économique n'est nullement en contradiction avec l'intérêt des patients : supprimer un examen obsolète ou inutile, c'est éviter un gaspillage économique, mais c'est d'abord agir dans l'intérêt du patient ; de même, privilégier une stratégie médicale qui apporte le plus d'efficacité à dépense constante, c'est contribuer à la fois à améliorer les performances économiques du système, à réduire les charges de l'assurance-maladie et s'assurer que l'utilisation des ressources apporte le maximum de bénéfices pour les patients.

Deux confusions sont trop fréquentes. La première est celle qui réduit l'évaluation économique au seul souci de "faire des économies" : développer une nouvelle stratégie de santé source d'accroissement de dépenses peut s'avérer rentable économiquement dès lors que ce surcoût entraîne un bénéfice supplémentaire pour la collectivité (par exemple, un gain d'espérance de vie ou de qualité de vie). La deuxième confusion qui découle souvent de la première est celle qui identifie l'intérêt économique collectif avec le point de vue d'un agent particulier, tel la caisse d'assurance maladie ou l'hôpital : la prise en charge de patients atteints d'une maladie chronique d'une façon ambulatoire ou par un organisme de soins à domicile peut constituer pour l'hôpital une source d'économie budgétaire, mais elle ne garantit pas automatiquement l'intérêt de la collectivité, car la prise en charge à domicile peut augmenter les coûts pour les patients ou leurs familles.

Une clarification méthodologique des principaux éléments de l'évaluation économique peut s'avérer utile aux cliniciens afin d'optimiser leurs décisions. Nombre de stratégies de santé obéissent inévitablement à une loi des rendements décroissants, c'est-à-dire l'augmentation des moyens de production entraînant un rendement supplémentaire moindre.

Les progrès de la réanimation néonatale illustrent cette évolution : ce que coûte aujourd'hui la réanimation d'un prématuré de moins de 1 kg est sans commune mesure avec ce qui était consacré il y a une décennie pour un nouveau-né; il a été estimé que réanimer un prématuré de moins de 1 kg revenait deux fois plus cher que pour un prématuré de 1 kg à 1,5 kg et sept fois plus si on l'estime en années de vie sans handicap.

Parce qu'il n'est pas envisageable de "tout faire à tout le monde", la question des seuils légitimes à fixer à la stratégie de santé se pose fréquemment. Il peut résulter une tension entre l'exigence médicale que "le médecin ne peut dans le contexte d'un cas individuel placer les intérêts de la société au-dessus de ceux de l'individu" et la gestion des ressources médicales. Contrairement à l'idée reçue qui oppose éthique médicale et économie, cette notion est inscrite dans le code de déontologie médicale par son article 2, qui précise que le médecin est au service de l'individu et de la santé publique. Les cliniciens sont étroitement concernés pour trouver les moyens de concilier leur autonomie de décision et l'intérêt collectif du système de soins.

Là encore une meilleure connaissance des outils de l'évaluation économique constitue un pré-requis indispensable à des débats transparents sur la limite de la stratégie de santé et sur les systèmes de régulation.

I - LES DIFFERENTS TYPES D'ANALYSES ECONOMIQUES

Le principe de base de l'analyse économique dans le cadre de la théorie de l'économie appliquée aux investissements publics est la détermination d'un choix entre différentes alternatives d'utilisation de ressources. Elle permet de dégager la stratégie la plus efficace.

L'évaluation économique ne peut donc se limiter à une analyse descriptive de coûts ; elle implique d'analyser les différentes options en considérant simultanément les coûts et les conséquences, ce qui la distingue d'autres formes d'évaluation médicale (tableau 1). Dans certains cas l'alternative peut être la simple absence de programme.

Trois types d'analyse peuvent être définis en fonction de la nature du problème et du choix à effectuer (tableau 2).

A - L'analyse de minimisation des coûts

L'analyse de minimisation des coûts a pour but d'identifier la stratégie de santé la moins coûteuse pour assurer un certain service. Elle s'exprime en unités monétaires par patient traité.

Cette analyse ne se justifie que s'il a été démontré que les différentes stratégies de santé étudiées ont les mêmes conséquences en termes d'efficacité thérapeutique, ce qui est rarement le cas.

De plus, il est souvent important de vérifier qu'à efficacité égale les conséquences sociales sont également similaires.

Ce type d'étude peut être adapté à la situation hospitalière. En effet, on se place ainsi du point de vue de l'institution produisant des soins et on compare les coûts médicaux directs de deux procédures, par exemple de deux traitements antibiotiques, l'un étant administré par voie orale, l'autre par voie intraveineuse.

B - Les analyses coût-efficacité et coût-utilité

L'analyse coût-efficacité permet de comparer des stratégies qui diffèrent par leurs coûts et leurs effets. Elle s'exprime en unités monétaires par indicateur d'efficacité médicale (par exemple, en euros par année de vie sauvée).

Par exemple, une étude française réalisée par le groupe PREMISS (Protocole en Réanimation d'Evaluation Médico-Economique d'une Innovation dans le Sepsis Sévère) [1] a évalué le ratio coût-efficacité de la drotrécogine alfa (DA) comparée à la prise en charge conventionnelle dans le traitement du sepsis sévère en pratique réelle. Grâce à un modèle de type «avant» / «après», un ratio coût-efficacité de 20 300 euros par année de vie gagnée a été mis en évidence. Le seuil retenu était celui d'acceptabilité proposé par l'Organisation mondiale de la Santé (3 fois le PIB par tête). Pour cette valeur de disposition à payer, la probabilité pour que la DA soit coût-efficace était de 79 %.

De même, si la transplantation rénale, la dialyse en milieu hospitalier ou la dialyse à domicile diffèrent quant au nombre d'années de vie sauvées, il est possible de les comparer grâce à leur ratio d'années de vie sauvées par euro dépensé.

Dans certains cas, les indicateurs d'efficacité à dimension unique apparaissent inadaptés, notamment les programmes évalués agissent à la fois sur la durée et la qualité de la vie des personnes qui en bénéficient.

Pour traiter de ces situations se sont développées les études de type "coût-utilité" où l'indicateur de résultat devient le nombre d'années de vie gagnées ajustées sur la qualité de la vie liée à la santé (QALY: Quality Adjusted Life Years).

Pour être utilisables dans les études de type coût-utilité, les mesures de qualité de vie liées à la santé doivent être déterminées sur une échelle d'intervalle comprenant la parfaite santé et la mort : on parle alors de préférences cardinales, par opposition aux préférences ordinales résultant de l'utilisation d'une méthode de mise en rang (questionnaire psychométrique de qualité de vie). On utilise pour cela des outils expérimentaux innovants de détermination des préférences conduisant à une estimation indirecte de la qualité de vie des patients. L'utilisation d'une échelle visuelle analogique, du pari standard ou de l'arbitrage temporel, sont les trois outils les plus couramment utilisés et qui répondent aux exigences méthodologiques sous-tendant la réalisation des analyses de type coût-utilité.

Une option alternative à ces méthodes consiste à utiliser des systèmes de classification d'états de santé multi-attributs qui sont pré-scorés et qui évitent donc le travail long et délicat d'évaluation des préférences individuelles des patients : les trois systèmes principaux sont le Quality of Well Being (QWB), le Health Utilities Index (HUI) et l'Euroquol (EQ-5D), dont les fonctions « scorage » sont fondées respectivement sur des mesures d'échelle catégorielle pour le QWB, de pari standard pour le HUI et d'une échelle visuelle analogique pour l'Euroquol.

Par exemple, une étude de type coût-utilité évaluant différentes stratégies thérapeutiques dans le cas des formes périanales fistulisées de la maladie de Crohn a été réalisée [2]. Dans ce travail, un modèle de Markov a été utilisé pour simuler une période de traitement d'un an. La stratégie de référence, constituée de l'association 6-mercaptopurine et metronidazole, était comparée à 3 interventions : il s'agissait pour la première de 3 perfusions d'infliximab (à 0, 2 et 6 semaines) avec l'association 6-mercaptopurine et métronidazole en seconde ligne en cas d'échec ; de 3 perfusions d'Infliximib (également à 0, 2 et 6 semaines) avec une reperfusion épisodique en cas d'échec, pour la deuxième ; et enfin pour la troisième, de l'association 6-mercaptopurine et métronidazole avec des perfusions d'Infliximab (à 0, 2 et 6 semaines) plus ou moins associées à des reperfusions épisodiques en cas d'échec. Le critère d'efficacité choisi était la survie ajustée sur la qualité de vie dont l'unité est le QALY. Pour une efficacité quasi similaire, les 3 interventions évaluées étaient beaucoup plus coûteuses. Les ratios coût-utilité étaient respectivement de 355 450 $/QALY (Intervention I), 360 900 $/QALY (Intervention II) et de 377 000 $/QALY (Intervention III). La véritable question émergeant de cette analyse était : la société est-elle prête à payer plus de 350 000 $ pour chaque année de vie ajustée sur la qualité de la vie ?

Les analyses coût-efficacité ou coût-utilité peuvent conduire à 4 types de situation :

- La stratégie étudiée est moins coûteuse et au moins aussi efficace que la stratégie de référence, donc efficiente (et dite «dominante») : elle peut être recommandée d'un point de vue médico-économique.
- La stratégie étudiée est plus coûteuse et moins efficace que la stratégie de référence, donc n'est pas efficiente (dite «dominée») et ne peut être recommandée.
- La stratégie étudiée est moins coûteuse et moins efficace que la référence. Cette situation pose la question de savoir si les économies dégagées sont suffisantes pour compenser une baisse de l'efficacité.
- La stratégie étudiée est plus coûteuse et plus efficace que la référence. C'est la situation la plus fréquente. Il est alors nécessaire d'arbitrer afin de savoir si l'augmentation du coût est acceptable du point de vue adopté par l'étude en regard du gain d'efficacité obtenu.

Cet arbitrage s'effectue au travers du calcul d'un ratio coût-efficacité différentiel (Incremental Cost Effectiveness Ratio – ICER) qui fait le rapport de la différence du coût de 2 stratégies sur la différence de leur efficacité. Il s'interprète comme le surcoût

engendré par la stratégie pour gagner une unité d'efficacité supplémentaire par rapport à la stratégie de référence. Bien qu'il soit possible de comparer de simples ratios des coûts par rapport aux résultats pour chacune des stratégies de santé, la comparaison correcte est celle des coûts différentiels avec les résultats différentiels en vue d'établir une mesure appropriée des résultats finaux et constituer un critère d'aide à la décision.

Une fois le ratio coût-efficacité différentiel calculé, il s'agit de statuer sur le caractère acceptable ou non de la somme qui devrait être allouée pour obtenir ce gain supplémentaire d'efficacité. Cela pose la question de la détermination du ratio seuil, c'est-à-dire du ratio jusqu'au niveau duquel la collectivité est prête à aller pour obtenir ce gain supplémentaire. Selon les pays, ce seuil varie. En Angleterre, le National Institute for Clinical Excellence (NICE) utilise une fourchette de seuils allant de 20 000 £ à 30 000 £ par QALY gagné, mais le débat est toujours en cours à propos de l'utilisation et du niveau des valeurs seuils de l'ICER. Certains pays ont essayé de dériver une valeur seuil de l'ICER implicite en se basant sur des décisions passées en matière d'affectation des ressources. L'Australie a estimé une valeur seuil de 69 900 AU$/QALY, la Nouvelle-Zélande de 20 000 NZ$/QALY et le Canada a défini une fourchette d'acceptation allant du coût d'une intervention dominante jusqu'à 80 000 CAN$/QALY avec une fourchette de rejet allant de 31 000 à 137 000 CAN$/QALY (Valeurs seuils pour le rapport coût-efficacité en soins en santé, KCE reports 100B, 2008). En France, l'utilisation de «deux fois le produit national brut (PNB)» par habitant est évoquée par certains auteurs qui suggèrent qu'en dessous de 50 000 euros par année de vie gagnée on pourrait accepter sans discuter certaines stratégies [3]. Il serait nécessaire de débattre de leur bien-fondé pour celles se situant entre 2 et 6 à 8 fois le PNB par tête. Enfin, seraient rejetées celles dépassant le chiffre de 150 000 à 200 000 euros par année de vie gagnée.

C - L'analyse coût-bénéfice

L'analyse coût-bénéfice vise à comparer le coût d'une stratégie et son bénéfice. Dans une étude de ce type, les coûts réels et les conséquences sont exprimés en unités monétaires. Cette analyse vise à calculer le bénéfice net pour la collectivité de la stratégie de santé étudiée et à déterminer celle des différentes stratégies envisageables qui "maximise le surplus social net". Elle aide à déterminer si un certain objectif mérite d'être réalisé.

Plusieurs techniques de mesure du bénéfice sont disponibles. Dans la première, apparue dans les années 60 et qualifiée de méthode du «capital humain», le bénéfice est mesuré en tant que résultat de la production améliorée (ou détériorée) par le programme de santé. Cette méthode a été fortement critiquée, notamment en raison de son assertion que le but principal de la société est l'augmentation du produit national par tête, ce qui implique que les personnes non productives ne peuvent avoir d'amélioration de leur état de santé. Une autre technique dite «préférence révélée» est basée sur l'estimation des arbitrages monétaires qu'effectue l'individu en termes de prise de risque. Cette méthode vise, par exemple, à estimer le consentement à payer pour une réduction de risque morbide ou mortel, par l'observation, sur le marché du travail, des différentiels de salaires associés à un différentiel de risque. On peut ainsi déduire une estimation statistique du prix de la vie humaine. La dernière technique dite «préférence déclarée» est fondée sur l'expression des préférences déclarées par les individus sur un marché fictif. Parmi les techniques permettant d'estimer les préférences déclarées, l'évaluation contingente consiste à proposer à un individu une situation de marché hypothétique sur lequel la personne interrogée doit indiquer le montant monétaire maximum

qu'elle est prête à consentir pour accéder au bien proposé [4] [5]. Ce montant est un indicateur de l'utilité ou de la satisfaction que ce bien lui apporte.

Par exemple, une étude française a évalué, par simulation, l'opportunité en termes de coûts et de bénéfices cliniques, d'une campagne de vaccination contre l'encéphalite à tique chez les militaires français en mission au Kosovo, versus l'absence de vaccination [6]. Les auteurs ont valorisé le coût de l'absence de vaccination (dans l'optique du «capital humain»), et l'ont comparé au coût du programme de vaccination. Le critère de jugement de l'évaluation économique correspondait aux bénéfices de la vaccination en termes monétaires (valorisés par la méthode du capital humain) mesuré par le nombre de cas d'encéphalites à tique évités, c'est-à-dire au final en coûts évités. Ces coûts évités sont ceux liés à la mortalité et à la morbidité de la maladie. Les auteurs ont utilisé pour cela la valeur de la vie humaine exprimée en francs d'après la pension versée par le Ministère de la Défense aux ayants droits en cas de décès en service. Le coût des séquelles, quant à lui, a été obtenu par le guide-barème des invalidités versées au titre des pensions militaires (Guide-barème des invalidités applicable au titre du code des pensions militaires d'invalidité et des victimes de la guerre, fourni par le Secrétariat d'Etat aux anciens combattants). Les résultats étaient les suivants : 143 cas d'encéphalites à tique pouvaient être évités par la vaccination /4 ans, dont 3 décès et 17 patients présentant des séquelles invalidantes ; les coûts totaux de la vaccination étaient de 25,0 MF et les coûts totaux «évités par le programme de vaccination » s'élevaient à 27,1 MF.

Quel que soit le type d'étude, des incertitudes sur l'estimation des coûts ou des conséquences peuvent exister ; aussi, il est recommandé de réaliser dans chaque cas une analyse de sensibilité. Celle-ci vérifie si la modification des valeurs prises par les principales variables entraîne ou non une modification des résultats de l'analyse.

II - LES DIFFERENTES PERSPECTIVES D'ANALYSE ECONOMIQUE

Plusieurs points de vue sont possibles pour une analyse économique. Les coûts, les résultats, les bénéfices peuvent être envisagés de quatre manières : selon le payeur, le prestataire de service (l'hôpital), le patient et la société.

A - Le coût pour le payeur et pour l'hôpital

Le coût pour le payeur (assurance maladie) est égal à la tarification permise par celui-ci, tandis que ce même coût est pour l'hôpital le coût réel de la prestation de service quelle que soit la tarification. Pour déterminer le coût réel pour l'hôpital, il est souvent nécessaire de disposer d'une comptabilité analytique dans l'établissement, c'est-à-dire d'un dispositif comptable autorisant le calcul du coût par séjour du patient.

Par exemple, une étude a établi en déterminant le coût du travail du personnel soignant et les fournitures nécessaires, que délivrer un antibiotique en une dose journalière à la place de plusieurs doses permettait l'économie de 6 euros par patient et par jour pour la même efficacité thérapeutique. Cette économie intéressante pour l'hôpital est néanmoins sans répercussion pour le payeur.

B - Le coût pour le patient

Le coût pour le patient est la somme à payer non couverte par l'assurance maladie pour la prestation d'un service, celle entraînée indirectement par le traitement ou la maladie (journée de travail non effectuée, déplacement non pris en charge, …) et celle des coûts subjectifs

(anxiété, douleur …). Si la chirurgie ambulatoire, qui permet la réduction de l'hospitalisation, est source d'économie pour le payeur, elle peut être pour le patient source de dépenses supplémentaires variant en fonction de différentes modalités de remboursement. Les intérêts sont alors contradictoires selon la perspective d'analyse.

C - Le coût pour la société

Le coût pour la société est le coût total net pour les différents agents économiques de cette société, incluant la perte de productivité du patient et les dépenses totales entraînées par la maladie et sa prise en charge. Il représente ainsi le coût d'opportunité mesurant le sacrifice de ressources consenti pour un programme donné et ne pouvant être utilisé pour un autre effet.

Ainsi, la diffusion des innovations en matière de diagnostic prénatal des anomalies génétiques par les techniques de biologie moléculaire est actuellement freinée ; l'application de ces méthodes se traduit, pour les hôpitaux qui les pratiquent, par des charges supplémentaires sans contrepartie suffisante, alors que la rentabilité globale pour la société serait sans doute élevée.

III - LES DIFFERENTS TYPES DE COUTS

Différents types de coûts constituent le coût total : coût direct (médical et non médical) et coûts indirects (liés indirectement à la prise en charge).

A - Les coûts directs

Le coût direct est la valeur des ressources directement consommées pour le programme analysé.

Le coût direct médical inclut habituellement les frais d'hospitalisation, les médicaments, les examens biologiques et radiologiques, les honoraires médicaux, les soins de réhabilitation et les soins au long cours nécessaires.

Son estimation pose des problèmes méthodologiques, notamment parce que l'absence d'un véritable système de prix de marché dans le secteur sanitaire fait que les dépenses de santé ne sont pas toujours représentatives de la valeur réelle des coûts médicaux. En effet, le coût doit être distingué d'une tarification. La tarification est imposée par un système de régulation, qui souvent ne reflète pas le coût réel de production d'un bien ou d'un service.

De plus, de nombreux autres coûts directs non médicaux doivent être considérés : nourriture, transport non médicalisé vers les établissements de santé, équipement ou adaptation du domicile… Ils sont provoqués par la maladie ou le traitement mais n'ont pas entraîné de consommation d'un service médical. Certaines de ces sommes peuvent être à la charge directe du patient ou des proches du patient : une étude a montré que la famille d'un enfant cancéreux utilise un quart de son revenu pour des dépenses non médicales dues au traitement et habituellement non remboursées.

B - Les coûts indirects

Les coûts indirects représentent les pertes de productivité liées à la maladie, c'est-à-dire les heures de travail perdues consécutives à un épisode morbide.

La détermination de la perte de productivité se fait par la mesure du nombre d'heures ou de journées de travail perdues du fait de la maladie et de sa prise en charge. La mesure de ce nombre se fait par des enquêtes ad-hoc. Trois méthodes existent pour le chiffrage monétaire de ces pertes (Guide méthodologique pour l'évaluation économique des stratégies de santé, Collège des économistes de la santé, 2003) :

- La théorie du capital humain

Cette théorie conduit à chiffrer l'impact de la maladie par les pertes de production qu'elle induit, en multipliant le nombre de journées de travail perdues par la valeur de cette production, exprimée par le PIB par personne active, rapportée à la journée de travail. Simple à mettre en œuvre, cette approche est peu réaliste dans la mesure où elle repose sur l'hypothèse d'une économie de plein emploi au sein de laquelle la perte d'une journée de travail a un impact proportionnel et mécanique sur la production.

- L'approche des coûts de friction

Cette approche, plus réaliste que la précédente, considère que la perte de production n'est pas exactement proportionnelle au nombre de journées de travail et propose une modélisation macro-économique de l'impact des arrêts de travail sur le marché du travail. Elle requiert donc un travail empirique spécifique, important, réactualisé de façon permanente pour être appliquée.
Ces deux premières approches ne permettent pas de prendre en compte le travail non rémunéré, en particulier le travail domestique, ou le temps des personnes inactives comme les personnes retraitées ou en dehors du marché du travail en raison d'un handicap.

- La théorie du bien-être

La troisième approche est celle dérivée de la théorie du bien-être et s'applique dans le cadre d'études coût par QALY ou coût-bénéfice avec la propension à payer. Cette approche consiste à évaluer les inconvénients liés aux arrêts de travail en les intégrant dans l'évaluation faite par les patients de leur état de santé à la suite d'un traitement donné. Les coûts dits indirects sont inclus au dénominateur du ratio coût-résultats.

L'ampleur des coûts pris en considération doit être jugée en fonction du contexte et de l'objectif de chaque étude : se limiter aux seuls coûts médicaux directs peut être tout à fait adapté lorsqu'on compare deux protocoles thérapeutiques voisins dans un même contexte hospitalier ; en revanche négliger les coûts indirects biaise l'évaluation lorsqu'une stratégie de prise en charge hospitalière des troubles psychiatriques est comparée à une stratégie d'intervention communautaire.

IV – ACTUALISATION ET INFLATION

Actualisation :

Selon la théorie économique, une somme payée aujourd'hui n'a pas la même valeur si elle est payée plus tard : il existe une préférence pour le présent, car la disponibilité immédiate de ressources permet de les investir avec un intérêt, assurant plus tard une somme supérieure. Le raisonnement vaut pour les coûts et pour les résultats, lorsque ceux-ci sont valorisés de façon monétaire (analyse coût bénéfice).

L'actualisation est un calcul économique qui standardise les coûts dans le temps. Il repose sur un taux d'actualisation r qui relie C_n coûts supportés à l'année n, à leur valeur actuelle C_0 ; on peut utiliser des tables [7] ou une formule. En admettant que les coûts supportés en début d'année, donc non actualisés la 1ère année, la valeur actualisée VA d'un coût est :

$$VA = \Sigma C_n(1+r)^{-n} = C_0 + C_1/(1+r) + C2/(1+r)^2 + \ldots + C_n/(1+r)^n$$

Le taux r est souvent choisi à 5 % ; une étude de sensibilité est conseillée (cf ci-après) pour apprécier l'impact de différents taux possibles sur les résultats de l'étude.

Inflation :

Lorsque l'augmentation annuelle des prix est notable (exemple des années 70 en France), ou s'applique différemment aux actions ou programmes comparés, elle doit être intégrée dans l'analyse.

V - CONCLUSION

S'agissant de la définition des politiques publiques en matière de santé, de programmes nationaux de prévention, des allocations d'investissements en équipements ou moyens lourds, le recours à l'analyse économique est "naturel" pour aider à définir, comme on l'a vu, les modalités optimales d'une intervention.

S'agissant de l'intégration directe de l'évaluation économique dans l'optimisation des stratégies cliniques, il faut en revanche se garder de toute logique normative et accepter comme nous l'avons signalé plus haut des tensions entre éthique collective et éthique individuelle. C'est afin de favoriser la transparence de tels débats que nous avons proposé, tout au long de ce chapitre, un rapprochement entre réflexion économique et pratique clinique.

Tableau 1 - Spécificité de l'évaluation économique en médecine

		Y-a-t-il mesure des coûts et des conséquences des alternatives ?	
		oui	non
Y-a-t-il comparaison d'alternatives ?	oui	évaluation économique	essai thérapeutique
	non	comptabilité analytique étude clinique	statistique de consommation

Tableau 2 - Typologie des études d'évaluation économique de stratégies de santé

type	mesure des coûts	identification des conséquences	mesure des conséquences
minimisation des coûts	euros	conséquences identiques pour toutes les alternatives comparées	aucune
coût-efficacité	euros	indicateur d'efficacité à dimension unique	unité physique
coût-utilité	euros	indicateur d'efficacité à plusieurs dimensions	QALY Espérance de vie ajustée sur la qualité
coût-bénéfice	euros	indicateur d'efficacité à une ou plusieurs dimensions	euros

Références

1. Payet S, Riou-França L, Le Lay K, Vallet B, Dhainaut JF, Launois R et le groupe PREMISS, Evaluation coût-efficacité de la drotrécogine alfa comparée à la prise en charge conventionnelle dans le traitement du sepsis sévère en pratique réelle, Journal d'Economie Médicale 2007, vol 25, n°4 ; 207-223.

2. Arseneau KO, Cohn SM, Cominelli F, Connors AF, Cost-utility of initial medical management for Crohn's disease perianal fistulae, Gastroenterology 2001;120:1640-1656.

3. Moatti JP, Le plan cancer en France : une réflexion d'économiste, Bull Cancer 2003 ; 90 :1010-5.

4. Desjeux G, Colin C, Launois R, La mesure de la disposition à payer dans l'analyse coût-bénéfice : l'évaluation contingente, Journal d'Economie Médicale 2005, 23(5) : 293-306.

5. Allenet B, Sailly JC, La mesure du bénéfice santé par la méthode du consentement à payer, Journal d'économie médicale, n°5, 1999.

6. Desjeux G, Lemardeley P, Colin C, Pascal B, Labarere J, Etude coût-bénéfice de la vaccination contre l'encéphalite à tique chez les militaires français au Kosovo. Revue d'Epidémiologie et de Santé Publique;2001;49(3):249-257.

7. Drummond M, O'Brien B, Stoddard G, Torrance GW. Méthodes d'évaluation économique des programmes de santé. 2ème édition. Edition Economica 1998.

On pourra consulter également :

8. Gold MR, Siegel JE, Russell LB and Weinstein MC. Cost-effectiveness in health and Medicine, Oxford University Press, New York, 1996.

9. Bennett KJ, Torrance GW. Measuring health preferences and utilities : rating scale, time trade-off and standard gamble methods. In : Spliker B (ed). Quality of life and Pharmacoeconomics in clinical trials. Philadelphia : Lippincott-Raven, 1996;235-265.

10. Duru G, Auray JP, Beresniak A, Lamure M, Paine A, Nicoloyanis N. Limitations of the methods used for calculating Quality-Adjusted-Life-Years value. Pharmacoeconomics 2002;20(7):463-473.

11. Launois R. Les Arcanes décryptées de l'analyse médico économique à l'usage du décideur. Journal d'Economie Médicale 2008,vol.26,n°6-7,331-349.

CHAPITRE XI

LES ANALYSES DE DECISION

A Termoz, C Buron, I Jaisson-Hot, C Colin

L'évolution rapide des connaissances médicales au cours des trente dernières années, que ce soit sur le plan des outils diagnostiques ou celui des moyens thérapeutiques, est telle que, dans le domaine de la santé, tout décideur, médecin ou gestionnaire, est confronté à des choix de plus en plus difficiles. Dans le même temps, ces progrès techniques sont accompagnés d'interrogations de la part des consommateurs sur l'efficacité et l'innocuité du système de soins.

L'analyse de décision permet de décrire, pour une situation clinique ou de santé publique donnée, les différentes stratégies diagnostiques et thérapeutiques possibles. Elle permet de modéliser la décision médicale en intégrant à la fois les données expérimentales, les données épidémiologiques, les avis d'experts, et l'appréciation de l'état de santé du patient. Elle autorise de plus la prise en compte du point de vue du patient et de sa qualité de vie. A partir de ces éléments, l'analyse de décision s'attache à mettre en évidence une préférence pour une stratégie d'action dans une situation clinique ou de santé publique donnée.

L'objectif de ce chapitre est de présenter, à partir d'exemples simples de la pratique clinique, les fondements méthodologiques de l'analyse de décision.

PLAN DU CHAPITRE

Dans la pratique quotidienne actuelle des soins, les médecins utilisent les technologies les plus avancées et les approches scientifiques les plus élaborées. Ces nouvelles techniques médicales, souvent très coûteuses, ne se substituent pas toujours aux techniques antérieures. Le développement de ces technologies pose alors le problème éthique de l'évaluation de l'efficacité à améliorer l'état de santé de la population, et de l'accessibilité des patients à la technique.

Notre formation médicale initiale a été centrée sur une approche analytique du diagnostic positif et différentiel, c'est-à-dire l'examen précis et exhaustif de toutes les maladies pouvant affecter le patient. Cette formation a poussé les médecins à collecter le plus de données possible grâce aux tests de laboratoire et d'imagerie afin de porter un diagnostic positif et d'écarter des diagnostics différentiels. Une fois le diagnostic posé, les technologies et les procédures les plus récentes sont employées pour soigner et traiter le problème de santé. L'usage de certaines procédures intervient parfois avant que leur efficacité n'ait été prouvée, sur la simple hypothèse qu'une nouvelle technologie a une bonne probabilité d'être plus efficace qu'une ancienne.

Cette pratique des soins encourage l'idée selon laquelle plus on engage des ressources en examens paracliniques et en traitements pour un patient, meilleur est le résultat. Cette approche, qu'on pourrait appeler celle du maximum des moyens, fonctionne bien tant que les effets secondaires des actes médicaux sont mineurs et tant que la source de financement est suffisante. Cependant la distribution des soins médicaux change radicalement selon l'époque, le lieu, le type de pratique et la disponibilité des ressources ou des technologies nouvelles. L'augmentation des dépenses de santé dans les pays industrialisés, de même que le manque de ressources dans les pays en voie de développement, rendent impérative une approche évaluative de la stratégie diagnostique et thérapeutique.

I - REALISATION D'UNE ANALYSE DE DECISION

L'analyse de décision est une méthode quantitative utilisant les probabilités pour éclairer le processus de décision dans des situations d'incertitude. Elle est issue des sciences de gestion, dans lesquelles elle était appliquée pour déterminer les meilleures stratégies à utiliser dans un contexte d'optimisation des ressources.

L'intérêt de l'analyse de décision dans le domaine de la santé est de déterminer, avec le moins d'imprécision possible, la stratégie qui maximise le bénéfice attendu pour le patient mesuré par exemple par l'espérance de vie ajustée sur la qualité de vie.

Les étapes de l'analyse de décision sont les suivantes :

- la structuration du problème;
- l'identification des alternatives;
- la construction de l'arbre de décision;
- la détermination des probabilités liées aux événements consécutifs aux décisions prises;
- la détermination de la valeur des résultats de santé pour le patient;

- le calcul du bénéfice attendu de chaque stratégie;

- la réalisation de l'analyse de sensibilité.

A - Structuration du problème

Le point de départ de l'analyse est une situation clinique précise pour laquelle la stratégie diagnostique et thérapeutique est controversée. A l'instant de la décision médicale, l'ensemble des informations disponibles et les incertitudes relatives à l'une ou l'autre des stratégies ne permettent pas de connaître la stratégie qui maximise le bénéfice attendu pour le patient. Il est important de décrire précisément la condition du patient, en tenant compte de l'âge, du sexe, des antécédents, de la maladie présentée, des co-morbidités et du contexte social.

B - Identification des alternatives

L'analyse de décision repose sur la comparaison de plusieurs stratégies. La sélection de ces stratégies est faite après une analyse critique de la littérature. Les choix possibles sont ainsi sélectionnés en tentant de simplifier le plus possible les différentes attitudes envisageables pour le patient décrit. Pour limiter l'étendue des stratégies possibles, le choix du point de départ de l'analyse, c'est-à-dire la description clinique du patient dans l'espace et dans le temps, peut autoriser à sélectionner des stratégies entreprises une fois le résultat d'un test connu ou après l'application d'une thérapeutique. Il faut bien sûr veiller à ce que l'ensemble des alternatives identifiées recouvre l'ensemble des alternatives possibles.

On peut par exemple choisir comme point de départ un homme de 40 ans ayant présenté un syndrome coronarien aigu moins de 10 jours auparavant. Pour limiter le nombre de stratégies, on peut aussi choisir comme point de départ un homme de 40 ans ayant présenté un syndrome coronarien aigu moins de 10 jours auparavant et dont les facteurs de risque sont élevés.

C - Construction de l'arbre de décision

L'arbre de décision est une simplification d'une réalité souvent très complexe. Il est schématisé par un ensemble de noeuds reliés par des branches (fig. 1). L'arbre doit être assez complet pour représenter les éléments essentiels du problème, mais également assez simple pour la compréhension du modèle et les facilités de calcul. Les noeuds de l'arbre sont de trois types:

1 – Nœuds de décision (figurés par des carrés)

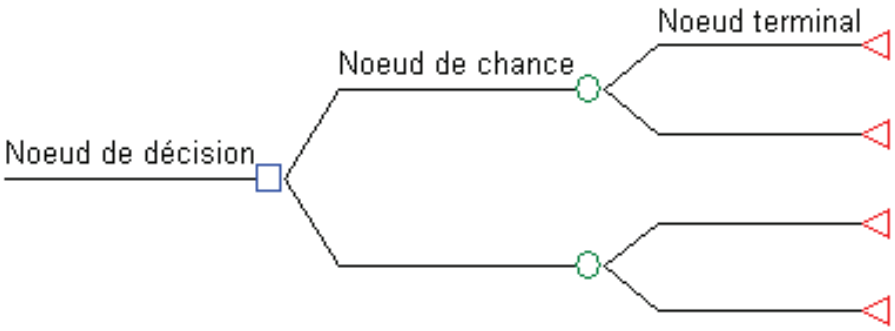
Ils représentent les choix à la disposition du décideur au moment où la décision doit être prise. A un noeud de décision sont associées dans l'arbre autant de branches qu'il y a de décisions possibles à ce niveau. L'arbre se construit de gauche à droite et le premier noeud à gauche est toujours un noeud décisionnel.

La question du problème doit être aussi limitée que possible. Les types de décision dans le champ de la médecine comprennent:

- la recherche d'une information supplémentaire, par exemple la prescription d'un test diagnostique;

- la mise en compétition d'options thérapeutiques, par exemple la chirurgie, le traitement médical, ou l'abstention thérapeutique.

Fig. 1 - Arbre de décision



2 – Nœuds aléatoires ou nœuds de chance (figurés par des cercles)

Ils correspondent à des phénomènes aléatoires qui ne sont pas sous le contrôle direct du décideur. De chaque noeud aléatoire sont issues autant de branches que le processus aléatoire admet d'événements. Ces événements doivent être exhaustifs et mutuellement exclusifs. Donc, si à chacune des branches est associée une probabilité, la somme des probabilités pour tout noeud aléatoire doit être égale à 1.

3 - Nœuds terminaux (figurés par des triangles)

Ils correspondent aux résultats de chaque trajet décisionnel. A chacun de ces noeuds terminaux est assignée une valeur numérique associée au résultat de la stratégie et exprimée dans la perspective du patient, de l'institution de soins, ou de la société. Il est habituel d'employer une échelle de valeurs homogène et arbitraire dont l'unité est définie identique pour chaque résultat de l'arbre (espérance de vie, mortalité, morbidité évitée, coût en euros, …). L'intervalle de temps sur lequel porte l'analyse est important car on doit tenir compte des résultats immédiats et de ceux estimés dans le futur.

D - Détermination des probabilités liées aux nœuds aléatoires

L'estimation des probabilités repose dans l'idéal sur des données objectives, chiffrées, provenant de la littérature ou d'études. Mais des données subjectives (avis d'expert) peuvent être utilisées lorsque les données objectives ne sont pas disponibles ou lorsque la controverse entre différentes études est trop grande.

E - Détermination de la valeur des résultats sur l'état de santé du patient

Différents types d'échelles de résultat peuvent être utilisés dans un arbre de décision:

- échelle arbitraire : de 0 à 1 ou 0 à 100;
- survie (0 à 1 ou 0 à 100%) : survie immédiate, à 1 an, ou à 5 ans;
- morbidité (temps libre de morbidité, ou temps d'hospitalisation);
- espérance de vie (analyse sur tableau de mortalité ou estimation de l'espérance de vie);

F - Calcul du bénéfice attendu de chaque stratégie

Le calcul du bénéfice attendu, c'est-à-dire la pondération des valeurs de résultats par les probabilités de chaque stratégie s'effectue de la droite vers la gauche. Le bénéfice attendu d'une stratégie est la somme des produits des probabilités par les valeurs des résultats de chaque branche. La différence observée entre les bénéfices attendus de chaque stratégie permet l'aide à la décision et montre le niveau de robustesse des résultats. Le choix de l'échelle de résultat reste cependant le point critique pour l'interprétation des résultats.

Prenons un exemple de la pratique courante : un homme âgé de 70 ans, qui a une artériopathie chronique des membres inférieurs développe un ulcère froid du pied droit. Il est hospitalisé pour intensifier le traitement médical. Après stabilisation de son état, vous demandez un avis chirurgical. Pour le chirurgien, une amputation au-dessous du genou devrait être faite immédiatement, attendu que l'échec du traitement médical pourrait obliger à effectuer plus tard une amputation au-dessus du genou, avec un taux de mortalité opératoire beaucoup plus élevé; la probabilité de la progression de l'ulcère sous une thérapie médicale intensive est de 50%; les taux de mortalité opératoire pour l'amputation au-dessous et au-dessus du genou sont respectivement de 1% et de 2% (fig. 2).

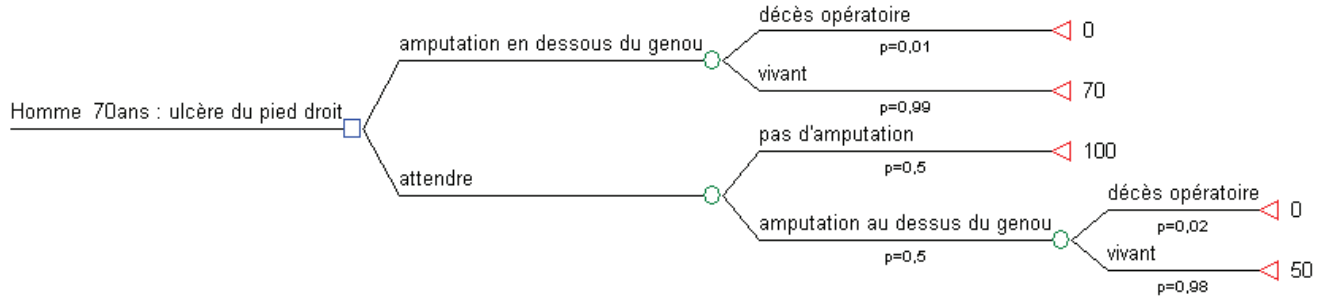
Le résultat de santé est apprécié sur une échelle arbitraire de 0 et 100 : une valeur de 0 a été choisie pour le décès, 50 au-dessus du genou, 70 pour l'amputation en dessous du genou et 100 pour la conservation des deux jambes.

Les bénéfices attendus des deux stratégies sont les suivantes (fig. 2):

- bénéfice attendu de l'amputation = 0 x 0,01 + 70 x 0,99 = 69,3
- bénéfice attendu de l'attente = 100 x 0,5 + (0 x 0,02 + 50 x 0,98)x0,5 = 74,5

Le bénéfice attendu est maximisé pour la stratégie « attente de l'amputation ».

Fig. 2 - Arbre de décision pour un ulcère froid du pied dans un contexte d'artériopathie chronique des membres inférieurs



G - Analyse de sensibilité

L'intérêt de l'analyse de sensibilité est d'examiner l'influence sur le résultat de la variation des paramètres entrant en jeu dans le processus d'analyse de décision. Tous les paramètres du processus peuvent être soumis à une variation, les probabilités comme les valeurs attribuées sur l'échelle de résultats.

Pour pouvoir être pratiqués de manière réaliste et reproductible, les calculs effectués pour une analyse de sensibilité sur une variable, deux variables, ou trois variables nécessitent une aide informatique. Les logiciels disponibles (*Data Tree-Age*…) permettent de plus une représentation graphique des seuils de décision proposés.

II - DE L'ANALYSE DE DECISION A L'ANALYSE COUT-UTILITE

L'évaluation médico-économique permet d'analyser les stratégies diagnostiques et thérapeutiques. Elle aborde la prise en charge du patient non plus par un examen des diagnostics différentiels pour parvenir au diagnostic "réel", mais plutôt par l'approche de conduite à tenir mettant en compétition différentes stratégies diagnostiques et thérapeutiques pour une situation clinique donnée. L'analyse coût-utilité introduit simplement cette réalité essentielle que chaque décision humaine peut être replacée dans un contexte de limitation des ressources.

L'utilité est définie par les économistes comme la satisfaction ou le bien-être associé à la consommation d'un bien ou d'un service. Appliqué à la santé et tenant compte de la notion nouvelle que le patient porte un jugement sur son état de santé, un indicateur définissant l'utilité doit tenir compte à la fois de l'espérance de vie et de la qualité de vie.

Pour l'estimation de l'utilité, il est possible d'ajuster l'espérance de vie sur la qualité de vie. C'est ce que l'on appelle couramment les années de vie ajustées sur la qualité de vie, les QALY (Quality Adjusted Life Years) ou définis comme une année pleine de vie sans limitation fonctionnelle ou symptômes morbides. Cet ajustement peut être fait de plusieurs manières. Pour la morbidité à long terme, on utilise une échelle catégorielle de qualité de la vie appréciée sur un tableau intégrant l'incapacité ou le handicap fonctionnel et l'état moral.

Dans notre exemple de l'ulcère froid, l'espérance de vie du sujet après intervention serait ajustée en fonction de la qualité de vie résultante de son amputation au-dessous ou au-dessus du genou. Cela permet de tenir compte de l'opinion du patient dans l'appréciation du résultat de la stratégie thérapeutique.

Prenons un autre exemple : un homme de 82 ans a des antécédents d'infarctus du myocarde postérieur compliqué d'insuffisance mitrale, et une fibrillation auriculaire. Quatre semaines après une première hospitalisation est apparue une défaillance cardio-respiratoire nécessitant son admission en soins intensifs. A l'examen il existe une fibrillation auriculaire avec une réponse ventriculaire rapide, un souffle de régurgitation mitrale, et une insuffisance cardiaque congestive modérée. Le traitement digitalo-diurétique et nitré a fait régresser les signes d'insuffisance cardiaque. L'échographie montre la même régurgitation mitrale avec une fraction d'éjection de 60% et une hypertrophie ventriculaire gauche modérée. La question du remplacement de la valve mitrale est posée, le cathétérisme cardiaque objective une fonction ventriculaire gauche normale (fraction d'éjection à 67%) et une sténose à 90% de la coronaire droite. Cet ancien avocat n'exprime aucune préférence personnelle eu égard à sa prise en charge médicale ou chirurgicale.

A - Sélection des différentes stratégies possibles

C'est la première étape de l'arbre de décision. Dans ce cas, il s'agit principalement de la prise en charge médicale et de la prise en charge chirurgicale.

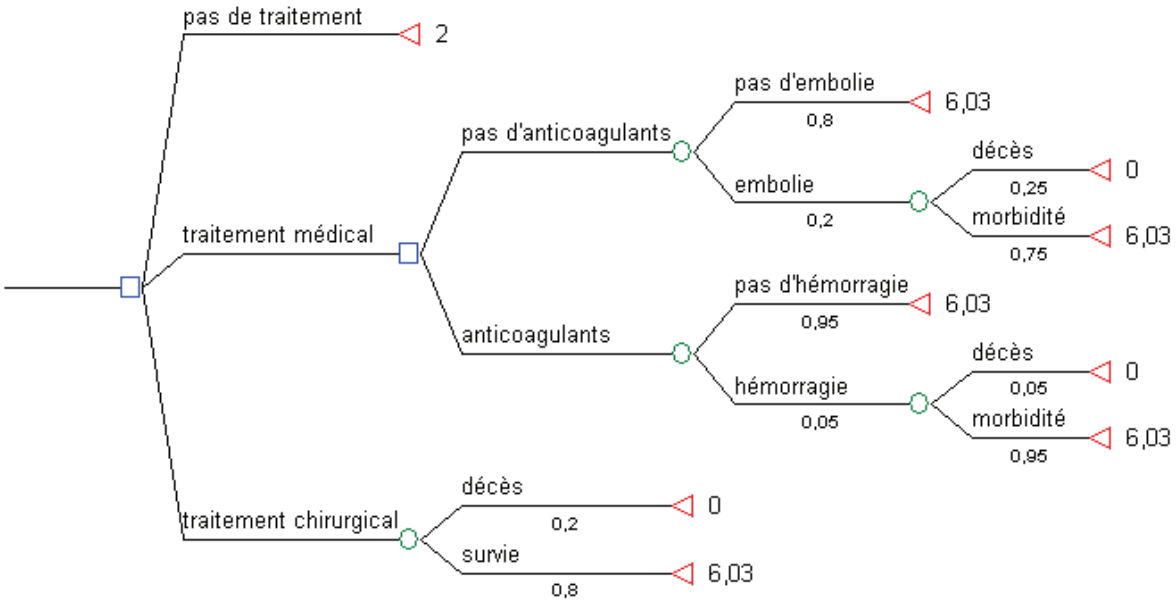
Un cardiologue a exprimé que ce patient peut bénéficier d'un remplacement mitral par une prothèse biologique, ce qui ne nécessiterait pas de traitement anti-coagulant.

Parce que la régurgitation mitrale est modérée, et compte tenu d'un risque de mortalité opératoire estimé à 20% pour ce patient dans ce centre, du fait également que la capacité fonctionnelle du patient ne devrait pas changer même après réussite chirurgicale, un autre expert a considéré qu'un traitement médical serait plus approprié.

Le choix est alors celui de l'opportunité ou non d'un traitement anti-coagulant. L'anti-coagulation entraînerait un risque d'hémorragie nécessitant l'hospitalisation dans 5% des cas, et parmi ces 5% un risque de décès de 5% (fig. 3). Ne pas décoaguler le patient entraînerait un risque d'accident embolique du fait de la fibrillation auriculaire de 20%, la mortalité due à cette embolie systémique est à peu près de 25%.

L'espérance de vie (EV) d'un homme de 82 ans peut être déterminée à partir des statistiques de mortalité et a été estimée dans ce cas à 6,03 années.

Fig. 3 - Arbre de décision pour un remplacement mitral dans un contexte d'insuffisance cardiaque



B - Survie et espérance de vie ajustée sur la qualité de la vie

Dans la plupart des études cliniques, la mesure habituelle de l'efficacité est la mortalité. Cependant, chez cet homme de 82 ans présentant une maladie qui peut affecter significativement chacune des stratégies, il n'est pas suffisant de ne tenir compte que de la mortalité. Il est important de tenir compte également de la qualité de chaque année de vie telle qu'elle peut être influencée par les symptômes et les limitations de mobilité liés à l'affection principale et aux co-morbidités.

La qualité de vie peut être mesurée en unités homogènes par année, ce qui permet la comparaison entre les différentes stratégies. Ces unités, qui sont appelées conventionnellement QALY (Quality Adjusted Life Years), correspondent au nombre d'années de vie gagnées ajusté par un facteur de pondération.

L'échelle utilisée pour la pondération chez ce patient est l'index de qualité de bien-être (Quality of Well Being Index: QWBI). Cet index exprime la qualité de vie en fonction des symptômes et des activités (activité sociale, activité physique, mobilité); il vaut 1 pour l'optimum asymptomatique et 0 pour la mort. Pour chaque stratégie, le total d'années de vie en bonne santé (QALY) est calculé en multipliant l'index de qualité de bien-être (QWBI) par l'espérance de vie que la thérapie est en mesure de produire.

D'autres techniques permettent de pondérer l'espérance de vie par les préférences du patient et notamment son estimation personnelle du risque de l'intervention médicale ou chirurgicale. On peut l'apprécier par la technique de la loterie ou jeu de hasard idéalisé: "je préfère vivre sûrement moins bien sans prendre le risque d'une intervention que vivre peut être mieux en prenant ce risque"; ou par la technique de l'arbitrage temporel : "je suis prêt à vivre moins longtemps dans un état meilleur que vivre plus longtemps dans un état moins bon".

Ces techniques fournissent des mesures plus sophistiquées que la simple mortalité, et fondées sur les préférences du patient.

Dans notre exemple, les QALY les plus élevées sont obtenues avec la thérapie médicale par anti-coagulant et sont un peu moindres avec la thérapie médicale sans anti-coagulant (tableau 1). Clairement la prise en charge médicale ou chirurgicale est de toute manière préférée à l'abstention thérapeutique. La prise en charge chirurgicale par prothèse biologique, efficace pendant 5 à 10 ans mais ne nécessitant pas de traitement anti-coagulant, procure un résultat un peu amélioré ajusté sur la qualité de vie que le traitement médical.

Tableau 1 - Calcul des Quality Adjusted Life Years (QALY)

Traitement	Probabilité	Espérance de vie	QWBI	QALY
Abstention	1,0	2	0,560 (an 1) 0,130 (an 2)	0,69
Médical sans anticoagulants		6,03		3,03
Pas d'AVC	0,8	6,03	0,734	2,75
AVC-morbidité	0,15	6,03	0,394	0,28
AVC-décès	0,05	0	0	0
Médical avec anti-coagulants		6,03		3,14
Pas d'hémorragie	0,95	6,03	0,673	3,00
Hémorragie-morbidité	0,0475	6,03	0,673	0,14
Hémorragie-décès	0,0025	0	0	0
Chirurgical		6,03		3,17
Vivant	0,8	6,03	0,878	3,17
Décédé	0,2	0	0	0

QWBI: Quality of Well Being Index

La différence de QALY entre la stratégie médicale avec anticoagulants et la stratégie chirurgicale est de 0,03 (tableau 1). Lorsque la différence est si faible, il devient intéressant d'intégrer les coûts dans l'analyse en leur appliquant une actualisation en fonction du temps. Il s'agit dans ce cas d'une analyse coût-utilité basée sur le coût par QALY. La stratégie la plus efficiente est le traitement médical par anticoagulants autant en coûts totaux qu'en coûts par unité de QALY (tableau 2). La chirurgie reste de loin la plus coûteuse.

Tableau 2 - Coût moyen par Quality Adjusted Life Years (QALY)

Traitement	Coût par espérance de vie	QALY	Coût / QALY
Abstention	0 €	0,69	0 €
Médical sans anticoagulants	22.412 €	3,03	7.397 €
Médical avec anticoagulants	7.224 €	3,14	2.301 €
Chirurgical	28.000 €	3,17	8.833 €

Une autre façon d'exprimer l'analyse coût-utilité est de considérer le coût marginal d'une année supplémentaire de vie ajustée sur la qualité de vie après chirurgie par rapport au traitement médical. C'est ce qu'on appelle une analyse incrémentale ou différentielle. Le coût supplémentaire de la chirurgie par rapport au traitement médical est de six cent quatre-vingt treize mille euros pour obtenir une augmentation d'une unité de QALY (tableau 3). Ceci ne signifie pas que la prise en charge chirurgicale entraîne une dépense d'un tel montant (coût

moyen). Cela permet de calculer le coût d'une unité supplémentaire (coût marginal) par rapport au coût moyen.

Tableau 3 - Gain différentiel et coût marginal

Traitement	QALY gagnées	Coût différentiel (CD)	CD par QALY gagnée
Abstention	----	----	----
Médical avec anticoagulants	3,14 - 0,69 = 2,45	7.224 – 0 = 7.224 €	7.224 / 2,45 = 2.949 €
Chirurgical	3,17 - 3,14 = 0,03	28.000 - 7.224 = 20.776 €	20.776 / 0,03 = 692.533 €

Ce cas illustre les concepts de base de l'analyse coût-utilité et montre comment ce type de réflexion s'intègre dans le choix de stratégies concurrentielles en tenant compte de l'efficacité thérapeutique, de la qualité de vie du patient, des préférences du patient et de la contrainte de ressources.

C - Analyse de sensibilité et actualisation

L'analyse de sensibilité permet de tenir compte de l'influence que pourrait produire le changement de valeur d'une des variables.

Dans notre exemple, si la différence artério-veineuse en oxygène est plus basse, les QALY pour les stratégies traitement médical avec et sans anti-coagulant augmentent à 3,68 et 3,63 respectivement (au lieu de 3,14 et 3,03). Une autre variable qui pourrait faire changer les résultats est la mortalité chirurgicale: si elle passe à 30%, les QALY chirurgicales baissent à 2,76 (au lieu de 3,17).

Ainsi après analyse de sensibilité, on remarque que le traitement médical reste le choix préféré.

Il est indispensable de prendre en considération dans l'analyse le facteur temps. Cela nécessite de choisir un cadre temporel et de fixer un taux d'actualisation.

Ce procédé permet d'apprécier la valeur actuelle d'un service qui sera effectué dans le futur. Il exprime la préférence d'une société entre la consommation ou l'investissement. La valeur actuelle du coût d'un traitement effectué aujourd'hui n'est pas équivalente à celle du même traitement effectué dans une décennie. Ainsi avec un taux d'actualisation i, un événement survenant dans n années a une valeur actuelle de $\frac{1}{(1+i)^n}$.

III – CONCLUSION

L'analyse de décision n'a pas pour objectif de modéliser le comportement humain en matière de décision. Elle n'apporte pas non plus la vérité scientifique sur un sujet déterminé (tableau

4). Elle permet d'apporter une aide à la décision dans un contexte d'incertitude en tenant compte à la fois des données épidémiologiques, des résultats d'études, et des opinions d'experts. Elle s'adresse autant à une question limitée dans le cadre de la relation médecin-patient qu'à un problème d'allocation de ressources en santé publique et représente un outil de nature scientifique dans l'arsenal des méthodes à la disposition des médecins.

Tableau 4 - Avantages et inconvénients de l'analyse de décision clinique

AVANTAGES	INCONVENIENTS
- Apporte une structure simple - Permet de combiner plusieurs sources de données - Autorise la considération de l'utilité (variable patient) - Permet d'examiner l'impact de données subjectives - Sépare un large problème complexe en plusieurs plus petits plus faciles à gérer - Fournit une représentation du raisonnement clinique	- Encourage les simplifications réductrices - Nécessite les données - Peu familier - Consomme beaucoup de temps - Fournit une représentation du raisonnement clinique

Bibliographie

Drummond M, O'Brien B, Stoddard G, Torrance GW. Méthodes d'évaluation économique des programmes de santé. 2ème édition. Edition Economica 1998.

Torrance GW, Blaker D, Detsky A, Kennedy W, Schubert F, Menon D et al. Canadian guidelines for economic evaluation of pharmaceuticals. Pharmaco Economics 1996;9:535-559.

Gold MR, Siegel JE, Russell LB and Weinstein MC. Cost-effectiveness in health and Medicine, Oxford University Press, New York, 1996.

Russell LB. Modelling for cost-effectiveness analysis. Statistics in Medicine 1999;18:3235-3244.

Keeler E. Decision trees and Markov models in cost-effectiveness research. In : Valving health care : costs, benefits and effectiveness of phamaceuticals and other medical technologies. Cambridge : cambridge university press;1995:185-205

CHAPITRE XII

LES POINTS CLE DE LA VALIDITE INTERNE D'UNE ETUDE BIAIS ET FACTEURS CONFONDANTS.

P Duhaut, J Schmidt

Toutes les méthodes de recherche en épidémiologie clinique (ou en médecine basée sur les faits) consistent à essayer d'établir les principes gouvernant notre pratique clinique sur des faits observés, quantifiables, mesurés, vérifiés, plutôt que sur des affirmations théoriques, de l'empirisme simple, voire des croyances médicales répandues ayant, d'une façon ou d'une autre, pris racine dans notre savoir ou nos attitudes. Cet effort a sans doute permis des progrès notables. Aucune méthode d'investigation ou d'observation n'est cependant parfaite, et la littérature regorge de résultats contradictoires ou différents. L'épidémiologie clinique, en essayant de dégager des tendances, n'établit pas de vérité 'dure', comme la découverte de l'ADN et son rôle dans la transmission du génome. Ce chapitre passe en revue quelques unes des limites d'interprétation des études.

L'Epidémiologie clinique, dont les méthodes sont en partie inspirées des méthodes d'Epidémiologie classique, a été développée afin d'aider le clinicien à répondre aux multiples questions relevant de sa pratique quotidienne dans les domaines de la prévention, du diagnostic, de la thérapeutique, de l'évaluation du pronostic, de la nosologie... :
Quels sont les facteurs de risque du cancer du colon ? Quelle est la meilleure façon de prévenir les complications de l'hypertension artérielle ? Quelle en est l'incidence ou la prévalence au sein de la population âgée ? Quel est le meilleur traitement de l'infarctus du myocarde pris dans les premières heures ? Quelle est la meilleure stratégie thérapeutique d'un cancer du sein stade II ? Quelle est la démarche diagnostique optimale face à une fièvre au long cours ? Quelle est la sensibilité et la spécificité des tests sérologiques de deuxième génération de l'hépatite C ?...

Chaque médecin, en fonction de son expérience, de ses habitudes thérapeutiques, de son école de formation, peut avoir une opinion fondée sur des bases lui paraissant raisonnables... Opinion souvent différente de celle d'autres médecins expérimentés, voire diamétralement opposée.

Les études cliniques cherchent à résoudre ces interrogations ou ces contradictions en quantifiant les faits sur des bases reproductibles à partir d'un nombre de patients suffisamment important pour obtenir une réponse statistiquement significative. Le résultat doit -en principe- être applicable au prochain patient atteint de la même pathologie.
Les outils méthodologiques de l'analyse épidémiologique et statistique ont été développés dans ce but. Force nous est de constater, cependant, qu'existent entre de nombreuses études apparemment bien conduites des différences de conclusion ou des contradictions aussi criantes que celles existant entre les opinions individuelles.

Les origines de ces différences sont multiples.
Elles relèvent tout d'abord d'un possible **manque de précision** d'une étude correctement construite.

Elles peuvent correspondre ensuite à la présence de **vices cachés**, appelés biais, plus ou moins inhérents à la structure des études en question. Les biais conditionnent la **validité interne** de l'étude.
Elles dépendent en troisième lieu des **caractéristiques des populations étudiées** : une même maladie peut avoir des causes différentes, et un même traitement peut avoir une efficacité variable, dans différentes populations. Ceci introduit la notion de **validité externe**, ou capacité à être généralisée, de l'étude.
Elles font appel enfin à la connaissance des limites d'applicabilité de l'outil épidémiologique et statistique à l'analyse et à la compréhension d'un phénomène.

Ces questions résolues -s'il est possible- laissent entier le problème de l'application à un individu de résultats obtenus au niveau d'un groupe. Problème particulièrement important pour le clinicien, qui souvent doit se référer aux études de groupe pour déterminer la cause de la maladie d'un individu, en choisir le traitement ou en inférer le pronostic... et paradoxe des études de groupe, dont le but premier et ultime est souvent d'aider le clinicien à répondre aux questions qu'il se pose face à l'individu.

II- NOTION DE PRECISION ; ERREUR LIEE AU HASARD :

Imaginons un essai randomisé cherchant à comparer un protocole A et un protocole B de chimiothérapie dans les cancers épidermoïdes du pharynx (excluant les cancers peu différenciés liés au Virus d'Epstein-Bar du nasopharynx), conduit dans deux unités de cancérologie différentes I et II.
Supposons ces deux unités de cancérologie ouvertes à tout patient susceptible de présenter un cancer ORL, sans distinction aucune de sexe, d'origine, d'âge, de milieu social, d'exposition à un facteur de risque particulier ou de tout autre facteur. Tous les nouveaux patients présentant un carcinome épidermoïde hospitalisés durant la période de réalisation de l'étude ont été inclus et randomisés entre les bras A et B en double aveugle, éliminant en principe tout biais de sélection. L'unité I a inclus 23 malades et l'unité II 19. Les données, exposées ci-dessous, sont analysées dans chaque service séparément.

	Unité 1		Unité 2	
Protocole	A	B	A	B
Patients inclus	12	11	8	11
Survie à 1 an	6	4	3	5
Survie à 1 an (%)	50%	36%	37%	45%

Ces deux essais, identiques, correctement conçus, sans biais évident, portant sur le même type de malades diagnostiqués de la même façon, semblent donner des résultats contradictoires. Quelles peuvent en être les raisons ?

Une première explication peut être donnée par le hasard d'échantillonnage : quoiqu'il n'y ait pas de biais de sélection au niveau de l'inclusion dans l'étude (puisque tous les patients ont été inclus et que les deux unités sont ouvertes à tous les patients) ou dans un bras (puisque l'allocation a été randomisée), les patients issus de l'Unité I ou de l'Unité II ne représentent que deux échantillons tirés au hasard de l'ensemble de la population des patients atteints de cancer ORL épidermoïde dans le monde. Ces deux échantillons peuvent par hasard ne pas être équivalents entre eux, ou ne pas être également représentatifs de la population globale. Ils peuvent par conséquent répondre de façon différente à un traitement A ou B sans que cela soit

lié au traitement lui-même : la disparité, réelle, observée entre l'Unité I et II dans l'efficacité du traitement correspondra en fait à la différence non connue, survenue au **hasard d'échantillonnage**, entre les deux groupes.

On peut supposer maintenant que la différence observée ne soit pas réelle, ne soit pas significative. Un petit nombre seulement de patients ont été inclus, et les différences de 14 % dans un sens et de 12 % dans l'autre observées dans l'Unité I et II ne tiennent chaque fois qu'à un individu : une survie supplémentaire survenue par hasard dans le bras B de l'unité I ramènerait le succès de B à 45 %, et une survie supplémentaire survenue par hasard dans le bras A de l'unité II le succès de A à 50 %, homogénéisant ainsi les résultats des deux unités... et faisant disparaître toute différence entre les traitement A et B, de façon concordante dans les deux centres!

Des résultats obtenus à partir d'un petit nombre de patients éveillent un certain scepticisme. Il existe cependant des "garde-fous" statistiques, exprimés sous la forme d'intervalle de confiance, permettant de "quantifier" le hasard dans l'obtention des résultats et par conséquent leur significativité. On admet en général que des résultats sont significatifs lorsqu'ils ont moins de 5 % de chances d'être dus au hasard. Ainsi, on appelle l'erreur toujours possible de conclure à une différence significative n'existant pas réellement : alpha.

En d'autres termes, on demande le plus souvent que **l'erreur alpha** soit inférieure à 5%.

Oublions maintenant l'unité II et admettons que 5 patients soient vivants à un an dans le bras B de l'unité I, ramenant ainsi le succès de ce bras à 45 %. Il n'y a plus de différence notable entre les traitements A et B. Est-ce à dire que les deux traitements sont d'efficacité (ou d'inefficacité) comparables ? Ou que l'étude a failli à montrer une différence existant réellement ?
Une étude doit inclure suffisamment de patients, être suffisamment puissante, pour avoir une chance acceptable d'arriver à montrer une différence existant réellement. Il existe là encore, un "garde-fou" statistique permettant de quantifier la puissance d'une étude, c'est-à-dire sa chance d'arriver à une conclusion reflétant la réalité que l'on cherche à estimer : une bonne étude doit avoir au moins 80 % de chances d'arriver à résoudre la question qu'elle s'était posée, soit une puissance de 80 %. L'erreur, toujours possible, de conclure à tort qu'il n'y a pas de différence entre les groupes comparés parce que l'étude a failli à montrer une différence existant réellement, est appelée **l'erreur bêta.**

Autrement dit, on demande en général que l'erreur bêta soit inférieure à 20 %, ou que la puissance soit supérieure à 80 %.
Puissance = 1 - bêta.

Hasard d'échantillonnage et résultats obtenus au hasard de l'erreur alpha ou bêta représentent donc les sources majeures d'erreur liées au manque de précision d'une étude par ailleurs correctement construite et menée. Il est possible, pour diminuer ce risque d'erreur liée au hasard, d'augmenter la taille de l'échantillon étudié : on augmentera ainsi la puissance de l'étude, on augmentera la significativité des résultats, et on améliorera les chances que l'échantillon soit véritablement représentatif de la population de patients atteints de la maladie que l'on se propose d'étudier.

Autrement dit, augmenter la taille de l'échantillon étudié diminue l'erreur alpha, diminue l'erreur bêta, et diminue les erreurs liées au hasard d'échantillonnage.
Une revue systématique des essais effectués en anesthésie a ainsi montré qu'en 2000, seuls 56 % des essais randomisés avaient une puissance suffisante. Le score s'est amélioré à 86 % en 2006, mais seuls 18 % des essais à résultats négatifs analysaient leur risque d'erreur beta [1]. Cette étude reflète assez bien la tendance générale et il est rare que la puissance d'une étude soit considérée avant de conclure, peut-être à tort, à l'absence d'effet.

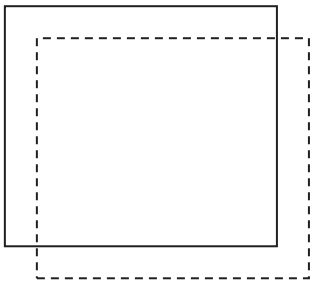
II- LES BIAIS :

Contrairement aux erreurs précédentes, les biais sont à l'origine d'erreurs relatives à de véritables vices de forme de l'étude, apparaissant à l'une quelconque des étapes de sa conception. Un biais est -en principe- un travers que l'on doit s'efforcer de réduire au minimum ou de faire disparaître - cela n'est pas toujours possible -, avant de débuter l'étude sur le terrain. Un biais n'est pas dû au hasard.
Les vices de forme sont nombreux et certains types d'études prédisposent plus particulièrement à tel ou tel type de biais. Olli Miettinen a proposé une classification en trois grandes catégories :

1- Les biais de sélection
2- Les biais de mesure, ou d'information, ou de mauvaise classification
3- Les éléments confondants

1- Les biais de sélection :

Les biais de sélection surviennent lorsque la population effectivement étudiée n'est pas représentative de la population que l'on voulait étudier, et à laquelle on voudrait pouvoir appliquer les résultats.



Trait plein : population que l'on veut étudier
Trait pointillé : population effectivement atteinte

Plusieurs facteurs peuvent concourir à produire un biais de sélection, qui peut être suffisamment important pour entacher la validité des résultats :

Les **critères d'inclusion et d'exclusion** peuvent ne pas avoir été adéquats. Imprécis, laissant une trop grande incertitude quant à leur interprétation et permettant ainsi, dans le cadre d'une étude multicentrique, l'inclusion d'une population hétérogène et différente selon les centres participants. Trop restrictifs ou trop précis au contraire, excluant de l'étude un grand nombre de sujets et mettant ainsi en péril la validité externe de l'étude : les résultats ne pourront être

appliqués au patient tout venant atteint de la pathologie étudiée, car trop différent du "modèle idéal" sur lequel l'étude a été réellement basée.
Dans une étude cas-contrôle, la sensibilité et la spécificité des tests ou des critères diagnostiques de la maladie étudiée sont déterminants pour la sélection des cas et des témoins : trop sensibles, ils aboutiront à l'inclusion, dans le groupe des cas, de patients non atteints de la maladie étudiée. Trop spécifiques, ils aboutiront à l'exclusion de patients réellement atteints de la maladie. Une bonne spécificité est cependant préférable à une bonne sensibilité pour la validité de l'étude [2].
La sélection des témoins dans une étude cas-contrôle représente un élément très important de la construction de l'étude, dont les résultats sont basés sur la comparaison du groupe de patients et du groupe de témoins. Le choix idéal des témoins est réalisé par un tirage au sort dans la population dont sont issus les cas (voir chapitre V). Ceci n'est pas toujours possible. Ainsi, de nombreuses études recrutent les témoins parmi la famille, les amis ou le voisinage des cas.
Le **biais de non-réponse** peut représenter un biais de sélection important, alors même que les étapes précédentes (choix des critères d'inclusion et d'exclusion, choix de la population témoin dans une étude cas-contrôle) ont évité les embûches. Il peut survenir lorsqu' un certain nombre de patients pressentis pour entrer dans l'étude ne répondent pas au courrier ou à l'appel téléphonique les y invitant, ou refusent de participer à l'étude après avoir été informés.
Les sujets non-répondeurs ou refusant de participer à l'étude peuvent être différents des sujets finalement inclus, qui dès lors ne seront plus représentatifs de l'ensemble de la population initialement ciblée. Certains auteurs se sont attachés à étudier les sujets non-répondeurs et ont effectivement mis en évidence des différences socio-démographiques suffisamment importantes pour modifier les résultats de l'étude entreprise [3].
Les biais de sélection possibles peuvent être évidents ou plus cachés :

Exemples :
- Le taux de réponse au questionnaire de Richard Doll sur la prévalence de la consommation excessive d'alcool et ses complications potentielles dans la cohorte des médecins britanniques a été de 73 %, versus 95 % pour les questionnaires s'intéressant aux complications du tabagisme. La réduction de 22 % correspond sans doute à un biais de sélection important, les non-répondeurs pour cette question socialement sensible étant sans doute différents des répondeurs. Dès lors, la mesure des complications du facteur de risque est biaisée et les résultats de l'étude difficiles à interpréter, du moins sur le plan quantitatif [4].
- Le *biais de survie sélective* constitue un autre biais de sélection possible : un traitement ne pourra s'appliquer, bien sûr, qu'à des patients vivants. Ceux-ci, dans la mesure où le temps 0 du début d'une maladie est rarement connu, peuvent ne constituer que la fraction survivante de la population initialement atteinte, et donc la fraction de meilleur pronostic. Une étude récente a cherché à quantifier ce biais dans l'expression des résultats de la prise en charge de patients admis à l'hôpital pour infarctus du myocarde : elle a pu montrer que le biais de survie augmentait avec le délai d'admission à l'hôpital, et que 'l'efficacité' mesurée d'un traitement objectivement inefficace pouvait, en réduction de mortalité, passer de 4 à 27 % si le caractère changeant de la maladie en fonction du temps passé était ignoré [5] : les survivants, par définition, ont un meilleur pronostic que les patients décédés rapidement, et ce meilleur pronostic peut à tort, être mis sur le compte du traitement. Ce biais peut survenir aisément dans toute étude cas-témoins ou de cohorte : ainsi, la mortalité dans la 'Cardiovascular Health Study' (CHS), pourtant basée sur la population, était inférieure de 40 % à celle observée dans plusieurs cohortes de patients enregistrés dans des programmes d'assurance-maladie de type Medicare aux Etats-Unis, *alors que les patients*

de la CHS étaient en moyenne plus âgés et que la proportion d'hommes y était plus importante [6]. La randomisation d'une population particulière pour un essai thérapeutique ne le fait pas disparaître, et divers biais de sélection peuvent toujours influencer les résultats de l'essai [7].

Lorsque l'inclusion de témoins par tirage au sort n'est pas possible, il vaut mieux, d'une manière générale, préférer les contrôles recrutés dans le voisinage du patient aux amis directs : ils sont plus susceptibles d'être issus de la population dont sont issus les cas, sans pour autant partager toutes les habitudes ou facteurs de risque des cas comme pourraient le faire les amis (notamment en terme de consommation tabagique ou alcoolique) [8].

Le **biais de détection** est un autre exemple de biais de sélection, souvent très difficile à mettre en évidence ou à quantifier.

Exemples :

- Un biais de détection possible a été longuement discuté dans les diverses études cas-témoins s'intéressant à la relation entre oestrogénothérapie substitutive de la ménopause et survenue d'un cancer de l'endomètre (voir chapitre V) [9], avec des résultats contradictoires entre études montrant pour le moins qu'une étude construite pour éviter un biais peut être prise à son propre piège [10].
- L'utilité du dépistage du cancer du poumon a été longtemps débattue, et les principales études menées en Europe ou aux Etats-Unis fondées sur la réalisation d'une radiographie pulmonaire simple par an n'ont pas montré d'avantage net en termes de survie des patients dépistés. Une étude récente effectuée au Japon, avec réalisation de scanners thoraciques systématiques, semble mettre en évidence une amélioration de la survie des patients dépistés. Deux questions non résolues restent posées : celle d'un biais de détection précoce du cancer allongeant tout simplement la durée d'observation des patients, et donc la survie observée, et celle d'un biais de détection tout court de petits cancers moins agressifs non diagnostiqués sur la radiographie pulmonaire simple, et dont le pronostic, potentiellement meilleur, modifie les données de survie auparavant collectées sur des patients plus graves [11].
- L'administration de finastéride, traitement de l'hypertrophie bénigne de la prostate, a été associée dans un essai randomisé à une diminution de 25 % de la prévalence de cancer prostatique objectivé par biopsie par rapport au placebo. En revanche, la prévalence de cancer de haut grade sous finastéride était supérieure à celle sous placebo. En essayant d'en comprendre les raisons, les auteurs ont montré que la distribution du PSA dépendait du volume prostatique avec une diminution de la surface sous la courbe ROC (voir chapitre IX) associée à l'augmentation du volume prostatique, et ceci en fonction du grade du cancer. Autrement dit, les PSA étaient plus performants dans le diagnostic de cancer de haut grade sur des prostates de petit volume, par comparaison aux prostates augmentées de volume, et l'augmentation apparente de cancer de la prostate de haut grade dans le groupe finastéride était en partie au moins, liée à leur détection facilitée par la diminution de volume de la prostate… elle-même en relation avec la prise de finastéride [12].

Un biais de détection peut donc, en fonction de la pathologie étudiée et des conditions d'étude, jouer dans le sens du diagnostic de pathologies moins graves avec amélioration apparente, mais artéfactielle, du pronostic, ou dans le sens du diagnostic de pathologies plus graves avec aggravation apparente du pronostic.

Chaque étude peut, en fonction de sa structure et de la pathologie impliquée, générer ses propres biais de sélection et en produire de nouveaux, non encore catalogués. Leur recherche doit être systématique dans la construction de l'étude et systématique encore lors de la lecture des résultats.

2- Les biais de mesure :

La sélection des patients (et des contrôles le cas échéant) étant réalisée de la façon la moins biaisée possible, peuvent survenir des **biais de mesure** encore appelés **biais d'information ou de mauvaise classification** : l'objet d'étude (le facteur de risque dans une étude cas-témoins, l'incidence d'une maladie dans une étude prospective en cohorte, l'effet d'un traitement dans un essai randomisé, la prévalence d'un facteur de risque ou d'une maladie dans une étude horizontale) a été mal mesuré. Les raisons peuvent en être multiples.

Le diagnostic d'une maladie, la mesure d'un effet, la détermination d'un facteur de risque dépendent étroitement de la *sensibilité et de la spécificité de la méthode* mise en oeuvre pour les reconnaître (test diagnostique, examen complémentaire, échelle qualitative ou quantitative, qualité de la question posée à l'interrogatoire).

Exemples:

- Le cancer colorectal compte parmi les premiers cancers affectant les deux sexes, et le dépistage de polypes ou de lésions de petite taille, si possible par des moyens non invasifs, devrait permettre d'éviter l'évolution vers les cancers invasifs. La vidéocapsule dans un essai récent a été comparée à la colonospie conventionnelle. Pour des polypes de plus de 6 mm de diamètre, la Se (sensibilité) n'a pas excédé 64 % et la Sp (spécificité) 84 % ; des chiffres similaires ont été retrouvés pour les adénomes avancés. La vidéocapsule appliquée sur de larges populations conduirait pour l'instant à de nombreux faux positifs et faux négatifs, et donc, si l'on voulait comparer le pronostic d'une population dépistée et non dépistée, à un biais vers le nul par 'mixage' des deux populations [13].
- L'examen anatomo-pathologique est habituellement considéré comme le 'gold standard', ou étalon-or, pour le diagnostic de lésions bénignes ou malignes : nous n'avons pas mieux pour l'instant que le microscope, éventuellement aidé de méthodes de marquage ou de biologie moléculaire, pour le diagnostic étiologique d'une lésion. Quatre anatomo-pathologistes spécialistes en pathologie gastro-intestinale se sont intéressés à leur reproductibilité inter-observateurs sur l'examen de polypes colorectaux, pour trouver que le coefficient de reproductibilité kappa, prenant en compte les diagnostics concordants par simple hasard, ne dépassait pas… 49 % [14].

Il faut bien sûr s'entourer d'un maximum de précautions dans l'utilisation des tests diagnostiques, ou des mesures d'exposition. Malgré cela, l'étude précise des tests, y compris des tests étalon-or, met en évidence leurs limites et les résultats des études les utilisant doivent être interprétés avec ces limites présentes à l'esprit. L'étude de la reproductibilité, notamment, met en évidence les limites de validité d'un test, même lorsqu'il est appliqué par un médecin expérimenté.

Le **biais de souvenir** auquel sont particulièrement exposées les études rétrospectives est un autre inducteur fréquent de biais d'information.

Exemple : il a été rapporté que les avortements légaux pouvaient représenter un facteur de risque du cancer du sein. Lindefors-Harris et coll. ont comparé deux études cas-témoins, la première utilisant comme source de données des interviews de patientes et de leurs témoins, la seconde recueillant l'information de façon plus objective dans le registre national suédois des avortements légaux. Le risque relatif de cancer du sein en regard des antécédents d'avortement s'est avéré être 1,5 fois plus important dans l'étude basée sur les interviews que dans l'étude utilisant les registres (différence significative). De plus, le rapport entre la sous-estimation des avortements chez les témoins et la surestimation chez les cas (données d'interrogatoire comparées aux données objectives du registre) s'est avéré être égal à 22,4 ! [15].

Les sujets malades ont tendance à vouloir trouver une explication à leur maladie, alors que les témoins sains ont tendance à oublier leur passé médical... Le biais de souvenir peut donc expliquer des odds ratio anormalement élevés, reflétant mal ou peu la réalité et inducteurs de conclusions faussées. L'importance de l'erreur au niveau du résultat dépend cependant des conditions des facteurs examinés (facteur de risque dans une étude cas-témoins) et peut être minime lorsque la prévalence du facteur de risque est faible [16].

Les biais de mesure peuvent se rencontrer sur tous les types de variables étudiées, que ce soit des variables quantitatives de type dosage biochimique (les techniques de dosage ont leur imprécision), des variables semi-quantitatives comme l'appréciation d'un stade de tumeur (I, II, III, ou IV), car il existe des limites de détectabilité des localisations secondaires viscérales ou ganglionnaires, aux variables qualitatives comme certaines données d'interrogatoire (antécédents personnels ou familiaux « oubliés ») ou d'examen anatomo-pathologique ou radiologique (en faveur, ou non, de tel ou tel diagnostic).

Il existe enfin des **biais de mauvaise classification** liés aux erreurs de manipulation des données, aux erreurs de remplissage de questionnaires, voire aux données fausses fournies sciemment par des investigateurs peu scrupuleux...

Lorsque ce biais de mauvaise classification s'effectue de façon bi-directionnelle (patients du groupe A malencontreusement inclus dans le groupe B, et patients du groupe B malencontreusement inclus dans le groupe A), les deux groupes que l'on cherchait à comparer, que l'on supposait initialement différents, s'uniformisent. La différence disparaît alors artificiellement et peut à l'extrême s'annuler. Il s'agit du "**bias toward the null**" anglo-saxon. Cette vérité générale peut souffrir quelques exceptions, en cas d'exposition multiple en particulier [17].
Lorsque le biais s'effectue de façon uni-directionnelle, le résultat dépendra du sens du biais ! Dans une étude rétrospective avec surestimation du facteur de risque dans le groupe des patients et sous-estimation dans le groupe contrôle (cas de figure le plus souvent rencontré), il y aura accroissement de l'odds ratio et par conséquent surestimation du risque... Dans une étude prospective de cohorte où les perdus de vue se rencontreraient essentiellement chez les sujets exposés (exemple : employés exposés d'une entreprise ayant fait faillite, obligés de quitter la région pour des raisons professionnelles), le risque relatif peut être artificiellement abaissé par défaut d'information uni-directionnel sur les perdus de vue (cas de figure en fait inhabituel en pratique).

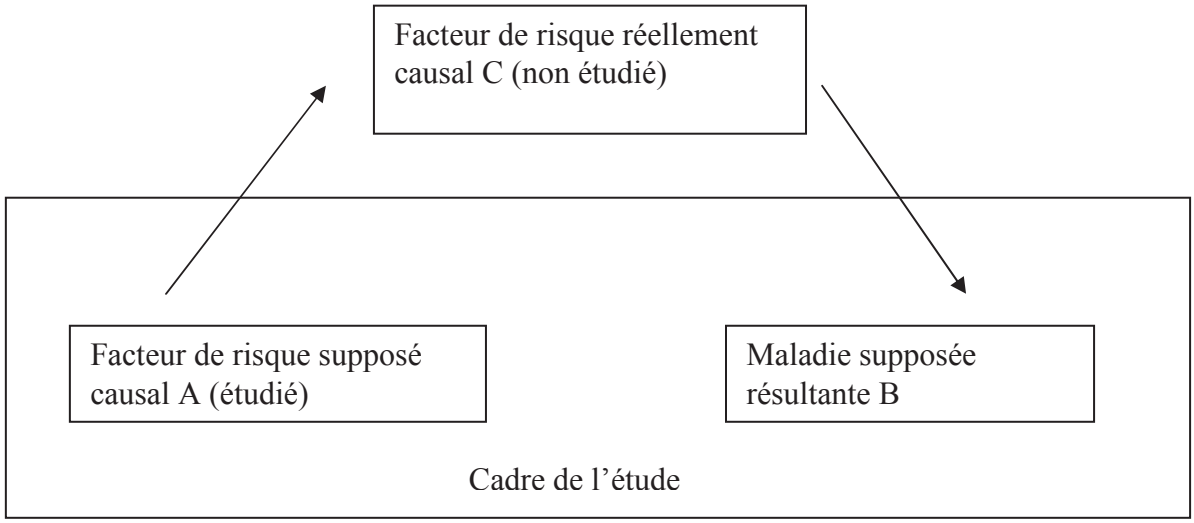
3- Les éléments confondants :

Indépendant des deux premiers types de biais, le **biais de confusion** se produit lorsque l'association observée sur le plan statistique ne correspond pas à une réalité biologique,

pathologique ou étio-pathogénique, mais s'explique en fait par un troisième facteur, réellement impliqué dans la physiopathologie de la maladie étudiée.

Imaginons une étude de cohorte étudiant l'effet du facteur de risque A sur la survenue de la maladie B. Une étude de cohorte ne peut pas tester de nombreuses hypothèses étiologiques (contrairement à une étude cas-témoins rétrospective), et le facteur de risque A doit donc être fortement suspecté d'être causal avant le début de l'étude prospective. Imaginons maintenant que l'on observe, dans la part de la population soumise au facteur de risque A, une incidence de la maladie B plus élevée que chez les sujets non exposés. En conclure que A est véritablement un agent étiologique semble logique : le facteur de risque était présent avant que la maladie ne se déclare, et représente justement le facteur distinctif entre les sujets devenus malades et les sujets restés sains.

Ceci est vrai à une condition : qu'il n'existe pas un ou plusieurs facteurs C, ou éléments confondants, non mesurés dans l'étude, présentant d'une part une association statistiquement significative (correspondant à une réalité biologique ou non) avec le facteur de risque A étudié, et réellement impliqués d'autre part dans la genèse de la maladie B.



Le facteur C, réellement causal mais ignoré dans le cadre de l'étude, fait donc apparaître du fait de son association avec le facteur A, ce dernier comme responsable de la maladie B dans une association fallacieuse, statistiquement significative mais biologiquement non fondée.

Conséquence immédiate :

- Le véritable agent étiologique n'aura pas été reconnu malgré le travail de recherche entrepris.

Conséquences possibles :

- Une action d'éradication du facteur A dans le but de prévenir l'apparition de la maladie B peut s'avérer tout à fait inefficace pour peu que l'association entre A et C ne soit que statistique et non pas biologique. Autrement dit, l'élimination du facteur A, si A n'est pas physiquement lié à C, n'entraînera pas l'élimination du véritable facteur étiologique et n'aura pas d'effet sur l'incidence de la maladie.

- Un agent thérapeutique élaboré contre le facteur A dans le but de guérir la maladie B n'atteindra pas l'effet espéré.
- Des programmes de recherche peuvent être lancés sur une fausse piste (situation non exceptionnelle permettant parfois de faire avancer la connaissance!).

Exemple : Le SIDA est une maladie dont l'étiologie est connue, mais la transmission du virus est plus fréquente dans des groupes dits à risque. La prévalence élevée dans la maladie dans le Bronx New-Yorkais a motivé la réalisation d'une étude visant à déterminer les facteurs de risque particuliers des sujets résidents utilisateurs de drogues. La prévalence du SIDA s'est avérée trois fois plus élevée chez les patients de race noire comparés aux malades de type caucasien, et la tentation de conclure à une susceptibilité augmentée à l'infection chez les premiers est facile...
En fait, le type de drogue employé n'était pas le même dans les deux groupes pour des raisons économiques. La cocaïne était utilisée préférentiellement par les Noirs, et l'héroïne principalement par les Blancs. La cocaïne, du fait de sa durée d'action plus brève, nécessitait trois à quatre fois plus d'injections que l'héroïne et le risque potentiel de transmission du virus se trouvait augmenté d'autant [18]. Une analyse des données contrôlée par le facteur confondant (type de drogue) aurait permis une conclusion plus précise.

Le **biais de confusion** peut se glisser insidieusement dans les meilleures études, à différentes étapes de sa réalisation. Ainsi, Vach et Blettner décrivent comment l'utilisation correcte de méthodes de traitement des données manquantes dans les études cas-contrôle peut amener à créer un biais de confusion indépendant de la construction même de l'étude, mais apparaissant lors de la phase d'analyse. Le remède consiste à étudier de façon plus précise, la répartition et le pourquoi des données manquantes dans les différents sous-groupes [19].

Les études réalisées à partir de bases de données informatisées constituées sont sujettes tout particulièrement au biais de confusion. Les bases de données représentent une source potentielle d'information souvent très riche, mais elles n'ont pas été construites en vue d'une étude précise. Elles ne peuvent donc pas prendre en compte les différentes possibilités de biais inhérentes à chaque étude, et n'offrent pas la possibilité d'y remédier.

Exemple : Le rôle des bêta-bloqueurs dans la prévention de la maladie coronarienne chez les patients hypertendus peut être analysé dans le cadre de bases de données déjà constituées rassemblant les prescriptions et la délivrance des médicaments, comme elles existent dans les pays scandinaves. Les bêta-bloqueurs cependant, sont prescrits dans différentes indications, y compris le traitement de l'angine de poitrine, et l'intrication des pathologies (HTA, angor, post-infarctus) chez un même patient rend l'analyse particulièrement difficile.
L'utilisation d'informations rassemblées dans un but différent de celui de l'étude que l'on désire réaliser (but économique, évaluation de la consommation médicamenteuse et de son évolution par exemple) est toujours sujette à caution, et les conclusions de l'étude doivent être interprétées avec autant de prudence [20].

Correction du biais de confusion :
Il faut, pour que l'élément confondant puisse être éliminé lors de l'analyse des données, qu'il ait été prévu de collecter les informations le concernant, lors de la construction de l'étude, de sorte qu'il puisse être contrôlé. Le contrôle consiste à établir des sous-groupes de patients en fonction du facteur confondant suspecté, et à calculer l'association entre la maladie et le facteur de risque étudié à l'intérieur de chaque sous-groupe.

Exemple constitué à partir de données fictives inspirées d'articles parus : le risque d'infection par le virus HIV semble augmenté chez les drogués d'un groupe ethnique A, par rapport à un autre (B)

	HIV +	HIV -	
Groupe A	287	163	450
Groupe B	163	287	450
	450	450	

$$OR = \frac{287 \times 287}{163 \times 163} = 3,1$$

Si l'on admet que les sujets du groupe A utilisaient préférentiellement la cocaïne et se piquaient quatre fois plus souvent, le type de drogue utilisé représente alors un élément confondant qu'il faut contrôler : on construit des sous-groupes sur la base de la drogue utilisée :

	Cocaïne		Héroïne		
	Groupe A	Groupe B	Groupe A	Groupe B	
HIV+	275	138	12	25	450
HIV-	25	12	138	275	450
	300	150	150	300	

Puis on calcule les odds ratio pour chaque sous-groupe :

1- Sous-groupe cocaïne : 2- Sous-groupe Héroïne :

$$OR = \frac{275 \times 12}{25 \times 138} = 0,95 \qquad OR = \frac{12 \times 275}{138 \times 25} = 0,95$$

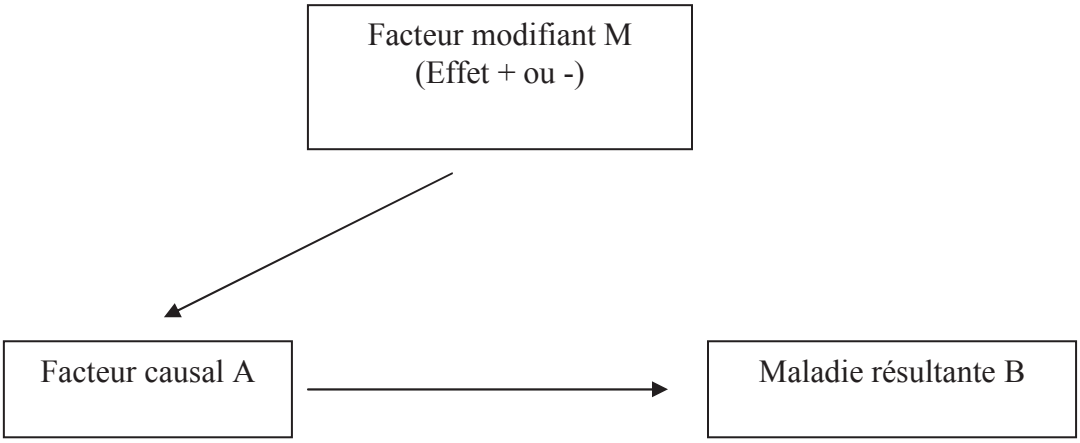
et l'on s'aperçoit que la différence observée initialement disparaît et était à mettre sur le compte, en fait, de la drogue utilisée et du nombre d'injections quotidiennes.

Le contrôle par l'élément confondant lors de l'analyse statistique n'est bien sûr possible que si l'élément confondant potentiel est déjà suspecté lors de la construction de l'étude et si les données sont recueillies en conséquence !

Une autre méthode de contrôle statistique de l'élément confondant peut se montrer plus puissante car elle ne nécessite pas la division de l'échantillon initial en sous-groupes : il s'agit de *l'analyse multivariée en régression logistique*, dans laquelle l'élément confondant potentiel peut être introduit comme variable indépendante. Si la plus grande part du pouvoir explicatif revient à l'élément confondant potentiel au détriment du facteur de risque supposé initial, l'élément confondant sera retenu dans le modèle comme significatif, et le facteur de risque supposé initial en verra sa significativité amoindrie, voire annulée.

D- MODIFICATEURS D'EFFET, OU FACTEURS MODIFIANTS :

Nous avons vu qu'un facteur confondant représente un facteur causal véritable, dont l'ignorance dans une étude fait attribuer à tort la responsabilité causale à un autre facteur mesuré dans l'étude et statistiquement mais non biologiquement associé à la survenue de la maladie.
Il faut bien différencier facteurs confondants et facteurs modifiants : un facteur modifiant est, par définition, un agent qui va modifier la relation existant entre un facteur réellement causal et la maladie qui en résulte. Il y a interaction entre le facteur causal et le facteur modifiant.
Plusieurs cas de figure peuvent se présenter :



1- **Le facteur modifiant M ne joue pas de rôle pathogène s'il est isolé, mais accentue le rôle du facteur causal A lorsque A et M sont présents simultanément.** Par exemple, le virus delta ne peut à lui seul, provoquer une hépatite. En association avec le virus de l'hépatite B, il renforce le rôle pathogène de ce dernier avec aggravation de l'hépatite et évolution plus fréquente vers la chronicité et la cirrhose. *M seul, ne peut pas provoquer la maladie B.*

2- **Le facteur modifiant M ne joue pas de rôle lorsqu'il est isolé, mais atténue le rôle du facteur causal A lorsqu'ils sont associés.** C'est le cas des mécanismes de protection génétiques ou acquis contre une affection : le gène de la drépanocytose ne joue aucun rôle dans l'infection palustre. Cependant, sa présence à l'état hétérozygote diminue la sévérité de la maladie due au Plasmodium par diminution de la capacité de ce dernier à parasiter les globules rouges.

3- **Le facteur modifiant M joue un rôle dans la genèse de la maladie, mais sa présence associée au facteur causal A résulte en une multiplication d'effets et non en l'addition de leurs effets séparés** : M et A sont synergiques. Le tabac et l'hypertension sont deux facteurs de risque bien connus des maladies cardio-vasculaires. Leur association aboutit cependant à une multiplication, et non pas une addition, du risque.

4- **Le facteur modifiant M joue un rôle dans la genèse de la maladie, mais sa présence associée au facteur causal A diminue son pouvoir pathogène** : M et A sont antagonistes. Cette situation, malheureusement assez rare en Médecine, pourrait illustrer la vieille théorie d'Hippocrate du combat du mal par le mal ? Il semblerait ainsi que l'augmentation du cholestérol au niveau du cerveau observée chez des patients alcooliques antagoniserait en fait, l'action de l'alcool sur les connections neuronales [21].

5- **L'effet antagoniste peut parfois s'exercer… au détriment d'un bénéfice final**. Plusieurs études ont montré une diminution du risque de cancer colo-rectal chez les patientes prenant du calcium et de la vitamine D en prévention de l'ostéoporose. Cet effet bénéfique n'a pas été retrouvé dans un essai randomisé, dont un bras comprenait également un traitement par oestrogènes. Il s'avère que dans cet essai, les oestrogènes ont agi comme modificateur d'effet antagoniste sur l'action de calcium et de la vitamine D, pour en annuler l'effet préventif bénéfique [22].

Comment, en pratique épidémiologique, différencier facteur de confusion et facteur modificateur d'effet ?

L'analyse contrôlée par le facteur de confusion (le type de drogue dans l'exemple précédent) permettait d'éliminer la responsabilité d'un facteur présumé causal à tort (le groupe ethnique). L'analyse contrôlée par le facteur modificateur d'effet permettra, de même, d'établir le rôle de chaque facteur causal. Imaginons qu'une étude de cohorte sur les méfaits du tabac en pathologie ORL donne les résultats suivants après un suivi de 10 ans :

	Tabac +	Tabac -
Cancer ORL +	200	60
Cancer ORL -	600	740
	800	800

S'agissant d'une étude de cohorte, nous calculons les incidences cumulées sur 10 ans des cancers ORL dans le groupe de patients tabagiques et le groupe de patients non tabagiques :

Groupe tabac : Incidence = 200/800 = 25 %
Groupe sans tabac : Incidence = 60/800 = 7,5 %

L'alcool étant connu comme facteur de risque des cancers ORL, les coordinateurs de l'étude avaient prévu d'en mesurer la consommation, de pair avec celle du tabac, chez tous les membres de la cohorte. Une analyse contrôlée par le facteur alcool est effectuée, qui donne les résultats suivants:

	Alcool +		Alcool -	
	Tabac +	Tabac -	Tabac +	Tabac -
Cancer ORL +	160	40	40	20
Cancer ORL -	240	360	360	380
Total	400	400	400	400
Incidence	40%	10%	10%	5%

Nous voyons tout d'abord que de contrôler par la variable alcool ne fait disparaître la relation entre tabac et cancer : l'incidence des cancers ORL chez les sujets tabagiques est toujours plus élevée que chez les sujets non tabagiques. L'alcool n'est donc pas un facteur confondant de la relation tabac-cancer, qui existe réellement.

Nous voyons ensuite que le groupe (tabac - alcool -) a un risque de cancer ORL de 5 %, les groupes (Tabac + Alcool -) et (Tabac - Alcool +) un risque équivalent de 10 %, et le groupe (Tabac + Alcool +) un risque non pas de 20 % (somme des risques séparés de l'Alcool et du

Tabac), mais de 40 %. Nous sommes dans la situation 3, où l'alcool est à lui seul un facteur de risque de la maladie (car le groupe sans alcool a un risque plus faible que le groupe avec alcool) et un facteur modifiant le risque toxique du tabac en l'augmentant.

Nous remarquons enfin que l'incidence brute de 25 % dans le groupe (Tabac+) est comprise entre l'incidence de 40 % dans le groupe (Tabac+ Alcool+) et l'incidence de 10 % dans le groupe (Tabac+ Alcool-). De même, l'incidence brute de 7,5 % dans le groupe (Tabac-) est comprise entre l'incidence de 10 % dans le groupe (Tabac- Alcool +) et l'incidence de 5 % du groupe (Tabac- Alcool-). Cette relation, que l'on peut résumer comme suit, est caractéristique de rôle de facteur modifiant de la variable M par laquelle est fait le contrôle dans l'analyse :

Incidence groupe M - < incidence brute < Incidence groupe M +

Lorsque le rôle exact (facteur confondant ou facteur modifiant ?) d'une variable dans la genèse d'une maladie est inconnu, l'analyse contrôlée par cette variable permet d'une part de reconnaître son caractère, et d'autre part, dans le cas d'un facteur modifiant, de quantifier le sens (synergie ou antagonisme ?) et l'ampleur exacte de la modification d'effet d'un facteur causal.

Dans notre exemple, tabagisme ou alcoolisme pris isolément multiplient le risque de cancer ORL par 2 (10 % versus 5 %), et, associés, par 8 (40 % versus 5 %).

E- COMMENT INTERPRETER UNE ETUDE EPIDEMIOLOGIQUE ?

Les biais structurels des études cliniques que nous venons de détailler expliquent sans doute nombre de résultats divergents observés dans la littérature. D'autres facteurs cependant ne doivent pas être oubliés :

E1- La qualité de l'analyse statistique :

Le type d'analyse statistique, les tests employés, l'application aux questions posées doivent être définis dès la construction de l'étude. Cela permet d'une part de recueillir les données sous la forme adéquate à l'analyse, évitant ainsi les manipulations inopportunes, et d'autre part de tester de façon rigoureuse les hypothèses initiales en évitant la multiplication des calculs effectués au hasard des données. En d'autres termes, la multiplication des comparaisons, des calculs de test d'association expose à découvrir et à décrire comme statistiquement significatives des différences ou des associations pour lesquelles le coefficient de significativité n'est significatif que par hasard.

Exemples :

- *Certains logiciels de statistique permettent d'établir des matrices de corrélations* entre toutes, ou une partie, des données collectées. Il s'agit de corrélations statistiques, ne correspondant pas obligatoirement à une réalité biologique ou épidémiologique. Si l'on accepte un coefficient d'erreur alpha de 5 % et si l'on effectue au hasard 100 tests de corrélation, 5 de ces tests peuvent s'avérer faussement significatifs car correspondant aux 5 % d'erreur admise. Le risque est grand ensuite, si l'on ne publie que les 5 résultats revenus positifs sans indiquer qu'ils ont été observés du fait du hasard dans un lot de 100 tests, de ne porter à la connaissance des lecteurs qu'un artefact statistique. Il s'agit de la publication de résultats ne répondant pas à une hypothèse de recherche pré-établie.

- Tous les logiciels rendent l'analyse statistique très facile et rapide, par rapport au temps nécessaire aux calculs effectués manuellement… Et donc, ils exposent tous au risque de *tests multiples sans rationnel pré-établi*, et au risque de valeurs de p significatives juste… par hasard.

L'erreur alpha communément admise à 5 % relève d'un choix arbitraire. Le risque au terme d'une analyse statistique bien conduite d'obtenir un résultat positif par chance ou hasard subsiste toujours, et il suffit parfois de l'adjonction de quelques cas supplémentaires à une série pour qu'alpha passe à 6 ou 7 % et que les résultats soient tenus comme non significatifs... Abaisser alpha augmente la fiabilité des résultats.

E2- Les études non publiées :

Une étude à résultats négatifs a généralement moins de chance d'être publiée qu'une étude à résultats positifs (les grandes études multicentriques mises à part). Il existe donc dans la littérature un **biais vers les résultats positifs**, particulièrement important sur le plan pratique pour les études de thérapeutique alors qu'il serait important de savoir que l'efficacité d'un traitement n'est pas retrouvée de façon régulière dans différents centres investigateurs. Le **biais de publication** est un biais invisible s'il n'existe pas de registre de l'ensemble des études réalisées, publiées ou non. Lorsqu'un tel registre existe, comparer les résultats publiés aux résultats effectivement obtenus, mais parfois non publiés, met en évidence une différence significative et une influence manifeste du biais de non-publication [23].

E3- Toute étude doit être interprétée dans le contexte plus large des connaissances déjà acquises :

B. A. Hill proposait en 1965, **9 critères d'interprétation** d'une étude épidémiologique [24] :

1- **Force de l'association observée** : une association a d'autant plus de chances d'être réelle que son amplitude mesurée est plus importante. Des risques relatifs inférieurs à 2 sont faibles.

2- **Reproductibilité** : un résultat sera d'autant plus fiable qu'il est retrouvé, comme en biologie, par différents investigateurs.

3- **Spécificité** : une relation de causalité spécifique entre un facteur de risque et une maladie donnée est un argument supplémentaire pour admettre la responsabilité du facteur. Exemple : BK et tuberculose, alors que l'on hésitera à reconnaître comme pathogéniques d'une lésion pulmonaire des germes retrouvés dans les crachats... Un facteur de risque peut cependant intervenir dans différentes maladies.

4- **Temporalité** : établir que la cause supposée précède l'effet constaté représente un argument de poids !

5- **Gradient biologique** : l'augmentation ou la diminution de l'incidence d'une maladie sous l'influence de l'augmentation ou la diminution de la fréquence ou de l'intensité de l'exposition, rend la responsabilité pathogénique de l'exposition plus probable.

6- **Plausibilité** : sa définition souffre de toute évidence de nombreux biais, mais le bon sens reste utile même dans les démarches scientifiques les plus avancées !

7- **Cohérence** : des résultats seront plus vraisemblables s'ils s'intègrent dans le cadre des connaissances acquises comme solides sans trop les bouleverser... Il y a bien sûr des exceptions historiques notables.

8- **Expérimentation** : elle n'est pas toujours possible en Epidémiologie, mais peut venir conforter des données observées sur le terrain.

9- **Analogie** : ce critère a lui aussi ses limites, mais la ressemblance d'un phénomène observé avec d'autres phénomènes déjà connus peut accroître sa crédibilité.

E4- Application des résultats d'une étude à un individu :

Les études de groupes sont nécessaires afin de mieux cerner la réalité d'ensemble d'une pathologie ou d'un traitement, impossible à appréhender à partir d'un seul ou de quelques patients. Elles permettent d'observer au niveau du groupe des tendances générales, et l'on appliquera les conclusions au niveau de groupes avec des résultats raisonnablement prévisibles *car escomptés sur les données recueillies au niveau d'un groupe* :

Exemple : La vaccination anti-variolique a permis d'éradiquer la maladie de la surface du globe, donc de réduire l'incidence à 0/an/million d'habitants.
Corollaire: les individus semblent actuellement protégés de façon définitive contre la maladie au prix cependant d'un certain nombre d'encéphalites vaccinales graves, survenues chez des patients qui peut-être n'auraient jamais été atteints de la maladie : *le bénéfice individuel peut être dissocié du bénéfice de groupe*.

En pratique, le clinicien est souvent confronté à des situations où le poids respectif du bénéfice et du risque pour un individu précis est beaucoup plus difficile à apprécier, malgré la réalisation d'études extensives :

Exemple : La plupart des études de la thrombolyse à la phase aiguë de l'infarctus du myocarde ont montré une réduction de la mortalité hospitalière chez les patients traités. Il s'agit donc de résultats confirmés, et d'autant plus fiables que l'amplitude de la réduction est semblable d'une étude à l'autre (environ 25 %). Dans l'étude GISSI, la mortalité globale passe de 13 % dans le groupe non traité, à 10,7 % dans le groupe traité (réduction de 2,3 % en valeur absolue, de 18 % = [(10,3 - 13)/13] x 100 en valeur relative) [25].
Les accidents neurologiques sous thrombolyse surviennent cependant de façon non exceptionnelle dans 0,94 à 1,33 % des cas [26]. Traiter 1000 patients par thrombolyse à la phase aiguë revient à éviter 23 décès durant la période d'hospitalisation, mais à provoquer 10 à 13 accidents neurologiques présumés hémorragiques. Le bénéfice est supérieur au risque et le traitement donc licite, mais le clinicien face à son patient n'a que très peu de moyens de savoir, même en respectant les contre-indications des thrombolytiques, si :

1- son malade fait partie des 23 patients courant un risque mortel en cas de non-administration du traitement;
2- son malade ne fait pas partie des 13 patients courant un risque d'hémorragie cérébrale sous traitement;
3- son malade ne fait pas partie des 977 patients pour lequel le bénéfice du traitement en termes de survie est nul.

L'épidémiologiste peut chercher à mieux définir les facteurs pronostiques, les indications du traitement, les caractéristiques des patients non améliorés. Cependant, son outil méthodologique n'est pas adapté à l'étude particulière de l'individu et il ne peut fournir qu'une orientation générale (précieuse) à moduler par le clinicien en fonction du cas présent. Les arbres de décision cherchent à combler cette lacune en intégrant les particularités propres à chaque patient, mais ils restent d'application difficile et ne peuvent pas être employés de façon routinière.

Conclusion

Les techniques épidémiologiques appliquées à la recherche épidémiologique pure ou à la recherche clinique ont donc, comme toute technique, leurs limitations tant dans leur champ exploratoire que dans leurs possibilités d'interprétation. Elles n'en demeurent pas moins indispensables et inégalées jusqu'à présent chaque fois que l'on veut mesurer un phénomène affectant le vivant dans son environnement normal, avec l'ensemble des interactions le caractérisant, en-dehors des conditions artificielles d'isolement d'un laboratoire. Elles restent le seul outil capable de vérifier sur le terrain, les hypothèses engendrées *in vitro* ou par l'expérimentation. Elles permettent enfin d'attribuer à chaque facteur le poids qui lui revient dans la réalité dans son action pathogénique, physiopathologique ou thérapeutique... mais elles doivent, du fait de la complexité des phénomènes qu'elles survolent et tentent de décrire de façon globale, être maniées avec la connaissance de leurs défauts. Elles sont à la médecine, ce que le solfège est à la musique : nécessaire, mais insuffisant pour interpréter la partition et l'adapter au moment présent.

Références

1 Greenfield ML, Mhyre JM, Mashour GA, Blum JM, Yen EC, Rosenberg AL. Improvement in the quality of randomized controlled trials among general anesthesiology journals 2000 to 2006: a 6-year follow-up. Anesth Analg 2009;108:1916-1921.
2 Brenner H, Savitz DA. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. AM J Epid 1990;132:181-192.
3 Holt VL, Daling JR, Stergachis A, Voigt LF, Weiss NS. Results and effect of refusal recontact in a case-control study of ectopic pregnancy. Epidemiology 1991;2:375-379.
4 Doll R, Peto R, Hall E, Wheatley K, Gray R. Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. BMJ 1994; 309: 911–918.
5 Austin PC, Mamdani MM, van Walraven C, Tu JV. Quantifying the impact of survivor treatment bias in observational studies. J Eval Clin Pract 2006;12:601-612.
6 Di Martino LD, Hammill BG, Curtis LH, Gottdiener JS, Manolio TA, Powe NR, Schulman KA. External validity of the cardiovascular health study: a comparison with the Medicare population. Med Care 2009;47:916-923.
7 Ioannidis JP, Polycarpou A, Ntais C, Pavlidis N. Randomised trials comparing chemotherapy regimens for advanced non-small cell lung cancer: biases and evolution over time. Eur J Cancer 2003;39:2278-2287.
8 Robins J, Pike M. The validity of case-control studies with non-random selection of controls. Epidemiology 1990;1:273-284.
9 Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. NEJM 1978; 299:1089-1094.
10 Hulka B, Grimson RC, Greenberg BG, Kaufman DG, Fowler WC, Hogue CJR, Berger GS, Pulliam CC. "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen. Am J Epidemiol 1980;112:376-387.

11 Sagawa M, Usuda K, Aikawa H, Machida Y, Tanaka M, Ueno M, Sakuma T. Lung cancer screening and its efficacy. Gen Thorac Cardiovasc Surg 2009;57:519-27.
12 Elliott CS, Shinghal R, Presti JC Jr. The influence of prostate volume on prostate-specific antigen performance: implications for the prostate cancer prevention trial outcomes. Clin Cancer Res 2009;15:4694-4699.
13 Van Gossum A, Munoz-Navas M, Fernandez-Urien I, Carretero C, Gay G, Delvaux M, et al. Capsule endoscopy versus colonoscopy for the detection of polyps and cancer. NEJM 2009;361:264-270.
14 Wong NA, Hunt LP, Novelli MR, Shepherd NA, Warren BF. Observer agreement in the diagnosis of serrated polyps of the large bowel. Histopathology 2009;55:63-66.
15 Lindefors-Harris BM, Eklund G, Adami HO, Meirik O. Response bias in a case-control study : analysis utilizing comparative data concerning legal abortions from two independent Swedish studies. Am J Epid 1991;134:1003-1008.
16 Drews CD, Greeland S. The impact of diffential recal on the results of case-control studies. Int J Epid 1990;19:1107-1112.
17 Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value ? Am J Epid 1990;132:746-748.
18 Schoenbaum EE, Hartel D, Selwyn PA, Klein RS, Davenny K, Rogers M, Feiner C, Friedland G. Risk factors for human immunodeficiency in intraveinous drug users. NEJM 1989; 321:874-879.
19 Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad-hoc methods of correcting for missing values for confounding variables. Am J Epid 1991;134:895-907.
20 Psaty BM, Koepsell TD, Siscovick D, Wahl P, Logerfo JP, Inui TS, Wagner EH. An approach to several problems in using large databases for population-based case-control studies of the therapeutic efficacy and safety of anti-hypertensive medicines. Statistics in Medicine 1991, 10, 653-662.
21 Crowley JJ, Treistman SN, Dopico AM. Cholesterol antagonizes ethanol potentiation of human brain BKCa channels reconstituted into phospholipid bilayers. Mol Pharmacol 2003;64:365-372.
22 Ding EL, Mehta S, Fawzi WW, Giovannucci EL. Interaction of estrogen therapy with calcium and vitamin D supplementation on colorectal cancer risk: reanalysis of Women's Health Initiative randomized trial. Int J Cancer 2008;122:1690-1604.
23 Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009;302:977-984.
24 Hill BA. The environment and disease : association or causation ? Proc Royal Soc Med 1965;58:295-300.
25 Gruppo Italians per studio della Streptochinasi nell' infarto miocardio (GISSI): Effectiveness of intraveinous thrombolytic treatment in acte myocardial infarction. Lancet 1986;1:397-402.
26 Gruppo Italiano per lo Studio della Sopravinenza nell' Infarto miocardico : GISSI 2 : a factorial randomized trial of alteplase versus streptokinase and heparin versus non heparin among 12 490 patients with acute myocardial infarction. Lancet 1990;336:65-71.

Troisième partie :
L'Elaboration du Protocole

TROISIEME PARTIE: L'ELABORATION DU PROTOCOLE

Lorsqu'il s'engage dans un projet de recherche, le chercheur doit développer un protocole, qui est un plan ou un guide décrivant son étude.

A travers la description des différents types d'étude qu'il a à sa disposition, on a vu l'importance pour le chercheur de trouver la meilleure adéquation entre l'objectif de sa recherche et le type d'étude choisi, mais aussi le meilleur compromis entre un niveau de preuve optimal lié à un plan d'étude très complexe et la faisabilité du projet.

Les données d'une étude en constituent le point central. C'est à partir d'elles que l'on évalue la taille d'un effet, une différence entre deux groupes, ou plus généralement que l'on apporte des informations utiles à la communauté scientifique.

Si l'analyse des données est capitale, leur mode de recueil l'est tout autant. Celui-ci fait intervenir des mesures, et donc des instruments de mesure, qui doivent être décrits et évalués dans le cadre du protocole. Le questionnaire constitue un type particulier de données.

Un protocole doit conduire à une description exacte du phénomène étudié. Toute recherche clinique comporte le risque de s'éloigner de la vérité, en surestimant, sous-estimant, affirmant à tort ou ignorant à tort un effet. Des contradictions majeures peuvent exister entre les conclusions de différentes études sur un même sujet.

Il est donc nécessaire, au stade de la conception du protocole, d'identifier et de prendre en compte toutes les sources d'erreurs potentielles, biais et difficultés d'interprétation de l'étude.

Il faut de plus se donner les moyens de mettre en évidence un effet s'il existe. C'est la contrainte fondamentale de la puissance de l'étude, liée à la taille de l'échantillon étudié, et dépendant de l'hypothèse que le chercheur a formulée.

Enfin, le chercheur ne vit pas que "d'amour et d'eau fraîche". Son étude ne verra le jour que s'il a prévu et résolu par anticipation les contraintes matérielles.

CHAPITRE XIII

LA REDACTION DU PROTOCOLE

Gilles Landrivon, François Delahaye

Le protocole est le document qui décrit la méthode de l'étude que l'on se propose de réaliser, de la justification aux objectifs, de l'hypothèse aux contraintes méthodologiques. Il définit ses conditions de réalisation et son déroulement.

Il est essentiel de rédiger le protocole à un stade précoce du processus de conception de l'étude, aussi bien pour l'investigateur et ceux qui participent à l'étude, que pour les comités d'éthique ou les commanditaires. Son caractère innovant, sa pertinence, l'importance de la question posée, ainsi que l'impact attendu du résultat de l'étude sur la population ciblée faciliteront l'obtention d'un financement adapté au budget requis.

Dans ce chapitre, il est proposé un plan de protocole dont chaque étape est décrite. A travers la rédaction du protocole, on retrouve toutes les notions développées dans les différents chapitres du livre.

Plan du chapitre

XIV - Les implications éthiques

XV - Le personnel

XVI - Le budget

XVII - Le calendrier

XVIII - Les annexes éventuelles

XIX - Les références

Le plan proposé dans ce chapitre est un plan général qui peut être utilisé dans toutes les situations que l'on peut rencontrer en recherche clinique. Pour une étude de conception plus simple qu'une étude analytique, toutes les étapes qui sont décrites ici ne sont pas forcément utiles ou pertinentes.

Chacune de ces dix-neuf étapes ne se présente pas dans un ordre figé. Si pour les cinq premières et les neuf dernières l'ordre proposé est relativement logique, il en va différemment de la sixième à la dixième étape où on a beaucoup plus de liberté pour envisager et ordonner ces informations selon la séquence qui paraît la plus naturelle dans le contexte où l'on se trouve.

Ces étapes sont les suivantes:

1 - Le titre

2 - Le(s) objectif(s)

3 - La justification de l'étude

4 - La (les) hypothèse(s)

5 - Le type d'étude

6 - Le(s) facteur(s) étudié(s)

7 - Le(s) critère(s) de jugement

8 - Les causes d'erreur: biais et facteurs de confusion

9 - Les sujets

10 - L'analyse des données

11 - La récolte et la gestion des données

12 - L'analyse des données

13 - Une éventuelle étude pilote

14 - Les implications éthiques

15 - Le personnel

16 - Le budget

17 - Le calendrier

18 - Les annexes éventuelles

19 - Les références

I - Le titre

Il résume le problème qu'on se propose d'étudier. Il doit être clair, précis, suffisamment court et en même temps suffisamment informatif.

Il peut contenir une information sur le type d'étude proposé.

Après le titre doivent être mentionnés le nom du ou des investigateurs et celui de leur institution, ainsi que la date de mise au point et le numéro de la version du protocole.

II - Le(s) objectif(s)

Il existe quatre grands types d'objectifs, qui doivent être formulés très clairement et très précisément:

- Le premier concerne le pronostic, ou l'évolution d'un état pathologique.

L'objectif est de connaître et de comprendre les événements qui vont survenir chez un patient entre le moment où la maladie s'est déclarée et le moment où l'histoire clinique se termine (par la guérison, la mort ou l'installation du patient dans un autre état physique, mental ou social).

- Le deuxième concerne l'étiologie, ou la causalité.

La compréhension d'une relation de causalité est fondamentale pour le clinicien qui sera guidé dans son travail de diagnostic, de traitement ou de prévention.

L'objectif est de mettre en évidence une relation de causalité entre deux événements ou de calculer la force de l'association entre deux facteurs.

- Le troisième concerne la performance des tests diagnostiques.

L'objectif est d'évaluer une stratégie diagnostique, ou d'améliorer l'interprétation des résultats d'un test.

- Le quatrième concerne l'impact d'une intervention.

L'objectif est d'évaluer une intervention de thérapeutique, de dépistage, de prévention ou d'éducation: fait-elle plus de bien que de mal, quel est son rapport coût-utilité ?

S'il peut y avoir des objectifs secondaires, il ne doit y avoir qu'un seul objectif principal. Celui-ci doit être formulé en termes clairs, et les résultats de l'étude s'y réfèreront.

III - La justification de l'étude

La recherche coûte cher et prend du temps. Elle peut en outre mettre des patients dans des situations d'inconfort ou de risque. Il faut donc mettre l'accent sur les retombées directes de ses résultats.

Le processus de justification commence par une recherche bibliographique très complète sur le sujet, afin d'identifier les faiblesses de l'information disponible sur la question qu'on se propose d'étudier.

Il convient de définir l'importance du problème - incidence, prévalence, morbidité, mortalité, … - ainsi que son cadre, dans le temps, en termes de répartition géographique et de population.

Le coeur de la justification est constitué par:

- la description de l'inadéquation entre ce qui est observé et ce qui devrait être,

- l'explication (ou la tentative d'explication) de ce décalage,

- une proposition de solution pour supprimer ce décalage.

Il faut insister sur:

- le caractère actuel de la question posée, et le caractère nouveau, inédit, de l'information qui sera fournie;

- le niveau de morbidité ou de gravité de l'événement considéré;

- l'importance de la population affectée par le problème (sous-groupe très spécifique *versus* très large population);

- la possible insertion du projet de recherche dans des programmes actuellement en cours, ou ses répercussions sur ce programme;

- le caractère multidisciplinaire du projet: médical, paramédical, économique, psycho-social ou politique.

En somme, un protocole de recherche est bon, et justifié, lorsqu'il est nouveau, pertinent, faisable, et éthique.

IV - La (les) hypothèse(s)

Tout protocole de recherche, s'il s'agit d'une étude analytique, doit explicitement formuler une hypothèse. Il n'y a pas d'hypothèse dans une étude descriptive, qui vise essentiellement à décrire la distribution des caractéristiques d'une population.

Une hypothèse est une affirmation (et non pas une question), sur une possible relation entre le ou les facteurs étudiés et le ou les critères de jugement. Cette affirmation doit découler logiquement de l'argumentation contenue dans la justification.

En général l'hypothèse est proposée sous la forme de l'hypothèse nulle: "il n'y a pas d'association entre le facteur étudié et le critère de jugement" afin que le test statistique, construit à partir des données collectées, permette de calculer la probabilité pour que l'association observée survienne par chance seule.

La proposition: "il y a une association entre le facteur étudié et le critère de jugement" constitue l'hypothèse alternative.

La formulation de l'hypothèse doit s'accompagner de la description des conditions dans lesquelles l'hypothèse est supposée être vraie.

Il peut y avoir plusieurs hypothèses à tester dans une même étude, mais le plus petit nombre est le mieux.

V - Le type d'étude

Le chercheur doit opter pour un type d'étude. Le choix du type d'étude le plus approprié dépend des objectifs et de la question posée, mais aussi des ressources.

Il existe deux grandes catégories d'étude:

- études descriptives, essentiellement études transversales ou études de prévalence;

- études analytiques: les études d'observation, avec les études cas-témoins et les études de cohorte, et les études expérimentales, avec les essais libres et les essais comparatifs.

Certains types d'étude conviennent mieux à certaines questions (tableau 1). Le niveau de preuve des conclusions de l'étude est d'autant meilleur que le type d'étude est mieux adapté à la question posée.

VI - Le(s) facteur(s) étudié(s)

Le facteur étudié se définit comme un événement, un état, une exposition ou une intervention susceptible, d'après l'hypothèse, d'être associé à un problème de santé, une maladie ou tout autre critère de jugement digne d'intérêt.

Il s'agit donc d'une variable à mesurer. Il peut y en avoir plusieurs pour une même étude. Le chercheur doit en fournir la liste dans le protocole. Il peut s'agir:

- de détails descriptifs d'un individu;

- de données, obtenues par questionnaire, sur les antécédents, les habitudes ou les symptômes d'un individu;

- de données obtenues par mesure;

- de données descriptives sur une exposition, une intervention diagnostique, thérapeutique, de prévention, de dépistage, d'éducation ou d'information.

Des précisions doivent être apportées sur la façon de mesurer ces variables:

- questionnaire;

- s'il s'agit d'une mesure instrumentale, description de l'instrument, et des conditions dans lesquelles sont effectuées les mesures (à jeun, assis, après repos, …).

Ces modes de mesure dont dépendent la qualité des données et la validité des résultats doivent également tenir compte du coût financier et humain.

VII - Le(s) critère(s) de jugement

Le critère de jugement, ou facteur résultant, se définit comme la situation ou l'événement supposé être le résultat de l'influence du facteur étudié.

Ces événements dignes d'intérêt pour le patient comme pour le clinicien ou l'épidémiologiste se répartissent en cinq catégories:

- la mort,

- la maladie,

- le handicap,

- l'inconfort,

- l'insatisfaction,

auxquels on peut ajouter un sixième élément, la destitution, qui a une dimension plus sociale. Il peut aussi s'agir de leur inverse: survie, guérison, absence de handicap, …

La définition de ces critères de jugement doit être aussi précise que possible.

Un critère de jugement est lui aussi une variable. C'est la variable dite dépendante, par rapport au facteur étudié qui est la variable indépendante.

Le même soin doit être apporté à la méthode de mesure que pour les facteurs étudiés.

VIII - Les causes d'erreur: biais et facteurs de confusion

Un biais est une erreur systématique qui contribue à produire des estimations systématiquement plus élevées ou plus basses que la valeur réelle des paramètres à estimer. Il intervient par exemple au niveau de la sélection des patients qui participeront à l'étude, ou sur la mesure des paramètres à étudier. Tous les biais potentiels sont bien entendu à identifier, anticiper et éviter lors de la conception de l'étude.

Un facteur de confusion est un facteur qui modifie les effets du facteur étudié sur le critère de jugement, du fait de son lien à la fois avec le facteur étudié et le critère de jugement. Au stade de la rédaction du protocole, le chercheur doit établir la liste de toutes les variables connues susceptibles de se comporter comme des facteurs de confusion, et choisir un type d'étude, ainsi qu'une stratégie d'analyse, pour contrôler leur éventuelle influence. (C'est la contrainte des études d'observation. Celle-ci n'existe pas dans les essais randomisés qui offrent la possibilité de contrôler ces facteurs de confusion, même non mesurés, ou inconnus, du fait de la randomisation.)

IX - Les sujets

La population étudiée et les sujets impliqués dans l'étude doivent être décrits dans le protocole, en prenant en compte plusieurs points:

- est-on sûr de pouvoir généraliser les résultats à partir de la population étudiée ? En d'autres termes, les sujets étudiés sont-ils représentatifs de la population à laquelle on veut rapporter les résultats ?

- y aura-t-il suffisamment de sujets à étudier ?

- le taux de réponse des sujets sollicités pour l'étude sera-t-il correct ?

- la population est-elle stable (caractéristique particulièrement importante pour une étude longitudinale) ?

- si l'étude nécessite des sujets témoins, leur mode de sélection garantit-il qu'ils soient semblables aux cas ?

Ainsi le protocole doit présenter:

- les critères d'inclusion, ou critères d'éligibilité, qui définissent les principales caractéristiques de la population impliquée dans l'étude;

- les critères d'exclusion, qui définissent un sous-groupe de sujets qui ne satisfont pas aux critères d'inclusion, ou qui pourraient y satisfaire, mais qui présentent certaines caractéristiques qui pourraient interférer avec la qualité des données ou l'interprétation des résultats (forte probabilité d'être perdu de vue, inaptitude à fournir des données correctes, …).

Les populations souvent utilisées sont:

- des populations de patients hospitalisés;

- des groupes professionnels;

- des catégories particulières de fonctionnaires (militaires, …);

- des clientèles de médecins généralistes;

- ou de véritables échantillons de la population générale.

X - La taille de l'échantillon

L'échantillon doit être représentatif de la population à laquelle vont s'appliquer les conclusions de l'étude. La technique d'échantillonnage doit être décrite au stade du protocole.

La détermination de la taille de l'échantillon est un élément essentiel du protocole. D'elle dépend en partie la validité des résultats. Elle a une influence fondamentale sur le budget. Elle dépend:

- de l'amplitude de l'effet qu'on espère mettre en évidence

- de la puissance du test utilisé pour détecter une telle différence, si elle existe

- du seuil de significativité choisi pour mettre en évidence une telle différence

- des ressources budgétaires disponibles

XI - La récolte et la gestion des données

Des précisions ont déjà été données sur les informations à collecter concernant le(s) facteur(s) étudié(s) et le(s) critère(s) de jugement. Des informations supplémentaires doivent aussi être fournies sur:

- le mode de collecte et le calendrier;

- les instruments de mesure utilisés;

- les techniques de laboratoire;

- les méthodes de travail sur le terrain;

- le contrôle de qualité;

- le matériel informatique utilisé ainsi que la technique: type d'ordinateur et de logiciel, mode de codage et de saisie.

XII - L'analyse des données

Le protocole doit comporter une description des analyses statistiques programmées dans l'étude et leur justification (pour la prise en compte, par exemple, des facteurs de confusion).

Le choix du matériel informatique est lié à celui des analyses statistiques.

Les analyses statistiques sont de deux ordres:

- statistiques descriptives, qui analysent la distribution des variables;

- statistiques analytiques, pour tester la ou les hypothèses originales et fournir des informations sur le caractère significatif de l'association mise en évidence entre les variables étudiées.

XIII - Une éventuelle étude pilote

Elle n'est pas toujours nécessaire.

Elle doit être réalisée sur un échantillon représentatif. Elle est utile pour:

- entraîner et tester le personnel impliqué dans l'étude;

- évaluer l'acceptabilité des procédures;

- évaluer le taux de réponse;

- estimer l'amplitude de la différence à observer (ce qui est utile pour la détermination de la taille de l'échantillon).

L'étude pilote doit donc être prévue assez tôt pour permettre d'adapter des modifications au déroulement de l'étude si cela est nécessaire. Il faut estimer son propre coût et sa durée.

XIV - Les implications éthiques

Au stade de la réalisation du protocole, les considérations éthiques mettent l'accent sur le respect de la personne, à travers les problèmes de confidentialité et le principe de la minimisation du risque.

Les mesures de sécurité concernant les données doivent être décrites: anonymat, accès limité, destruction des données après la fin de l'étude, identification des sujets impossible lors de la publication des résultats.

Les mesures prises pour garantir l'absence d'exposition du patient à un risque supérieur à un risque encouru en routine sont également précisées.

L'obtention d'un consentement éclairé de la part d'un patient impliqué dans l'étude doit être également garantie.

Enfin, une liste des noms et fonctions des membres de l'équipe de recherche doit être fournie avec le protocole.

XV - Le personnel

Qui fait quoi et quand ? Tout le personnel participant à l'étude doit être identifié (cette étape est importante pour l'estimation du budget):

- investigateur principal
- coordinateur
- promoteur
- assistant de recherche clinique
- technicien (de laboratoire, électroradiologiste, …)
- responsable des interrogatoires
- secrétaire
- opérateur de saisie
- programmateur informaticien
- statisticien

Pour certains, notamment ceux qui sont responsables des questionnaires, des informations doivent être fournies sur leur formation et leur compréhension des objectifs spécifiques de l'étude.

XVI - Le budget

Ce paragraphe réalise la synthèse de toutes les étapes de l'étude du point de vue du coût. La qualité de sa présentation, sa pertinence et sa justification sont essentielles pour l'acceptation du projet par un organisme de financement.

Il peut être présenté en deux rubriques:

- le matériel
- la gestion et le personnel

XVII - Le calendrier

Il doit préciser la durée et le moment de chaque étape.

XVIII - Les annexes éventuelles

Elles réunissent un certain nombre de documents:

- la lettre d'information au patient
- le formulaire de consentement éclairé
- les documents d'approbation des différents décideurs ou institutions impliqués dans l'étude (hôpitaux, ministères, employeurs, syndicats, …)
- le manuel d'opération (questionnaires, feuilles de saisie, lettre de relance, …)
- description des méthodes de mesure
- l'approbation du comité d'éthique, lorsque celle-ci est acquise

XIX - Les références

Dans cette section sont données les références des divers documents cités dans les différentes parties du protocole.

Tableau 1 - Type d'étude le plus adapté selon la nature de la question posée

Nature de la question	Type d'étude le mieux adapté
Prévalence	Etude transversale
Incidence	Etude de cohorte
Risque	Etude de cohorte, étude cas-témoins
Pronostic	Etude de cohorte
Etiologie, causalité	Etude de cohorte, étude cas-témoins
Intervention	Essai
Diagnostic	Etude transversale, essai

CHAPITRE XIV

LES OUTILS DE MESURE
Instruments et Questionnaires

Anne-Marie Schott, François Chapuis, Gilles Landrivon, Eric-Nicolas Bory, François Delahaye

La formulation de la question de recherche nécessite une bonne définition des facteurs étudiés et des critères de jugement. Ces derniers vont définir, dans le cadre du protocole, les données qui seront collectées et leur mode de recueil. La collecte des données fait intervenir des mesures et des questionnaires.

L'évaluation des instruments de mesure est essentielle à un stade précoce de la recherche. En effet, les défauts d'un instrument non seulement ont des conséquences sur la puissance de l'étude et la taille de l'échantillon nécessaire, mais ils peuvent compromettre aussi la validité même des conclusions de l'étude en introduisant des erreurs irréversibles dans l'estimation des variables (facteurs étudiés, critères de jugement, facteurs de confusion).

Le questionnaire constitue un mode particulier de collecte des données. Initialement utilisées en sociologie, psychologie et psychiatrie, les technologies de questionnaires intéressent aujourd'hui l'ensemble des spécialités médicales.

De la qualité d'un questionnaire dépend la qualité de la transmission, du stockage et de l'interprétation des renseignements recueillis. A travers le questionnaire, l'investigateur effectue une mesure du phénomène étudié. Un mauvais questionnaire, ou des questions mal posées, peuvent induire des biais de mesure : il peut y avoir alors une sur- ou une sous-estimation de l'importance du phénomène observé, invalidant totalement les conclusions de l'étude.

Plan du chapitre

PREMIERE PARTIE : INSTRUMENTS DE MESURE

Toute mesure est entachée d'un certain degré d'erreur. Il est fondamental de connaître les différents types d'erreurs, leurs origines, la façon de les quantifier et de les réduire.

Il n'y a pas d'instrument de mesure universellement bon. Le choix d'un instrument, parmi ceux qui existent et qui ont déjà été utilisés, dépend de plusieurs éléments dont les trois suivants qu'il faut prendre en compte : **l'objectif de la recherche, la maladie étudiée et la population cible.**

I - CHOIX D'UN INSTRUMENT DE MESURE

A - Définition de l'objectif

1 - Utilisation transversale dans un but diagnostique ou pronostique

On peut vouloir mesurer un symptôme ou une caractéristique permettant de classer une population en plusieurs groupes de sujets (dépistage). Il s'agit de déterminer la différence entre les sujets, pour une variable donnée, de façon transversale.

Par exemple:

- pour connaître la prévalence d'une affection dans une population déterminée ou pour dépister une affection dans une population déterminée, en épidémiologie;

- pour décider au niveau individuel d'une intervention thérapeutique ou pour déterminer si un individu est éligible dans un essai, en clinique.

Ainsi, on utilise le frottis cervical pour le dépistage du cancer du col utérin et pour en évaluer la prévalence dans une population déterminée.

Certains examens sont destinés à évaluer le risque du développement d'une affection chez des sujets sains ou de l'aggravation d'une maladie chez des patients. Il s'agit dans ces cas de mesurer des facteurs de risque et de pronostic. Par exemple, on recherche plusieurs types d'anticorps chez des patients atteints d'hépatite B. La présence et le taux de ces anticorps permettent d'apprécier le potentiel évolutif de la maladie et ainsi d'en établir le pronostic.

2 - Utilisation longitudinale pour évaluer l'efficacité d'une intervention

Il s'agit de déterminer l'évolution d'un sujet de façon longitudinale en comparant deux états d'un même sujet. Le plus souvent l'évaluation se fait avant et après une intervention thérapeutique dont on désire évaluer l'efficacité.

B - Recherche des instruments existants

Cette étape est obligatoire pour tous les types de mesure.

Paradoxalement, lorsqu'on s'intéresse à un objectif précis pour une population déterminée, il devient très difficile de trouver l'instrument adéquat. On est alors tenté d'être très critique vis-à-vis des instruments existants et de sous-estimer la difficulté du développement d'un nouvel instrument. Une erreur fréquente des cliniciens est d'éliminer trop facilement les échelles existantes et de se lancer dans le développement d'un nouvel instrument avec l'idée optimiste qu'ils peuvent mieux faire. En fait, le développement d'un instrument de mesure nécessite un investissement considérable de temps et d'argent. L'étape de revue exhaustive de tous les instruments développés dans le domaine est donc obligatoire.

A l'inverse, l'autre erreur fréquente est de considérer qu'une échelle est bonne parce qu'elle a été "validée", sans tenir compte des objectifs et des conditions dans lesquelles la validation a été faite. Une échelle validée pour déterminer le niveau des fonctions cognitives chez des personnes âgées n'est certainement pas valide pour mesurer l'état cognitif chez des étudiants de 20 ans.

Une fois cette étape de recherche dans la littérature et auprès des experts terminée, il faut alors choisir parmi les instruments existants celui qui semble le plus adapté à l'objectif qu'on s'est fixé. Il faut pour cela déterminer leur pertinence par rapport à ce qu'on veut mesurer. Il faut également rechercher si l'exactitude et la reproductibilité de ces instruments ont été évaluées de façon correcte, dans un contexte identique et pour des objectifs similaires.

C – Définir les phénomènes à mesurer

Le phénomène que l'on veut mesurer peut être exprimé selon divers types d'échelle:

- Données qualitatives: elles sont exprimées en catégories qui ne peuvent être ordonnées les unes par rapport aux autres (couleur des yeux, catégories professionnelles…). Lorsqu'il n'existe que deux catégories possibles, les données sont dites dichotomiques ou binaires (statut vital, sexe).

- Données ordinales: à l'inverse des précédentes, ces données peuvent être classées par ordre croissant ou décroissant et combinées comme des nombres (par exemple, multipliées ou additionnées pour former des index).

    • Elles peuvent être discontinues (ou discrètes), si elles prennent seulement certaines valeurs entières, qui sont soit des catégories (différents stades de la dyspnée, grades successifs d'évolution d'une tumeur cancéreuse), soit des valeurs disposées régulièrement le long d'une échelle comportant un intervalle constant entre chaque valeur (nombre de crises d'épilepsie par mois, nombre d'articulations inflammatoires).

    • Elles peuvent être continues, si elles peuvent prendre virtuellement toutes les valeurs possibles entre les deux valeurs extrêmes de l'échelle de réponse (pression artérielle, poids). Les intervalles sont constants et connus.

Le choix du type d'échelle de mesure dépend des variables concernées et des objectifs. Dans un questionnaire destiné à recueillir des informations sur la consommation de calcium, on peut quantifier le lait consommé et construire un indice de consommation de calcium en mg par jour. Si l'objectif est simplement de partager les individus en groupes, on peut exprimer les résultats en catégories ordonnées (moins de 500 mg/j, de 500 à 1.000 mg/j, plus de 1.000 mg/j). Les catégories ont l'avantage d'avoir une signification pratique en clinique. Elles ont

cependant un caractère plus arbitraire et peuvent aboutir à méconnaître des différences significatives entre les groupes si elles n'ont pas été judicieusement choisies.

Il faut s'assurer que les variables qu'on mesure sont appropriées à l'objectif de la recherche. Pour cela on consulte un groupe d'experts, puis un échantillon de la population sur laquelle on désire appliquer l'instrument.

Une fois l'instrument choisi, il faut évaluer ses qualités dans le cadre des objectifs particuliers de la recherche envisagée.

II - EVALUATION DE LA REPRODUCTIBILITE D'UN INSTRUMENT

On distingue classiquement deux grands types de caractéristiques liées à un instrument de mesure: la reproductibilité et l'exactitude (tableau 1).

La reproductibilité d'un instrument est sa capacité à fournir une mesure identique de façon répétée (capacité du thermomètre à indiquer la même température de façon répétée, …). La reproductibilité est essentiellement liée à l'erreur aléatoire. Plus l'erreur aléatoire est petite, plus la reproductibilité est bonne.

L'exactitude d'un instrument est sa capacité à fournir une mesure exacte du phénomène à mesurer (capacité du thermomètre à indiquer la température exacte, …). L'exactitude est liée à la fois à l'erreur aléatoire et aux biais. L'exactitude (appelée encore « validité ») est définie par la sensibilité, la spécificité et est développée dans le chapitre IX.

A - Evaluation de la reproductibilité

Elle comporte les étapes suivantes:

- Enumérer les sources d'erreurs potentielles (inclinaison du rayonnement par rapport au corps du patient pour les radiographies, heure de la journée pour le poids ou la taille, saison pour l'évaluation de l'activité physique, conditions dans lesquelles se déroule la mesure de la pression artérielle, …).

- Classer ces sources d'erreurs potentielles par ordre d'importance décroissante, en se fondant sur l'avis d'experts, les données de la littérature, et les conditions dans lesquelles on veut utiliser l'instrument. Dans un essai multicentrique, il est indispensable d'étudier la variation entre les centres au cours d'une étude pilote. En revanche, si l'étude envisagée a lieu dans un seul centre, cette source d'erreur n'existe pas.

- Déterminer si ces variations existent effectivement en pratique, et le cas échéant mesurer leur importance et tenter de les réduire au maximum.

Il est impossible d'établir ici une liste exhaustive de toutes les sources de variation possibles. Certaines sont spécifiques du type de mesure, d'autres existent, à des degrés variables, avec presque tous les types de mesures:

- la variation inter-investigateurs (variation entre des mesures du même phénomène par des investigateurs différents);

- la variation intra-investigateur (variation entre des mesures du même phénomène par le même investigateur à différents moments);

- la variation intra-patient (variation de la mesure ou du phénomène lui-même chez un même patient à deux moments différents).

Pour évaluer ces sources d'erreurs potentielles, on réalise des mesures répétées en faisant varier une seule source d'erreur à la fois. Il faut s'assurer qu'entre ces mesures répétées, la valeur réelle n'a pas changé. Ceci a une implication dans le choix du temps écoulé entre deux mesures répétées:

- si on mesure des variables labiles, il faut choisir un intervalle de temps très court entre les mesures successives afin d'éviter un réel changement de valeur;

- en revanche, pour des variables très stables telles que la taille à l'âge adulte ou la densité osseuse, le temps écoulé entre deux mesures peut être plus long, en conservant l'assurance qu'il n'y a pas eu de changement de la valeur réelle.

Prenons l'exemple du diagnostic radiologique de cancer pulmonaire. Pour évaluer la variabilité inter-investigateurs, chaque radiographie est lue par chacun des radiologues en insu et on mesure la concordance de leurs conclusions. La variabilité intra-investigateur est estimée en montrant plusieurs fois au même investigateur la même radiographie. La variabilité intra-patient est estimée en faisant plusieurs radiographies du même patient et en les faisant interpréter par un seul radiologue.

B - Mesure de la reproductibilité

Dans le domaine de la biologie, les laboratoires effectuent continuellement des mesures de reproductibilité en partageant les sérums en deux et en mesurant la variabilité de la mesure pour un même sérum. Cette variabilité est exprimée en écart-type autour de la moyenne. Par exemple, la reproductibilité pour le dosage du sodium est de ± 2,3 mmol/l. Comme les valeurs normales sont comprises entre 130 et 150 mmol/l, il est facile de juger de l'acceptabilité de cette erreur.

En clinique, on est souvent amené à travailler avec des données discontinues. Supposons qu'on réalise des radiographies pulmonaires chez 110 sujets exposés pour déterminer la présence ou l'absence de signes de pneumoconiose. Les radiographies sont montrées successivement à deux radiologues et on examine dans quelle mesure leurs avis sont concordants (tableau 2). Dans cette table 2×2, le pourcentage de concordance entre les radiologues est représenté par la case en haut à gauche (présent-présent) et celle en bas à droite (absent-absent). Il est de $\frac{14 + 81}{110}$, soit 86 %.

Il peut y avoir plus de deux catégories. Prenons l'exemple de 100 femmes atteintes de cancer du sein examinées indépendamment par deux cliniciens qui doivent attribuer un degré de sévérité sur une échelle de I à IV selon des critères prédéfinis (tableau 3). La concordance parfaite entre les cliniciens est: $\frac{25 + 14 + 17 + 10}{100}$, soit 66 %, ce qui ne paraît pas très bon. Cependant, lorsqu'on étudie les 34 cas dans lesquels les cliniciens ne sont pas d'accord, la différence n'est le plus souvent (27 fois) que d'un stade.

Pour ce type de données à plus de deux catégories de réponse, la concordance parfaite n'est pas un très bon reflet de la situation, car elle ne donne aucune indication sur l'importance de la discordance. Elle est également trop dépendante du nombre de catégories: plus celui-ci est grand, plus les chances de parfaite concordance diminuent.

Enfin, le simple calcul du pourcentage de concordance ne tient pas compte des cas où la concordance est due au hasard. Pour pallier ces inconvénients, plusieurs index statistiques de concordance ont été proposés (*cf infra*).

Ces index de concordance ne peuvent être appliqués que dans des études simples et lorsqu'on utilise des variables discontinues. Lorsque le plan d'étude est plus complexe, qu'il y a plus de deux observateurs concernés ou que les sources de variation sont plus nombreuses, il est préférable d'utiliser des méthodes mathématiques adaptées, permettant d'étudier plusieurs sources de variation à la fois.

Ces méthodes, fondées sur l'analyse de variance, évaluent la part de la variation liée aux différences réelles - soit entre les sujets, soit chez un même sujet après un traitement ou un intervalle de temps suffisamment long - et la part de la variation due aux erreurs de mesure.

On utilise également les principes de l'analyse de variance lorsque les variables étudiées sont continues. De ces méthodes sont dérivées des statistiques utilisées spécifiquement pour l'évaluation des instruments de mesure.

Une fois les sources d'erreurs identifiées et leur importance évaluée, il s'agit de les réduire. Dans l'exemple précédent, on peut réduire l'erreur due aux variations inter-investigateurs en n'utilisant qu'un seul radiologue pour lire toutes les radiographies ou encore en réalisant une formation identique de tous les radiologues. On peut diminuer l'erreur due à la variation intra-patient en appliquant des règles et des standards stricts pour la réalisation des radiographies (positionnement des patients, constantes radiologiques…).

C - Importance de la reproductibilité

L'exactitude d'un instrument est limitée si sa reproductibilité n'est pas suffisante (fig. 1). On doit donc évaluer d'abord la reproductibilité avant de se lancer dans une étude de validation complète car une mesure non reproductible est inutile.

Le manque de reproductibilité peut avoir de sérieuses conséquences sur tous les types d'études:

- dans les essais thérapeutiques, cela diminue la puissance de l'étude en augmentant la variance de la variable d'intérêt;

- dans les études épidémiologiques d'étiologie ou de causalité, on recherche le rôle d'un facteur ou d'une exposition dans la survenue d'une maladie. Si la mesure du facteur de risque ou celle de la maladie étudiée n'est pas reproductible, leur association est sous-estimée. Cela peut masquer une relation réelle et significative entre facteur de risque et maladie (augmentant ainsi le risque de ne pas détecter une relation réelle). Lorsqu'on veut mesurer les facteurs de confusion, cela réduit la capacité de contrôler ces facteurs, biaisant ainsi les conclusions de façon non prédictible.

D - Importance de l'échantillonnage

Certaines règles sont à respecter concernant le choix des sujets et la technique d'échantillonnage.

La reproductibilité et l'exactitude d'un instrument peuvent varier selon le type de sujets sur lesquels elles sont déterminées. Il faut donc les évaluer sur des sujets similaires à ceux à qui on veut appliquer l'instrument. Les mesures de densité osseuse sont moins reproductibles chez les sujets ostéoporotiques que chez les sujets sains, car les vertèbres sont moins nettement visibles, et la mesure de leur contour est entachée d'une plus grande erreur. Si on veut utiliser la densitométrie osseuse pour distinguer les sujets sains des sujets ostéoporotiques, il faut évaluer la reproductibilité de cet examen dans la population générale. En revanche, si on veut distinguer, parmi des sujets ostéoporotiques, différents stades de gravité, il faut évaluer la reproductibilité et l'exactitude de l'appareil dans un échantillon de sujets ostéoporotiques et non dans la population générale.

E - Méthodes statistiques pour l'évaluation de la reproductibilité des instruments de mesure

Pour étudier la reproductibilité d'un instrument, on utilise la concordance entre des valeurs obtenues par le même instrument dans différentes conditions. Pour étudier la variation entre différents radiologues étudiant les mêmes radiographies, on compare les résultats obtenus par chaque radiologue et on regarde dans quelle mesure ces résultats sont concordants en utilisant des statistiques de concordance.

Il faut distinguer la concordance entre deux variables d'autres relations plus vagues et moins strictes telles qu'une simple association ou corrélation. Cette distinction importante entre concordance et association est la base des principes généraux utilisés pour l'analyse statistique des mesures.

1 - Principes généraux

Pour décrire la relation qui existe entre deux variables dans une population donnée, les index utilisés habituellement représentent la mesure dans laquelle les variations d'une des variables sont similaires à celles de l'autre variable. Ces deux variables peuvent ou non être exprimées dans des échelles de mesure similaires.

Exemples:

- Si on s'intéresse à la relation entre taux de cholestérol et âge (le cholestérol étant une variable continue et l'âge étant exprimé en catégories de 10 ans), un indice d'association décrit dans quelle mesure le taux de cholestérol monte (ou descend) d'une décennie à l'autre. Ces indices d'association sont le coefficient de corrélation linéaire (r), le coefficient de régression, le coefficient rho de Spearman (ρ), le coefficient tau de Kendall (τ).

- On veut étudier la relation entre le taux d'hémoglobine et le taux de créatinine chez les patients atteints d'insuffisance rénale. Le coefficient de régression indique avec quelle force la valeur de la créatinine sérique influence celle du taux d'hémoglobine (le coefficient de régression correspondant à la pente de la droite de régression), et le coefficient de corrélation r détermine la relation linéaire qui existe entre ces deux variables, c'est-à-dire à quel point ces deux variables varient ensemble et de la même façon.

Concernant les instruments de mesure, lorsque deux instruments sont sensés mesurer le même phénomène, il ne suffit pas d'apporter la preuve que leurs mesures sont simplement corrélées

au-delà du simple hasard. Ces mesures devraient en théorie être identiques s'il n'y avait pas d'erreur de mesure. Ainsi la relation qu'on veut mettre en évidence doit être très forte, il ne s'agit pas d'une corrélation mais d'une concordance. Les index d'association classiques sont inadaptés pour décrire la concordance parce que deux variables peuvent avoir une relation très proche sans jamais être en parfait accord. Deux instruments A et B mesurant la densité osseuse peuvent varier dans le même sens et dans la même proportion même si A donne toujours des mesures plus élevées que B. Par ailleurs, deux variables peuvent avoir une très forte corrélation négative, ce qui correspond à une très mauvaise concordance malgré la très forte corrélation.

Plusieurs index de concordance ont été proposés pour évaluer la reproductibilité et l'exactitude des instruments. Seul le coefficient kappa est étudié ici.

Il est nécessaire, pour utiliser des indices de concordance, que les deux variables soient exprimées dans la même unité. Si on doit comparer deux variables qui s'expriment sur des échelles différentes, on ne peut utiliser des index de concordance, mais seulement des mesures d'association.

2 - Type des données

Les index de concordance utilisés dépendent du type des données.

a - Données qualitatives

On peut calculer simplement le pourcentage de concordance. Cependant, cette mesure englobe le pourcentage de concordance qui existerait par hasard, même si les deux variables étaient totalement différentes et indépendantes. Le coefficient kappa est plus approprié car il mesure la concordance entre les variables en tenant compte de l'effet du hasard.

Ne pas prendre en compte l'effet du hasard peut conduire à de fausses conclusions en surestimant la concordance réelle. Pour une variable binaire, la concordance prévisible par pur hasard est donnée par la formule:

$[p_1 \times p_2] + [(1-p_1)\,(1-p_2)]$

où p_1 est la proportion de personnes ayant une caractéristique selon une mesure X, et p_2 la proportion de personnes ayant la même caractéristique selon une mesure Y.

Le coefficient kappa est défini ainsi:

$$\text{Kappa} = \frac{\text{concordance observée - concordance prévisible}}{\text{1 - concordance prévisible}}$$

Lorsque deux mesures sont concordantes à un degré qui n'est pas supérieur au pur hasard, la valeur de kappa est zéro. Lorsque les deux mesures sont parfaitement concordantes, kappa est égal à 1.

Exemple: une étude compare l'information sur l'observance d'un traitement à la réserpine obtenue par l'interrogatoire des patients d'une part et par les dossiers médicaux d'autre part (tableau 4).

Pourcentage de concordance due au hasard = $\frac{(21\times 39)+(196\times 178)}{(217)^2} = 0{,}7583$

Pourcentage de concordance observée = $\frac{14 + 171}{217} = 0{,}8525$

Kappa = $\frac{0{,}8525 - 0{,}7583}{1 - 0{,}7583} = 0{,}39$

Dans cet exemple, ne pas prendre en compte le hasard aurait conduit à surestimer nettement la concordance (85 % au lieu de 39 %).

b - Données ordinales

Dans ce cas également, kappa est la statistique de choix mais on peut le modifier en kappa pondéré. Le principe est de prendre en compte l'importance de la discordance en considérant tous les différents degrés de concordance partielle.

Par exemple, s'il y a quatre catégories de réponse (tableau 3): une discordance entre un stade I selon l'observateur A et un stade II selon l'observateur B est moins grave que si l'observateur A estime le stade à I et l'observateur B le stade à IV.

c - Données continues

Les index à utiliser sont les coefficients de corrélation intra-classes qui associent une mesure de corrélation à une statistique testant la différence entre les moyennes. En termes de régression linéaire, cela signifie que ces coefficients testent non seulement la similarité des pentes des droites de régression, mais également que la droite passe par 0.

L'utilisation d'une analyse de variance à plus de deux entrées permet d'étudier les variations provenant de plusieurs sources simultanément et correspond à des méthodes qui sont au delà des objectifs de ce chapitre.

III – QUESTIONNAIRE : SAVOIR EXPRIMER CE QUE L'ON SOUHAITE

Questionnaire: de quoi s'agit-il ? Du document imprimé ? De l'ensemble constitué des questions posées à un individu ? Ou du concept global "entretien-questions-réponses-enregistrement des réponses" qui conduit *in fine* à l'élaboration du document imprimé ?

Dans un essai clinique, la forme papier ou électronique du questionnaire, souvent nommée "bordereau" ou "formulaire" est destinée à stocker les informations récoltées lors de l'étude. Cette version est d'autant plus facilement élaborée qu'elle suit la séquence chronologique de l'essai (par exemple, un questionnaire par visite médicale, dit bordereau de visite) et qu'elle est destinée à consigner des données ne prêtant pas sujet à interprétation (par exemple, les valeurs diastolique et systolique de la pression artérielle).

Il en va tout autrement d'un questionnaire destiné à l'évaluation de données subjectives concernant le vécu des patients, leur douleur, leur qualité de vie. Comment mesurer objectivement le subjectif ? Mesure-t-on bien ce qu'on souhaite mesurer ? Le vocabulaire

utilisé est-il compréhensible par tous et de la même manière (en Afrique centrale, fièvre signifie à la fois paludisme et fièvre) ?

Le questionnaire, qui constitue un exercice de communication pour les deux parties, ne peut être élaboré que lorsqu'une des parties sait ce qu'elle veut communiquer et obtenir de la part de l'autre. Il n'est pas envisageable de le construire avant que la question de recherche à laquelle on souhaite répondre, ainsi que le plan de l'étude ne soient connus avec précision

A - Définition et validation du champ d'exploration

Bien avant la rédaction des questions, il est primordial :

- de définir avec précision ce qui sera mesuré. Un questionnaire est d'abord destiné à la mesure de l'élément d'intérêt (par exemple : douleur, satisfaction…);

- et de s'assurer qu'un terme employé recouvre le même concept et la même définition pour tous.

La définition correcte des termes représente un tel obstacle que l'on fait souvent appel à un groupe multidisciplinaire d'experts lors de la construction d'un nouveau questionnaire.

L'étude de la littérature est d'un précieux secours car de nombreux questionnaires ont déjà été validés. Ils peuvent, sous certaines conditions, être réutilisés en l'état.

B - Les divers types de questions

On distingue trois grandes catégories de questions, non exclusives. Le choix dépend de la nature de l'information recherchée et de la façon dont l'analyse ultérieure est envisagée.

1 - La question ouverte

Exemple : "Comment décririez-vous votre douleur ?"

C'est la moins contraignante pour le sujet questionné. Il lui est possible d'employer ses propres mots pour exprimer toute la panoplie de ses sentiments. Ceci est particulièrement important lorsque l'expérimentateur cherche avant tout à définir l'étendue de ce qui est ressenti par le patient, la variété des opinions de l'équipe soignante …

La synthèse, à des fins de codage et d'analyse statistique, d'un ensemble aussi divers de réponses à une même question est compliquée. Elle peut conduire à des erreurs d'interprétation. La grande latitude laissée au sujet dans sa réponse expose l'investigateur à sa bonne volonté et/ou sa capacité à exprimer et transmettre ce qu'il ressent. Le codage, qui implique une réduction (par un processus de standardisation) de l'information, conduit à un appauvrissement de l'information initialement recueillie.

2 - La question semi-ouverte

Exemple : "Lors de votre dernière consultation médicale, votre médecin a-t-il également examiné un (ou plusieurs) membre(s) de votre famille: père, mère, frère(s), sœur(s), enfant(s), autre(s) ?"

Avec ce type de question, l'effort de mémoire est minimum. La réponse est suggérée et orientée.

3 - La question fermée

Exemple : "Etes-vous né(e) un vendredi 13 ?"

Elle impose une réponse univoque parmi celles proposées par l'équipe qui a rédigé le questionnaire (oui, non, ne sait pas). Aucun effort (ou presque) de mémorisation n'est demandé. Le codage de la réponse est facile et (au risque d'erreur près) fiable. En revanche, aucun choix n'est possible en dehors des réponses proposées. Ce n'est pas dommageable pour l'exemple donné ci-dessus. Cela l'est certainement pour l'évaluation des émotions, du comportement, de l'apport calorique, de la douleur…

C - Un vocabulaire approprié

Y a-t-il des mots à éviter ou d'autres à employer plus particulièrement ? Chaque mot simple et familier, faisant référence à un concept clair peut être utilisé. Dans le cas contraire, il doit être évité.

Comment le savoir ? Une équipe expérimentée choisit le vocabulaire en fonction du public visé et évite, lors de la construction d'un questionnaire destiné au "patient moyen", l'emploi de termes culturellement marqués, nouveaux, de termes scientifiques ou médicaux précis mais trop spécialisés, de termes insuffisamment répandus dans la population, ou à l'inverse trop flous ou recouvrant un sens large et imprécis, ou diversement interprétables, … Un individu peut s'avérer incapable de décrire son "statut marital" alors qu'il est assurément capable de dire s'il est célibataire, concubin, marié, divorcé, ou veuf.

Un vocabulaire simple, clair et précis dont le sens est identique pour tout sujet potentiel quel que soit son statut socio-économique, sa catégorie socioprofessionnelle ou son niveau d'éducation est un gage de qualité du questionnaire.

D - Formulation de la question

Il faut tendre vers la clarté et l'absence d'ambiguïté. Il s'agit avant tout d'éviter les formulations trop générales, complexes, ambiguës, litigieuses. Quelques exemples sont donnés dans le tableau 5.

E - Organisation des questions

Le but d'un questionnaire est de recueillir des données de qualité, communiquées librement par le sujet qui participe à l'étude. Le souci est d'éviter que le questionnaire soit mal ou incomplètement rempli.

Il faut intéresser le patient, lui donner envie de répondre et donc se concentrer sur la conception du fond autant que de la forme du questionnaire:

- au début figurent les questions auxquelles on tient absolument à avoir une réponse et auxquelles le patient a envie de répondre. C'est l'accroche initiale, attractive, garantie de réponse;

- puis viennent les questions par thème, groupées, car cela facilite grandement le déroulement de l'entretien (avez-vous déjà été opéré de la prostate ? Si oui, répondez aux 6 questions qui suivent, sinon allez directement à la question 22);

- les questions les plus personnelles prennent en général place à la fin, lorsque le sujet est confiant et cherche moins qu'au début à organiser ou à contrôler l'information qu'il délivre.

F - Codage

Convertir les réponses en chiffres, plus facilement utilisables à des fins d'analyse, implique l'emploi d'un système de conversion.

Deux possibilités s'offrent au chercheur :

- l'élaboration d'un système personnel de codage, indispensable si le champ de l'étude est nouveau ou très spécialisé,

- ou l'application de systèmes pré-établis et déjà testés : connaître la prévalence des maladies et leurs complications dans un hôpital peut se faire en appliquant une classification nationale ou internationale, par exemple la Classification Internationale des Maladies de l'Organisation Mondiale de la Santé, dont les codes sont accessibles à tous.

Le codage est fonction du type de question.

Une question fermée, imposant une réponse parmi celles proposées, permet un codage facile d'emblée.

En revanche une question ouverte ne permet pas de prévoir *a priori* l'ensemble des réponses. Les réponses sont alors regroupées par catégories logiques, de façon subjective, à l'issue de l'étude, par l'investigateur.

Où faire figurer les instructions de codage ? Deux options s'offrent à l'investigateur:

- sur le questionnaire lui-même

Au niveau de chaque question, figurent les valeurs de codage pour chaque variable. Il est utile pour l'épidémiologiste, dans la gestion des variables, ainsi que plus tard pour le statisticien au cours de l'analyse, de situer la place relative occupée par la valeur codée de la variable dans l'ensemble du questionnaire (tableau 6). Le codage figure en permanence sous les yeux, les risques d'erreur de transcription sont moindres. Mais l'espace nécessaire impose une augmentation de la longueur du questionnaire, particulièrement lorsque les instructions de codage sont longues et complexes. Le volume, le poids ainsi que le coût généré sont généralement plus élevés.

- dans un document annexe (tableau 7)

Le questionnaire est allégé dans sa présentation: ne figure que la question. Mais l'information de codage est moins facilement et moins rapidement disponible, un document annexe s'égare, …

Le codage est aussi fonction de la question de recherche et de l'analyse statistique prévue. Il est de tradition de réserver certains codes à des valeurs particulières: "non" prend souvent la valeur 0, "oui" la valeur 1, les valeurs manquantes les valeurs 9, 99, …

La question de recherche peut cependant conditionner les valeurs des variables d'une façon particulière. Le statisticien peut, lors de l'analyse, recoder des variables (par transformation) si le codage initial s'avère inadéquat.

Consulter un épidémiologiste ou un biostatisticien à ce stade est certainement sage.

G - Utilisation d'un questionnaire étranger

Le chercheur peut être tenté d'utiliser un questionnaire développé à l'étranger pour un problème de même ordre. Outre la difficulté d'une traduction parfaite se pose la question du transfert et de l'applicabilité des concepts culturels qui ont présidé à l'élaboration du questionnaire.

La demande exprimée par un patient devenu lombalgique à la suite d'un accident du travail n'est pas identique selon qu'il est heureux ou non dans l'accomplissement de son travail quotidien. Le chercheur doit en tenir compte dans l'analyse des données qu'il a recueillies sur ce thème. Mais a-t-il initialement pris soin de vérifier qu'une même demande était exprimée de façon identique par les sujets de la population de l'étude ? Des patients de cultures différentes n'expriment pas des difficultés similaires par les mêmes symptômes. Conduire une recherche dans ce domaine impose de tenir compte de ces particularités culturelles dès la phase d'élaboration du questionnaire.

De même, un questionnaire destiné à l'évaluation de divers aspects de l'état de santé de patients britanniques, bien qu'ayant été validé dans d'autres pays d'Europe du Nord et ayant conduit à des résultats similaires, ne traduirait probablement pas le même concept s'il était appliqué tel quel dans les pays latins de l'Europe du Sud: on ne se fait pas culturellement la même idée de la santé dans chacun de nos pays européens. Il en est de même entre Américains et Français.

Pourquoi utilise-t-on alors parfois des questionnaires étrangers ?

Parce que cela est inévitable dans certaines circonstances, par exemple lorsqu'une étude multicentrique se déroule simultanément dans plusieurs pays dont la langue est différente, ou lorsqu'il s'avère nécessaire de fonder une nouvelle étude sur la méthodologie et le questionnaire d'une étude précédemment publiée, considérée comme la référence en la matière, afin de la répliquer et d'en comparer les résultats. Quelques étapes sont nécessaires à la bonne transposition d'un questionnaire d'un pays (d'une culture) à un autre (tableau 8).

IV - EXACTITUDE ET REPRODUCTIBILITE

Deux qualités sont indispensables à un questionnaire : son exactitude et sa reproductibilité.

L'exactitude est la capacité du questionnaire à fournir une mesure exacte de ce qui est à mesurer, comme la capacité pour le champion de tir à l'arc à atteindre le centre de la cible, ou celle d'un thermomètre à indiquer l'exacte température.

Par analogie avec un test diagnostique, le questionnaire parfaitement exact serait celui qui aurait une sensibilité et une spécificité parfaites (100 %) et pour lequel les valeurs prédictives

positive et négative seraient également de 100 %. Un tel test a-t-il jamais existé ? Le questionnaire administré aux candidats de l'examen du code de la route est non seulement supposé dépister les candidats qui connaissent bien le code de la route (sensibilité du test) et éliminer ceux qui le connaissent insuffisamment (spécificité du test), mais également prédire qu'un candidat qui a réussi le test connaît bien son code de la route (valeur prédictive positive) ou qu'un autre candidat a échoué parce qu'il le connaissait mal (valeur prédictive négative).

Si on ne sait pas quelle mesure choisir, parce que plusieurs peuvent s'appliquer à ce concept, il faut en utiliser plusieurs. Si on ne sait pas quelle dimension du concept mesurer, il faut les mesurer toutes.

La reproductibilité est la capacité du questionnaire à fournir une mesure identique de façon répétée, comme pour le tireur à l'arc à mettre toutes ses flèches au même endroit, ou pour le thermomètre à indiquer la même température de façon répétée. La condition de reproductibilité est vérifiée dès que le même test, appliqué plusieurs fois de suite dans les mêmes conditions, conduit à chaque fois au même résultat. Il s'agit donc d'une qualité bien différente de la précédente mais tout aussi importante.

La figure 1 illustre ces deux concepts.

Comment s'assurer de la qualité d'un questionnaire avant son utilisation dans l'étude ?

- En le testant pendant une période de rodage (pré-test) et en l'appliquant à quelque temps de distance (re-test). La concordance entre les réponses doit être élevée, si l'information recherchée n'a eu aucune raison de changer. Ceci teste la reproductibilité.

- Puis en comparant les résultats obtenus par son propre questionnaire à ceux d'un questionnaire de référence précédemment publié par une autre équipe dans des circonstances similaires sur une population similaire. Ceci teste l'exactitude.

- Ou en utilisant une méthode de mesure déjà validée par une autre équipe: il en existe parfois des dizaines pour une même situation à évaluer. Il faut malgré tout re-tester la méthode de mesure dans ce nouveau contexte puisqu'elle a été validée dans des circonstances différentes.

V - PRESENTATION DU QUESTIONNAIRE

Tout questionnaire doit être très clairement identifié.

Le nom de l'étude (par exemple, ETCI2 - ETude de Comparaison de 2 Interventions) et le titre du questionnaire (par exemple, "Evaluation de la qualité de la vie à 6 mois") doivent figurer en gros sur la couverture ainsi que le numéro de code du centre (dans une étude multicentrique), la date à laquelle il a été complété, et l'identification du patient.

Si plusieurs questionnaires différents sont utilisés chez un même patient, chaque questionnaire doit pouvoir être distingué, par exemple grâce à une couleur spécifique, une identification spécifique facilement reconnaissable (par exemple, | Q | U | A | V | 0 | 6 | et | Q | U | A | V | 2 | 4 | représentent respectivement les questionnaires d'évaluation de la qualité de la vie à 6 et à 24 mois).

Les questionnaires du même type, utilisés chez des patients différents, sont identifiés par un code unique qui respecte l'anonymat tout en évitant les confusions entre questionnaires. On

utilise souvent le format suivant: | K | E | R | C | E |, où les trois premières lettres représentent les trois premières du patronyme et les deux suivantes les deux premières du prénom.

Les questions doivent être numérotées dans un ordre séquentiel, sans omission. Il en est de même pour les pages.

Tout symbole qui peut favoriser la rapidité de compréhension du questionnaire, sans en altérer le sens, est le bienvenu. Ceci est particulièrement vrai d'une flèche qui évite l'emploi d'une suite de mots, d'un point (•) ou d'un astérisque (*) placé devant chaque idée afin de la distinguer de la précédente. Il en est de même, pour éviter de se tromper de ligne, des pointillés qui relient la fin de la question à la case dans laquelle figurera la réponse.

Né(e) un Vendredi 13 ? (non = 0; oui = 1; ne sait pas = 9) …… | |

La lisibilité et la compréhension sont toujours facilitées par un espacement suffisant entre les questions.

VI - LE PRE-TEST

Le pré-test représente la phase ultime de la préparation du questionnaire. Il peut être difficilement vécu car la remise en question est parfois douloureuse ! Il représente cependant une étape indispensable.

Il est conduit dans des conditions identiques à celles de l'étude réelle et doit porter sur au moins 20 à 30 individus. Combien de pré-tests ? Il est rare que le questionnaire soit d'emblée parfaitement rectifié : un second pré-test est probablement à prévoir dès la planification de l'étude.

A l'occasion de ce pré-test sont obtenues pour la première fois:

- des données quantitatives, comme le temps nécessaire pour compléter le questionnaire (durée moyenne, durées minimale et maximale), le nombre de questions qui n'obtiennent pas de réponse, le nombre de réponses incohérentes (en désaccord avec une réponse à un précédent item), l'existence d'éventuels questionnaires non complétés, …

- des données qualitatives, peut-être les plus importantes, comme l'impression générale de facilité, ou non, à compléter le questionnaire, ou les difficultés liées au niveau culturel des questionnés (problème plus souvent lié au vocabulaire employé qu'au concept abordé par la question), ou relatives à la séquence logique des questions, …

Les commentaires des questionnés sont très riches d'informations. Ne jamais les négliger et tenter d'en obtenir le plus possible et de la meilleure qualité sont deux règles primordiales. Ces commentaires peuvent aboutir à la résurgence de concepts initialement abandonnés ou ignorés, ou à l'émergence de nouveaux concepts. Ils peuvent aussi concerner la lisibilité du questionnaire, la répétition de questions, l'insuffisance de place laissée pour la réponse aux questions ouvertes…

Conclusion

Les instruments de mesure jouent un rôle important, parfois sous-estimé, dans la qualité d'une étude. Pour déterminer avec précision les relations entre facteurs étudiés et critères de jugement, ces facteurs doivent être mesurés de façon reproductible et exacte. La validation des instruments de mesure est donc capitale et correspond à une démarche délicate et souvent très longue.

Le clinicien ou l'équipe de recherche qui utilisent un questionnaire poursuivent deux buts:

- obtenir l'information nécessaire à leur étude ou à leur enquête,
- mesurer cette information avec le maximum de qualité (en termes d'exactitude et de reproductibilité),

tout en respectant la dignité des patients. La règle d'or est certainement d'établir une relation de confiance dès les premières questions.

Tableau 1 - Les deux grands types de caractéristiques liées à un instrument de mesure

Biais	erreur systématique	exactitude (ou validité)
Chance	erreur due au hasard	reproductibilité (ou fiabilité)

Tableau 2 - Concordance des avis de deux radiologues pour 110 sujets suspects de pneumoconiose

		Radiologue n° 2		
		Présent	Absent	
Radiologue n° 1	Présent	**14**	7	21
	Absent	8	**81**	89
		22	88	110

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Tableau 3 - Concordance des avis de deux cliniciens sur la sévérité de l'atteinte de 100 femmes présentant un cancer du sein

		Clinicien n° B				
		I	II	III	IV	
Clinicien n° A	I	**25**	7	2	1	35
	II	4	**14**	5	0	23
	III	3	6	**17**	3	29
	IV	0	1	2	**10**	13
		32	28	26	14	**100**

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Tableau 4 - Concordance de l'information sur l'observance d'un traitement à la réserpine obtenue par l'interrogatoire des patients d'une part et par la lecture des dossiers médicaux d'autre part, chez 217 patients

		Dossier médical		
		Oui	Non	
Interrogatoire	Oui	**14**	7	21
	Non	25	**171**	196
		39	178	217

Mis en forme : Police :Gras

Mis en forme : Police :Gras

Tableau 5 - Formulations de questions à éviter

- la double négation : pensez-vous que les gens qui n'ont pas d'emploi ne sont pas heureux ?
- la double interrogation (deux questions à la fois) : êtes-vous favorable à l'adjonction systématique, moyennant 2 euros par mois, de fluor dans l'eau de boisson pour éviter les caries chez vos enfants ?
- la suggestion appuyée malgré la forme interrogative : ne trouvez-vous pas que la fluoruration systématique de l'eau de boisson est une atteinte à la liberté individuelle ?
- la combinaison des deux précédentes (suggestion appuyée et deux items explorés par une même question) : préférez-vous une source d'énergie propre comme le nucléaire à une centrale à charbon polluante ?
- le(s) mot(s) au sens imprécis : êtes-vous souvent constipé(e) ?
- la question imprécise : seriez-vous favorable à une réduction modérée de l'apport calorique chez les personnes inactives ?

Tableau 6 - Exemple de codage inclus dans le questionnaire

N° Question	Libellé de la Question	Codage	Repérage
06	En quelle année êtes-vous né(e) ? (Indiquer l'année complète)	|__|__|__|__|	(58-61)
07	Etes-vous né(e) un Vendredi 13 ? (non = 0, oui = 1, ne sait pas = 9)	|__|	(62-62)
08	Etes-vous né(e) en France métropolitaine ? (non = 0, oui = 1, ne sait pas = 9)	|__|	(63-63)
09	Si oui, dans quel département ? (Indiquer le numéro du département)	|__|__|	(64-65)

Tableau 7 - Informations minimales lorsque les règles de codage figurent dans un document annexe au questionnaire

- numéro de la variable
- son nom complet (par exemple, jour de naissance)
- son abréviation: le plus souvent huit symboles au maximum, pour des raisons de compatibilité des fichiers informatiques entre eux (par exemple, JOURNAIS)
- un bref descriptif (par exemple, jour de la semaine lors de la naissance)
- son type: caractère, numérique, logique, date, … (par exemple, numérique)
- sa largeur une fois codée, incluant éventuellement la virgule et les décimales (par exemple, 4 caractères - 2 chiffres, virgule, 1 décimale - pour la variable "température corporelle" - 37,5°C)
- la page (et éventuellement la référence) du questionnaire à laquelle elle se réfère
- la base de données (fichier informatique) à laquelle elle appartient, au cas où il y aurait plusieurs fichiers
- enfin et surtout, les modalités de codage (par exemple, non = 0, oui = 1, ne sait pas = 9)

Tableau 8 - Etapes d'une validation "trans-culturelle"

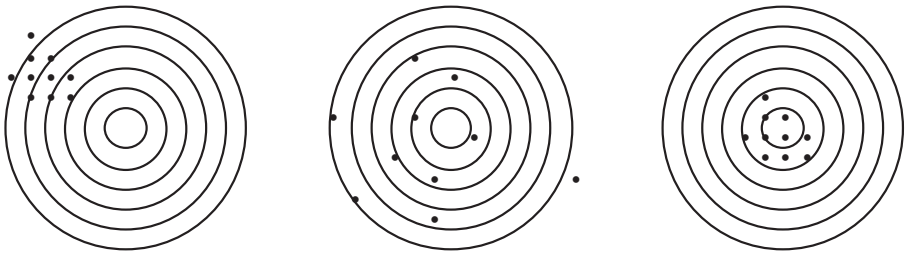
- sélection du questionnaire étranger le plus susceptible de répondre au besoin
- double traduction parallèle et indépendante (de la langue d'origine à la langue actuelle), en évitant le piège de la traduction littérale, par deux personnes connaissant bien l'idée sous-jacente à chaque question (en général, par au moins un des chercheurs bilingues de l'équipe)
- double traduction inverse (de la langue actuelle à la langue d'origine), par d'autres traducteurs que les premiers, afin de vérifier que les concepts véhiculés par chaque question sont encore présents
- arbitrage des différences lors d'une réunion de consensus
- pré-test du questionnaire auprès de futurs interrogés potentiels
- modifications finales

Fig. 1 - Exactitude et reproductibilité: la classique illustration de la cible

Sur la cible du centre, le tir est soumis à une erreur aléatoire et tous les impacts se distribuent au hasard autour du centre: le tir est très peu reproductible, et donc peu exact.

Sur la cible de gauche, les impacts sont regroupés mais systématiquement à côté du but recherché: le tir est très reproductible mais très inexact.

Sur la cible de droite, les impacts sont regroupés autour du centre: le tir est exact et reproductible.



Reproductible, mais non-exact

Non reproductible, et donc peu exact

Reproductible et exact

Chapitre XV

La mesure de la qualité de vie en recherche clinique

CaroleVuillerot, Isabelle Hodgkinson

L'échec relatif des indicateurs objectifs (mortalité, morbidité) à rendre compte de certains bénéfices de la médecine, ou de la chirurgie, a amené le milieu médical à développer des indicateurs subjectifs comme la qualité de vie.

La qualité de vie (QDV) est un critère de jugement. Ce critère peut être utilisé dans tous les types d'étude : cas-témoins, cohorte, essai thérapeutique…. Mais il doit être justifié. L'interprétation des résultats en dépend. Il existe deux approches de la QDV, l'une générale définie par l'OMS, « perception qu'a un individu de sa place dans l'existence, dans le contexte de la culture et du système de valeurs dans lequel il vit et en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes », l'autre plus spécifique au domaine médical, dite QDV liée à la santé « La qualité de vie liée à la santé est la valeur qui est attribuée à la durée de vie en fonction des handicaps, du niveau fonctionnel, des perceptions et des opportunités sociales modifiées par la maladie, les blessures, les traitements ou les politiques de santé »

Pour mesurer la QDV d'une personne dans un contexte de recherche clinique, il est habituel d'utiliser un questionnaire. Celui-ci sera élaboré en fonction des objectifs de l'étude, de la population étudiée et du concept de QDV choisi. Un auto-questionnaire est toujours préférable lorsque la situation le permet. Il doit être d'application facile, validé dans une population de sujets sains, et avoir des qualités métrologiques reconnues.

Malgré ces précautions, les biais d'interprétation sont nombreux. L'interprétation des résultats sera d'autant plus juste que la réflexion en amont aura été approfondie.

I. Introduction

L'échec relatif des indicateurs objectifs (mortalité, morbidité) à rendre compte de certains bénéfices de la médecine, ou de la chirurgie, a amené le milieu médical à développer des indicateurs subjectifs comme la QDV de manière à envisager l'existence d'une autre réalité.

La médecine prend dorénavant en compte l'individu dans sa globalité et non plus seulement dans ses aspects somatiques. Le seul avis des experts médicochirurgicaux dans l'évaluation médicale n'est plus suffisant. Le concept de QDV instaure une approche plus humaniste de la médecine, en modifiant le regard porté par le médecin sur sa relation au malade. Il oublie sa position classique de sujet supposé savoir en acceptant de porter son attention sur les points de vue proprement subjectifs qu'un individu a de lui-même, de sa situation actuelle, de son handicap, de ses attentes et de ses buts dans la vie [1]. Les préoccupations médicales ont évolué, la médecine s'adresse maintenant tant au sujet pensant (souffrant, anxieux) qu'au sujet vivant (dont le corps est malade) [2] [3]. La médecine doit permettre au patient de vivre plus longtemps, en bonne santé, et satisfait de sa vie.

De multiples outils de mesure validés ou non se développent dans le but de fournir des évaluations fiables, spécifiques et reproductibles. A l'heure actuelle, lors de la conception d'un protocole de recherche clinique, il devient indispensable d'ajouter un questionnaire de QDV comme garantie de l'intérêt porté au patient. Mais cette attitude systématique n'a que peu de valeur. Que la QDV soit le critère principal d'évaluation de l'efficacité d'un traitement, ou bien l'un des critères secondaires, dans tous les cas sa signification et l'interprétation des résultats dépendent de multiples facteurs, en particulier de la pertinence de l'outil utilisé et du cadre conceptuel de QDV choisi.

2. Qu'est ce que la qualité de vie ?

2.1. Concept général de qualité de vie

Il n'existe pas de définition universelle de la QDV ; chacun employant sa propre terminologie. La QDV est une notion subjective, dynamique et influencée par le contexte environnemental. Comme le bonheur ou la tristesse, lorsque l'on utilise ce terme, tout le monde le comprend, chacun ayant sa propre définition. La signification de ce concept est intuitive, spontanée et propre à chaque personne.

Une première approche est représentée par le ressenti de la personne sur sa propre existence. Il s'agit, en particulier, de sa satisfaction dans les différents domaines de sa vie ; domaines variables en fonction des âges de la vie. Ces différents domaines recouvrent « le large éventail des dimensions de l'expérience humaine depuis celles associées aux nécessités de la vie jusqu'à celles associées à un sentiment d'accomplissement, de réussite et de bonheur personnel » [4].

Une deuxième approche définit la QDV comme la différence entre les attentes d'un individu et sa situation présente. Avec en plus, dans le cas de l'enfant et de l'adolescent, les attentes des parents qui peuvent être différentes de celles de l'enfant. Pour nombre d'auteurs nord-américains, la qualité de vie se décompose en « being » (ce qu'on est), « becoming » (ce qu'on va -ou veut -devenir), et « belonging » (comment on se sent inséré dans un groupe d'appartenance ou la société : notions d'adaptation, d'acceptation sociale, d'appartenance) [5]. D.Curran [6] évoque la capacité de l'être humain à adapter ses attentes personnelles à ce qu'il perçoit comme compatible avec sa condition. Ces adaptations permettent aux personnes ayant des conditions de vie difficiles d'acquérir un sentiment de QDV convenable [7]. Ainsi chaque individu, selon son histoire, sa situation actuelle et son cheminement personnel, peut percevoir une situation de vie comme permettant une excellente ou une déplorable QDV. Le

jugement d'un médecin sur la QDV de son patient est alors rendu impossible. Plusieurs études confirment cette hypothèse, comparant l'estimation de la QDV par le médecin, et par le patient lui-même : le patient s'estime souvent beaucoup plus satisfait de sa vie que ne le juge son médecin. [8][9].

Ces approches permettent toutes deux d'affirmer que la personne concernée est seule capable d'évaluer sa propre QDV : « perception qu'a un individu de sa place dans l'existence, dans le contexte de la culture et du système de valeurs dans lequel il vit, et en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes »[10].

Cette définition de l'OMS s'organise ainsi autour de trois dimensions principales : la dimension physique ou physiologique, la dimension mentale ou psychologique et la dimension sociale ou environnementale.

La dimension physique est abordée sous l'angle des données subjectives. Il ne s'agit pas de mesurer un score fonctionnel ou un degré d'invalidité mais la façon dont le sujet perçoit et exprime sa dépendance physique et certains symptômes comme la douleur ou la fatigue.

La dimension psychologique explore l'état mental (dépression, estime de soi, structures de personnalité) exprimé par le sujet dans ses réponses au questionnement.

Les interactions sociales et familiales constituent la troisième dimension du concept de QDV. Chez l'adulte, contrairement à l'enfant, une des composantes majeures de cette dernière dimension est représentée par le statut socio-économique.

2.2. Concept de qualité de vie liée à la santé

Si la QDV est définie comme l'évaluation de la satisfaction et du bien-être de l'individu dans les différents domaines de sa vie, la QDV liée à la santé ou Health Related Quality Of Life (HRQOL) est alors considérée comme un sous-domaine de ce concept multidimensionnel. Sa

mesure est concentrée sur des aspects directement liés à l'état de santé. « La qualité de vie liée à la santé est la valeur qui est attribuée à la durée de vie en fonction des handicaps, du niveau fonctionnel, des perceptions et des opportunités sociales modifiées par la maladie, les blessures, les traitements ou les politiques de santé » [12].

Lorsque l'on mesure la QDV du sujet atteint d'une maladie très invalidante, on mesure en fait sa QDV liée à sa santé, tant la maladie est au cœur de l'existence du sujet.

Les définitions de QDV liée à la santé retrouvées dans la littérature comprennent souvent le statut fonctionnel et l'état de santé. Même s'il est certain que QDV, fonction et santé sont liées, cela ne signifie pas qu'elles sont interchangeables [11]. Pour illustrer ce propos, nous citerons le *Sickness Impact Profile* (SIP)[28]. Ce questionnaire comprend 136 items répartis en trois dimensions : la dimension physique (déambulation, soins du corps, mouvement, mobilité), la dimension psychosociale (interactions sociales, comportement émotionnel, communication) et la dernière dimension regroupant des sous-dimensions indépendantes (le sommeil, les repas, le travail, les travaux ménagers, les loisirs). Dans le cas de patients ayant une déficience motrice, avec ce questionnaire, nous mesurons l'indépendance fonctionnelle, les incapacités, les difficultés de la vie quotidienne, mais non la QDV du sujet. Nous obtenons l'avis du sujet sur des données objectives de sa vie bien loin de la notion de QDV telle que définie par l'OMS. Nous obtenons une information sur ce que la personne est capable de faire mais beaucoup moins sur ce qu'elle ressent.

3. La mesure de QDV

Mesurer « consiste (à établir) des règles d'attribution de nombres à des objets et à représenter des quantités d'attributs » [13]. Le fait de vouloir quantifier ou mesurer une valeur qualitative semble de prime abord suspect. Bergson, en 1889, [14] s'interrogeait déjà sur la volonté des médecins, pour rendre la psychologie « scientifique », de réduire le qualitatif au quantitatif. Il

s'opposait à la volonté des médecins de faire passer « des perceptions de grandeur pour des grandeurs de perception ». Le concept de QDV est par définition en dehors du spectre de la quantité, puisqu'il peut être défini comme ce qui fait l'essence de la vie. La préséance de la qualité sur la quantité n'est pas remise en question par la volonté de mesurer cette QDV. C'est le seul moyen que le monde médical a trouvé pour mesurer le gain ou la perte de bien-être engendrée par les interventions thérapeutiques. Nous devons mesurer la QDV de nos patients en nous fondant sur l'opinion des malades, c'est-à-dire sur des critères purement subjectifs. Pour convertir la QDV d'un patient en donnée chiffrée, on peut raisonner en considérant que le bien-être a plusieurs composantes ou dimensions et qu'à l'intérieur de chaque dimension il admet le plus et le moins. En d'autres termes, à l'intérieur de chaque dimension, il y a des degrés dans la qualité du fonctionnement qui permettent de préciser la QDV de l'individu. Nous convertissons alors une qualité en quantité en utilisant des questionnaires standardisés dont les réponses chiffrées des patients en termes de satisfaction constitueront notre profil ou index chiffré de QDV. Ces mesures associent des valorisations subjectives à des éléments descriptifs et sont appelées mesures composites.

Pour mesurer la QDV d'une personne dans un contexte de recherche clinique, il est habituel d'utiliser des questionnaires de QDV. On pourrait évaluer la QDV, en demandant à une personne d'estimer son niveau de QDV, sur une échelle de 1 à 10. D'abord le résultat obtenu ne serait que global, peu précis et difficilement interprétable. En effet, comment interpréter une évaluation de QDV à 4,3 ? De plus, face à une telle question, la réponse du sujet est rapide, sans réflexion approfondie sur son existence et son ressenti. Le fait d'utiliser des questionnaires avec des sous-dimensions permet au sujet de cheminer dans l'analyse des différentes facettes de sa vie. L'information obtenue est plus précise, on obtient alors des scores par dimension, à l'intérieur desquelles chaque question peut elle aussi être interprétée.

On peut aussi approcher de façon très différente la QDV d'un groupe d'individus par les mesures économétriques (d'utilité) ou en utilisant l'indicateur QALY (Quality Adjusted Life Years), c'est-à-dire le nombre d'années de vie gagnées ajustées sur la qualité de vie (liée à la santé) (voir chapitre X). Cet indicateur constitue une tentative d'estimation objective de la QDV sans tenir compte de l'opinion directe des intéressés. L'indice des QALY est habituellement utilisé pour évaluer plusieurs thérapeutiques médicales, en comparant les années de survie qu'elles offrent, « ajustées » par un jugement de la qualité de cette survie. Cette méthode permet d'étudier l'impact de certains états morbides et de comparer le coût-bénéfice de certaines interventions qui prolongent la vie au prix d'effets secondaires ou de séquelles physiques, des conséquences psychologiques et/ou sociales. Ces mesures permettent aux professionnels de santé de justifier certains choix thérapeutiques [15]. Elles ne s'intéressent pas au ressenti du patient dans son individualité, mais essaient d'estimer au niveau d'un groupe de patients l'impact d'une intervention thérapeutique et ce, dans un souci de rationalisation des soins et de politique de santé publique.

4. Le choix d'un questionnaire de QDV pour un projet de recherche clinique

Que la QDV soit le critère de jugement principal ou secondaire de l'étude, sa délimitation conceptuelle dans la population étudiée doit être précisée en fonction des objectifs poursuivis par l'étude. Et ce choix doit être justifié dans le protocole. Ainsi le questionnaire sera choisi en fonction des objectifs de l'étude, de la population étudiée et du concept de QDV choisi.

4.1. Questionnaire générique ou spécifique

Il existe deux types de questionnaires de QDV, les **questionnaires génériques** et les **questionnaires spécifiques**. Les questionnaires génériques sont élaborés à partir de questionnaires mis au point en population générale. Ils ne peuvent jamais rendre compte avec

détail de la spécificité de chaque pathologie abordée, mais permettent de comparer les résultats de plusieurs études sur différentes populations. Ils sont peu sensibles au changement de l'état de santé. Au contraire, les questionnaires spécifiques évaluent « la partie de la santé, du bien-être ou de la qualité de vie qui est affectée en priorité par la pathologie en question » [13]. L'instrument peut être spécifique d'une pathologie donnée ou d'une population donnée. Un instrument spécifique peut contredire les données apportées par un instrument générique dans la mesure où l'instrument générique ne peut tenir compte des particularités de la population étudiée.

Par exemple, pour un essai thérapeutique concernant l'évaluation d'un nouveau stylo à insuline dans une population d'enfants diabétiques, si on envisage de mesurer la QDV, le questionnaire le plus adapté est un questionnaire spécifique pour les enfants diabétiques en lien avec les aspects spécifiques de la maladie (contrainte des injections 3 fois par jour, rapport aux autres enfants, etc….). L'outil spécifique a dans ce cas toute sa valeur car il aborde toutes les facettes de la QDV de l'enfant en lien avec son diabète.

4.2 Auto-questionnaire versus hétéro-questionnaire

En règle générale, l'autoévaluation fait l'objet d'un consensus [16]. La personne elle-même est la mieux placée pour évaluer sa propre QDV. La QDV d'un sujet déficient ou atteint d'une maladie chronique ne doit pas être évaluée par un individu valide [16]. La projection de l'évaluateur de ses propres repères d'une vie de qualité fausse son appréciation de la QDV d'un sujet déficient ou malade. Si pour un sujet valide, le fait de pouvoir marcher est souvent inhérent à une vie de qualité, le raisonnement d'un sujet déficient est souvent bien différent.

Concernant l'évaluation de la QDV, l'ensemble du personnel médical et paramédical sous-estime la QDV du patient par rapport à sa propre évaluation [16] [17]. Dans le cadre des maladies neuromusculaires, pour Abresch et al. [18], la ventilation chez le patient porteur

d'une dystrophie musculaire de Duchenne n'a pu se développer que lorsque l'on a commencé à s'intéresser à l'avis du patient pour orienter les choix thérapeutiques. Les médecins jugeant la QDV des patients atteints d'une dystrophie musculaire comme très affectée, avaient tendance à ne pas proposer de traitement qui aurait pu prolonger leur vie. Gibson [19] rapportait aussi qu'une des raisons pour lesquelles les médecins ne recommandaient pas l'assistance ventilatoire aux patients atteints de dystrophie musculaire étaient qu'ils considéraient leur QDV comme mauvaise.

Chez l'enfant, la validité et la reproductibilité de l'évaluation de sa QDV par lui-même ont longtemps été remises en cause ; les parents et les soignants étaient considérés comme plus aptes à décrire la QDV de l'enfant [20] [21] [22]. Cependant, le jugement des adultes à propos des émotions de l'enfant est fondé sur leurs interprétations du comportement de l'enfant, ce qui constitue un biais supplémentaire à cette mesure. De plus, un des risques de l'hétéro-évaluation chez l'enfant est de ne pas prendre en considération des éléments qui sont pertinents pour l'enfant mais méconnus ou sous-estimés par son entourage. Alors que l'adulte perçoit souvent l'hospitalisation comme un événement de santé négatif, certains enfants la considèrent comme une marque de courage, liée à un sentiment d'estime de soi. Il est certain que les professionnels de santé et les parents peuvent fournir une évaluation intéressante du fonctionnement psychologique et physique d'un enfant. Mais nous devons également nous interroger sur la valeur du jugement d'un adulte au sujet du ressenti de l'enfant et sur ses propres perceptions du bien-être. Finalement, le répondant idéal, tout comme chez l'adulte, devrait être l'enfant. A partir de 8 ans, il semble que l'autoévaluation soit fiable, encore faut-il utiliser un questionnaire validé et adapté à cette population.

Mais parfois du fait du très jeune âge, de l'altération des capacités cognitives, ou des troubles du comportement, cette autoévaluation est impossible. L'hétéro-évaluation demeure la seule alternative si l'on veut mesurer la QDV des patients.

4.3. Questionnaire d'application facile et adapté à la population d'étude

Le questionnaire doit être un outil pragmatique et facilement applicable dans la population étudiée. En cas de difficulté de concentration, en particulier chez l'enfant, il ne doit pas être trop long. Les items du questionnaire doivent être facilement compréhensibles par les patients, en évitant par exemple les doubles négations. Le questionnement doit être orienté en fonction de la population étudiée sur des domaines importants de la QDV, domaines variables en fonction de l'âge, du sexe, de l'affection du patient. Par exemple chez l'adulte, un questionnement sur les ressources financières peut-être important [23]. Chez l'adolescent, le questionnaire doit contenir des items sur le corps et son acceptation, sur la relation avec les parents qui constituent des éléments pivots de la QDV à cet âge. Dans une population à motricité déficiente, il faudra veiller à l'absence d'items en lien direct avec la fonction motrice.

Enfin, il ne faudra pas sous-estimer les possibles sentiments de dévalorisation engendrés par les questionnaires de QDV chez le patient. Dans l'idéal, les questionnaires devront contenir à la fois des items négatifs comme« avez vous été stressé(e) ; déprimé(e) ; facilement découragé(e)….. » et des items « positifs » de la QDV comme « êtes vous content(e) ; satisfait(e) de votre vie ; entouré(e) par vos amis …. ». Ainsi les patients ne sont pas interrogés exclusivement sur ce qui ne va pas, mais aussi sur ce qui leur apporte du bien-être.

Les critères pour juger l'acceptabilité du questionnaire dans la population sont le temps de réponse au questionnaire, le nombre de refus de participer, la compliance (nombres d'items renseignés sur la totalité des items).

4.4. Un questionnaire possédant de bonnes qualités métrologiques.

Un bon outil de mesure doit satisfaire certaines qualités métrologiques pour être considéré comme pertinent : sa sensibilité au changement, sa fidélité et sa validité. Ces différentes propriétés sont analysées lors de l'étude de validation d'un questionnaire de QDV.

Ces études de validation sont longues et coûteuses mais elles constituent un passage obligé avant l'utilisation d'un outil dans un essai clinique.

Les questionnaires sont l'objet d'un copyright, ils ne peuvent donc être modifiés ou traduits sans l'autorisation de l'auteur. Dans le cas d'un outil traduit dans une autre langue, il faut d'abord vérifier que la traduction a été faite selon une méthodologie rigoureuse : la validation linguistique. Ensuite vérifier que la traduction a conservé les qualités métrologiques satisfaisantes de l'outil.

4.4.1. La fidélité

La fidélité (reliability) est la capacité du questionnaire à se comporter de manière fiable et donc de mesurer de manière reproductible la QDV. La fidélité intrajuge ou la reproductibilité (reproductibility) indique la cohérence des mesures. Elle est définie par la stabilité du résultat au cours des mesures répétées, l'état du sujet restant stable pendant ce laps de temps. Elle est à rattacher à la notion de précision de la mesure. La fidélité interjuge est étudiée en effectuant au même moment par deux observateurs une appréciation indépendante. Dans le cas de mesure qualitative comme la QDV, la fidélité des différentes mesures sera évaluée par le coefficient de concordance Kappa (qui s'applique aux jugements qualitatifs).

4.4.2. La cohérence interne

La cohérence interne (internal consistency) : le coefficient alpha de Crombach évalue la cohérence interne d'un ensemble d'items correspondant à une dimension clinique ; c'est-à-dire la force des inter-corrélations entre items d'une même dimension. Les items de chaque

dimension doivent former un tout cohérent. Plus les items sont liés entre eux, plus le coefficient alpha est proche de 1. En pratique, la cohérence interne doit être assez grande (alpha > 0,70-0,80) mais le coefficient ne doit pas être trop proche de 1 car cela signifie que plusieurs items sont redondants.

4.4.3. La validité

La validité ou pertinence d'un outil est sa capacité à bien mesurer ce qu'il est censé mesurer.

La validité présente plusieurs facettes. La **validité d'apparence** (face validity) représente le jugement subjectif (fonction de l'utilisateur) prenant en compte les aspects visibles de l'échelle de façon superficielle : la longueur du questionnaire, le libellé des items, les modalités de réponse… La **validité de contenu** (content validity) encore appelée **spécificité** concerne la pertinence du contenu de l'outil établie par un jugement d'experts. Elle juge si les questions sélectionnées représentent bien toutes les facettes du concept à mesurer. La sélection des items retenus pour composer l'outil doit avoir été réalisée par une équipe composée d'experts médicaux, de malades et de psychologues. La **validité contre critère** (criterion validity) représente la mesure d'intensité de lien statistique existant entre la mesure effectuée par l'échelle étudiée et la mesure effectuée par une échelle existante considérée comme la référence. Enfin, la **validité du construit** (construct validity) s'affirme au fur et à mesure que des expériences successives confirment les hypothèses théoriques émises lors de la conception de l'échelle.

4.4.4. La sensibilité au changement

Un instrument est dit sensible au changement s'il est capable de mesurer avec précision les variations en plus ou en moins du phénomène mesuré. Il doit permettre un classement suffisamment fin des individus et être capable de repérer des variations cliniquement perceptibles.

Un outil est sensible s'il mesure le phénomène étudié avec une finesse suffisante pour permettre de distinguer les individus ou des groupes d'individus. La sensibilité au changement est importante puisqu'elle permet d'évaluer l'évolution de la maladie avec les effets éventuels des thérapeutiques.

4.4.5. Score global ou profil de QDV

Seule l'homogénéité des questions autorisera l'établissement d'un score global ou index de QDV. Si les différentes dimensions ne sont pas homogènes, on obtiendra un score par dimension et un profil de QDV plus qu'un index global. Certains questionnaires comme le CHQ (Child Health Questionnaire) [24] par exemple ne fourniront qu'un profil de QDV alors que d'autres comme le VSP-A (Vécu Santé Perçu par l'Adolescent) [25] autoriseront l'établissement d'un index global de QDV sur 100.

5. Les biais d'interprétation des résultats

La QDV est un critère de jugement qui peut être utilisé dans tous les types d'étude : descriptive, cas-témoins, cohorte, essai thérapeutique….

Dans le cas d'une étude descriptive, la QDV des patients est mesurée à un instant t comme une mesure supplémentaire permettant de mieux définir la population. La question est de savoir si dans cette population spécifique du fait d'une affection aigüe, d'un traitement au long court ou d'une affection chronique, la QDV des patients est significativement différente de la population générale. Il ne s'agit pas d'interpréter les mesures de QDV en fonction de leur valeur, car il n'existe aucune norme de QDV. Pour l'interprétation d'une mesure de QDV dans une population donnée, nous devons disposer de résultats dans une population dite de référence.

Dans le cas d'une étude cas-témoin, l'interprétation est plus aisée car la comparaison est faite avec la population témoin.

Enfin, dans les études de cohortes et les essais thérapeutiques, intervient la notion de sensibilité au changement car plusieurs mesures de QDV sont réalisées à différents temps d'intervalle. Les questions posées peuvent porter sur l'évolution de la QDV de patients atteints d'une maladie chronique au cours du temps et sur l'impact de la thérapeutique sur cette évolution.

5.1 Biais d'interprétation en lien avec le concept même de QDV

Il n'existe aucune norme de QDV. Les questionnaires permettent dans la majorité des cas de calculer un score par dimension, voire un score global si l'homogénéité du questionnaire le permet. Ces scores sont souvent transformés linéairement en une échelle allant de 0 à 100, 100 indiquant la QDV « la plus favorable » et 0 la QDV « la moins favorable ». Comment interpréter un score de 85 par exemple ? Pouvons nous dire par exemple que le patient a une bonne QDV puisque qu'il se situe dans le quart maximum des résultats mais ne pourrions nous pas aussi dire que sa QDV n'est pas parfaite puisqu'elle n'atteint pas 100% ? Ce type d'interprétation n'a pas lieu d'être, car pour interpréter un score de QDV il faut disposer de données sur une population de référence appariée au moins sur l'âge et le sexe. Par exemple si on évalue la QDV chez les enfants, la période clé de l'adolescence est d'interprétation délicate. En effet, il a pu être montré chez les adolescents sains que, si leur évaluation de QDV était mauvaise, c'est finalement qu'ils allaient plutôt bien en s'inscrivant dans une attitude d'opposition propre à cet âge [4]. Sans la référence à une population d'enfants sains l'interprétation des résultats aurait été erronée.

Dans le cas d'une étude de cohorte ou d'un essai thérapeutique, les biais sont encore plus importants. Si entre deux mesures de QDV encadrant une intervention thérapeutique par exemple, on mesure une augmentation significative de la QDV, le lien de causalité sera évoqué mais difficile à affirmer. Car que s'est-il passé dans la vie de ces patients dans cet

intervalle, ils se sont mariés, ont trouvé un travail qui les épanouissait, ou peut-être ont-ils perdu une personne proche, autant de facteurs difficiles à contrôler dans un essai thérapeutique hormis par un entretien individualisé avec une psychologue. Il ne s'agit pas par ce propos de remettre en cause l'utilité de questionnaire de QDV en recherche clinique, mais seulement de relativiser certaines conclusions et de souligner la prudence avec laquelle il faut interpréter les résultats. Par exemple, dans une étude sur l'impact de la mise en place de la ventilation non invasive chez des patients porteurs d'une dystrophie musculaire, Young et al. [26] concluent à l'effet positif de la ventilation sur la QDV des patients car la QDV demeure stable après mise en place de la VNI, alors que les auteurs postulaient que la QDV devait se dégrader avec l'évolution de la pathologie. Voici un exemple de conclusion très discutable, car rien ne prouvait qu'en l'absence de cette ventilation la QDV de ces patients se serait dégradée [27].

5.2. Biais en lien avec la mesure de QDV

5.2.1. Défaut de sensibilité au changement.

L'utilisation d'une mesure de QDV qui n'est pas sensible au changement peut expliquer l'absence de significativité d'un test de comparaison. Conclure à l'inefficacité thérapeutique serait alors une erreur. Et ceci d'autant plus si la QDV fait partie des critères d'évaluation secondaires dont la progression attendue grâce au traitement n'a pas été prise en compte dans le calcul de la taille d'échantillon.

5.2.2. Concept du questionnaire non adapté à la question posée.

L'intérêt de l'étude peut être remis en cause selon le choix d'une échelle générique ou d'une échelle spécifique ou d'une échelle de QDV liée à la santé ; par exemple quel serait l'intérêt

d'utiliser une échelle de QDV dans laquelle on tiendrait compte des capacités fonctionnelles de déplacement dans l'évaluation d'une population de patients ayant un handicap moteur ?
Si cette étude comportait une comparaison à une population générale on en connaîtrait le résultat d'emblée. Si cette étude testait l'efficacité d'une thérapeutique sur la fonction, n'aurait il pas été plus pertinent d'utiliser une échelle de mesure de la fonction motrice ?

Conclusion

La notion de QDV apparaît comme un concept multidimensionnel, aux définitions variables en fonction de l'évaluateur et des objectifs de l'évaluation. La QDV d'une personne ne peut être assimilée à l'absence de symptôme objectif. Il apparaît que la meilleure évaluation de la QDV en recherche clinique est celle utilisant un questionnaire :

- pour lequel le concept de QDV a été parfaitement défini,
- qui possède de bonnes qualités métrologiques,
- qui s'adresse directement au patient permettant de ne pas passer à coté d'éléments méconnus ou sous-estimés par l'entourage.

Cependant, les profils ou index de QDV doivent être interprétés avec beaucoup de prudence, le bien-être ressenti par une personne tout au long de son existence ne peut être réduit à quelques données chiffrées, si précises et valides soient-elles. Une telle évaluation ne peut se concevoir sans un entretien avec le patient permettant de cerner son histoire et son environnement familial et social afin d'appréhender des domaines inaccessibles par un questionnement standardisé.

Références

[1] Bénony H. Mesure de la qualité de vie. Ann readapt Med Phys 2001;44 Suppl 1:72-84

[2] Fallissard B. Valeur scientifique des mesures de qualité de vie et autres mesures subjectives réalisées en recherche clinique ? E-mémoires de l'Académie Nationale de Chirurgie 2004;3(1):19-23

[3] Fallissard B. Peut-on mesurer la qualité de vie ? In: Groupe d'Etude et de Recherche sur l'Infirmité Motrice d'Origine Cérébrale(éd). « La Qualité de Vie », 15 et 16 mai 2006, Lyon

[4] Duverger P. Qualité de vie chez l'enfant malade chronique-le cas de la transplantation rénale. In: Groupe d'Etude et de Recherche sur l'Infirmité Motrice d'Origine Cérébrale(éd). « La Qualité de Vie », 15 et 16 mai 2006, Lyon

[5] Renwick R, Brown I. Quality of Life in Health Promotion and Rehabilitation. In: Brown I. (ed), The Center for Health Promotion's Conceptual Approach to Quality of Life: Being, becoming, and belonging. Thousand Oaks (CA): Sage Publications, 1996. 75-86

[6] Curran D. Analysis of longitudinal quality of life data with dropout (1998). In: "Three Country Corner" RSS Local Group meeting, The Janssen Research Foundation, 1er Avril 1998, Beerse, Belgique

[7] Chow SM, Lo SK, Cummins RA. Self-perceived quality of life of children and adolescents with physical disabilities in Hong Kong. Qual Life Res 2005 Mar;14(2):415-23

[8] Bach JR, Barnett V. Ethical considerations in the management of individuals with severe neuromuscular disorders. *Am J Phys Med Rehabil.* 1994;73(2):134-140.

[9] Bach JR, Vega J, Major J, Friedman A. Spinal muscular atrophy type 1 quality of life. Am J Phys Med Rehabil 2003;82:137-42

[10] The WHOQOL Group. The development of the World Health Organization quality of life assessment instrument (the WHOQOL). In: Orley J, Kuyken W (Eds), Quality of Life Assessment: International Perspectives. Heidelberg: Springer Verlag, 1994

[11] Drotar D. Validating measures of paediatric health status, functional status, and health-related quality of life: key methodological challenges and strategies. Amb Paediatr 2004;4:358-64

[12] Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press, 1992

[13] Leplège A. Les mesures de la qualité de vie. Paris: PUF, 1999. Que sais-je;3506

[14] Bergson H. Essai sur les données immédiates de la conscience. Paris Alcan, 1889

[15] Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993 Apr 15;118(8):622-9

[16] Bach JR, Barnett V. Ethical considerations in the management of individuals with severe neuromuscular disorders. Am J Phys Med Rehabil 1994 Apr;73(2):134-40

[17] Bach JR, Campagnolo DI, Hoeman S. Life satisfaction of individuals with Duchenne muscular dystrophy using long-term mechanical ventilatory support. Am J Phys Med Rehabil 1991;70:129-35

[18] Abresch RT, Seyden NK, Wineinger MA. Quality of life. Issues for persons with neuromuscular diseases. Phys Med Rehabil Clin N Am 1998;9:233-48

[19] Gibson B. Long-term ventilation for patients with Duchenne muscular dystrophy. Physicians' beliefs and practices. Chest 2001;119:940-6

[20] Achenbach T M, McConaughy SH, Howell CT. Child/adolescent behavioural and emotional problems: implications of cross-informant correlations for situational specificity. Psychol Bull 1987;101:213-32

[21] Herjanic B, Herjanic M, Brown F, Wheatt T. Are children reliable reporters? J Abnorm Child Psychol 1975;3(1):41-8

[22] Seaberg, JR. Child well-being scales: A critique. Soc Work Res Abstr 1988;24: 9-15

[23] Borgel F, Mémin B, Perret J. Réadaptation et concept de qualité de vie : critique des outils de mesure du subjectif. Ann Readapt Med Phys 1991;34:75-80

[24] Landgraf J, Abetz NL. Measuring health outcomes in paediatric population: issues in psychometrics and application. In: Spilker B (Ed), Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven, 1996. 793-802

[25] Simeoni MC, Auquier P, Antoniotti S, Sapin C, San Marco JL. Validation of a French health-related quality of life instrument for adolescents: the VSP-A. Qual Life Res 2000;9(4):393-403

[26] Young HK, Lowe A, Fitzgerald DA, et al. Outcome of noninvasive ventilation in children with neuromuscular disease. Neurology 200768:198–201

[27] Vuillerot C, Hodgkinson I, Bissery A, Schott-Pethelaz AM, Iwaz J, Ecochard R, D'Anjou MC–, Commare MC, Berard C .Self-Perception of Quality of Life by Adolescents with Neuromuscular Diseases. Journal of adolescent health 2010;46:70-76

[28] Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. The sickness impact profile: validation of a health status measure. Med Care. 1976 Jan;14(1):57-67.

CHAPITRE XVI

NOTIONS DE STATISTIQUES POUR LE CLINICIEN

Pierre Duhaut, Claire Andréjak

Les cliniciens sont souvent réfractaires aux statistiques. Or, elles sont indispensables pour arriver à observer tous les événements non forcément visibles 'à l'œil nu', et arriver à différencier la fréquence des événements auxquels on s'intéresse, par rapport à une survenue aléatoire. Les statistiques sont inutiles pour mettre en évidence l'action des antibiotiques sur la méningite à méningocoque, ou sur la tuberculose. Elles seront nécessaires pour mesurer l'incidence des accidents vasculaires cérébraux, pathologies graves s'il en est, sous anti-hypertenseurs par comparaison à des patients non traités. La différence entre les deux situations est que dans la première, on s'intéresse à une action rapide et individuelle, presque immédiatement visible. Dans la seconde, on s'intéresse avant tout à une tendance sur un groupe, qui se traduira par une amélioration à long terme du pronostic d'une partie seulement de la population traitée : le domaine des statistiques est celui de l'évaluation de tendances perceptibles uniquement au niveau d'échantillons ou de populations, mais pouvant bénéficier également, sur le plan clinique, à certains des patients traités ou observés à défaut de tous.
Ce chapitre passe en revue les questions cliniques les plus fréquentes et leur traduction en langage et en tests statistiques.

Biostatistiques et méthodologie sont indissociables : les données recueillies doivent être analysées… et doivent donc être recueillies de telle sorte à être analysables. La construction d'une étude doit prévoir le type d'analyse à faire en fonction de la question posée, et la façon de colliger les données doit obéir à deux impératifs parfois contradictoires, qu'il faudra concilier :

- Décrire la réalité au plus près,
- Rendre ces données compatibles avec un test statistique existant permettant la comparaison la plus exacte entre les groupes étudiés.

I. Questions posées, variables à colliger, analyse à faire

Partons d'une question clinique banale et d'une base de données effective pour rendre le propos plus intelligible :
Les thrombocytoses sont fréquemment rencontrées en médecine clinique. Un vieil aphorisme affirme qu'une thrombocytose supérieure à $1.10^6/mm^3$ correspond 'à tous les coups' à une thrombocytémie essentielle (TE) (syndrome myéloprolifératif). Le diagnostic étiologique peut cependant être difficile à poser, y compris par la biologie moléculaire (la mutation JAK2 n'est présente que dans 60 % des cas environ), et la biopsie de moelle peut ne pas montrer de myélofibrose évocatrice. Un 'gold standard' pourrait être le diagnostic porté après évolution du patient et élimination -si possible- de toutes les causes réactionnelles.
L'ensemble des thrombocytose supérieures à $600\,000/mm^3$ d'un hôpital ont été réunies dans une série. Cette base de données doit d'abord être décrite, et les facteurs prédictifs de diagnostics peuvent ensuite être analysés.

II Variables dichotomiques (proportions) :

Mille quarante sept patients ont été inclus : il faut connaître la distribution des sexes, le nombre de patients avec thrombocytose réactionnelle (TRe) ou essentielle (en définissant chaque classe comme exclusive de l'autre). Ces variables sont dites **dichotomiques** car elles ne peuvent prendre que deux valeurs (masculin-féminin, essentielle-réactionnelle, vrai-faux, oui-non). Elles sont également **catégorielles,** sans hiérarchie entre les deux valeurs. Elles permettent de calculer la **proportion** de femmes et d'hommes atteints par l'une, ou l'autre, des grandes causes de thrombocytose.

Application :
Notre série comprend 509 femmes et 538 hommes. Le diagnostic de TRe est posé chez 357 femmes (70,1 % de 509) et 461 hommes (85,7 % de 538). Ce taux est-il *significativement différent* entre les hommes et les femmes ? Autrement dit, 85,7 % - 70,1 % est-il égal à (ou proche de) 0 ?
Nous venons de poser *l'hypothèse nulle* : on conclura, si la différence entre les deux taux n'est pas significativement différente de 0, que les deux taux sont proches l'un de l'autre, similaires, et qu'il n'y a pas de différence entre patients de sexe masculin et féminin. Au contraire, si la différence est significativement différente de 0, les deux proportions, et par conséquent les deux groupes, seront estimés différents.
Les données peuvent s'exprimer dans une table 2×2 :

Table 1 : nombre de patients effectivement observés dans chaque catégorie

	Femmes	Hommes	Total
TE	152 *(29,9% de 509)*	*77 (14,3% de 538)*	229 *(21,9% de 1047)*
TRe	*357 (70,1% de 509)*	461 *(85,7% de 538)*	818 *(78,1% de 1047)*
Total	509 *(48,6% de 1047)*	538 *(51,4% de 1047)*	1047

1. Test de chi-2

Le **test de chi-2** permet la comparaison de proportions. Comme tout test statistique, son principe repose sur l'hypothèse nulle : s'il n'y a pas de différence entre hommes et femmes, alors hommes et femmes ne composent qu'un seul groupe dans lequel 229 patients (21,9 %) présentent une thrombocytémie essentielle, et 818 (78,1 %) une thrombocytose réactionnelle. Le test de chi-2 va mesurer, pour chaque cellule, le nombre de patients d'écart entre les patients *observés* ('O'tels qu'ils sont donnés par l'étude effectuée) et *attendus ('A')* si les proportions dans l'ensemble du groupe, et dans chaque sous-groupe, étaient égales. La table devient donc :

Table 2 : nombre de patients attendus dans chaque cellule)

	Femmes	Hommes	Total
TE	509 x 21,9 % (proportion attendue de TE dans les 1047 patients) = 111,5 (nombre Attendu = A)	538 x 21,9 % (proportion attendue de TE dans les 1047 patients) 117,5 (nombre Attendu = A)	229
TRe	509 x 78,1 % (proportion attendue de TRe dans les 1047 patients) = 397,5 (nombre Attendu = A)	538 x 78,1 % (proportion attendue de TRe dans les 1047 patients) = 420,5 (nombre Attendu = A)	818
Total	509	538	1047

Cette nouvelle table présente plusieurs caractéristiques notables par rapport à la précédente :

- Les totaux n'ont pas changé.
- Lorsque le nombre de patients attendus dans la cellule 1 est fixé, les nombres attendus dans les cellules 2, 3 et 4 sont déterminés sans liberté aucune : la somme des lignes, et la somme des colonnes, doit rester constante (car fixées par le nombre de patients effectivement observés dans chaque ligne et chaque colonne !).
- Donc, la détermination du nombre de patients attendus n'a pu se faire qu'avec *un seul degré de liberté,* celui de la cellule 1.

La somme des écarts entre observés et attendus pourrait s'écrire : **Σ (O - A)**.
On remarquera que cette somme est égale à 0 : effectivement, les patients retirés à une cellule ont été ajoutés à la cellule voisine : l'écart de la première cellule est l'opposé exact de la deuxième, celui de la troisième, l'opposé de la quatrième.

Pour contrer cet inconvénient, chaque écart est élevé au carré. La somme devient donc : **$\Sigma (O - A)^2$**.
Un écart absolu ne peut cependant à lui seul, rendre compte d'une différence : l'écart entre 10000 et 10002 est égal à celui entre 2 et 4. L'augmentation est de 1/5000 dans le premier cas, et de 50 % dans le second... il faut donc rapporter l'écart à un dénominateur. Le calcul du chi-2 le ramène au nombre d'événements attendus, et la formule du chi-2 devient :

$$\text{chi-2} = \frac{\Sigma (O - A)^2}{A}$$

Dans notre exemple,

$$\text{chi-2} = \frac{\Sigma (O - A)^2}{A} = \frac{(152-111,5)^2}{111,5} + \frac{(77-117,5)^2}{117,5} + \frac{(357-397,5)^2}{397,5} + \frac{(461-420,5)^2}{420,5} = 37$$

Il est ensuite possible de consulter les tables de distribution du chi-2 *pour un degré de liberté* et d'estimer quelle est la probabilité pour que 37 soit similaire à 0. Cette probabilité est largement inférieure à 1/1000 (en fait, inférieure à 10^{-7} !)

Le seuil de probabilité pour admettre l'hypothèse nulle est habituellement fixé, par convention, à 5 %. L'hypothèse nulle est acceptée au-dessus du seuil, rejetée en-dessous. L'hypothèse nulle dans notre exemple ayant moins d'une chance sur 10 millions d'être vraie, ($p < 10^{-7}$ *, < 0,05)* est rejetée. Les deux groupes n'étant pas similaires, et on en déduit qu'ils sont *probablement* différents. Il y a là une petite contorsion de logique, car toute la base du calcul repose sur l'axiome de l'égalité des deux groupes, et ça n'est que sur la base de cet axiome que la méthode de calcul est appropriée : or, la base de validité du calcul n'est pas respectée… Nous avons simplement montré *que les deux groupes n'étaient probablement pas similaires,* et en *déduisons qu'ils sont probablement différents*.
Les statistiques n'établissent pas de vérité : elles cherchent à circonscrire une incertitude, et ceci doit toujours rester présent à l'esprit lors de l'interprétation des résultats. Cependant, lorsque la valeur de p est aussi faible, on peut très raisonnablement penser que les deux groupes sont différents. Lorsque la valeur de p est proche de 0,05 et que changer quelques patients de groupe la fait passer au-dessus ou au-dessous de 0,05, la discussion reste ouverte !

Attention :

- Le calcul simple du chi-2 décrit ci-dessus n'est valable que si le nombre d'événements attendus est supérieur à 5 dans toutes les cellules de la table. Dans le cas contraire, un calcul exact faisant appel aux distributions géométriques doit être réalisé : c'est le *test de Fisher exact*, qui peut toujours être employé dans les cas douteux pour donner au calcul un maximum de rigueur.
- Plusieurs auteurs ont voulu rendre le calcul plus sévère (donc, à rapprocher la somme du chi-2 de 0), en introduisant des facteurs de correction dans la formule. *C'est le cas du chi-2 de Yates, qui soustrait ½ à chaque (A-O) avant de l'élever au carré.*
- Le test de Mantel-Haenszel est la variante du chi-2 pour les données stratifiées : nous l'aurions appliqué dans notre exemple si les diagnostics avaient été donnés par sexe et par tranche d'âge (20-40 ans, 40-60, 60-80, > 80). Les chi-2 sont alors calculés pour chaque strate séparément, puis additionnées sur l'ensemble des strates.
- On comprend aisément que la valeur du chi-2, et donc la significativité du test, dépend de la taille de la population analysée : refaites le calcul en divisant les nombres de chaque cellule par 10 pour atteindre une taille d'échantillon totale de 105 (fréquente dans les études cliniques) (proportions conservées), et vérifiez la valeur de p !

2. Généralisation du test de chi-2 à une table à n lignes et m colonnes :

La somme des écarts entre O et A peut toujours être calculée. Simplement, le nombre de degrés de liberté change, et devient égal à (n-1)(m-1) : le nombre A est toujours fixé sans liberté pour la dernière cellule de chaque ligne, et la dernière cellule de chaque colonne. La probabilité d'égalité entre la somme du chi-2 et 0 doit alors se lire sur la ligne de distribution du chi-2 au nombre de degrés de liberté correspondant (les logiciels fournissent la valeur exactement calculée de p).

Un chi-2 significativement différent de 0 dans une table n*m n'indiquera pas où se trouve la différence : elle peut se répartir de façon homogène entre les différentes cellules, ou, de façon beaucoup plus fréquente, entre quelques, voire deux, cellules de la table. Le risque d'atteindre un nombre d'événements attendus inférieur à 5 dans une table à multiples cellules augmente avec le nombre de cellules, ce qui rend la validité du calcul incertaine : mieux vaut alors prévoir un autre plan d'analyse.

III. Variables quantitatives

1. Moyenne, variance, écart-type, médiane, percentiles, mode.

Nous voulons maintenant décrire les âges des patients dans chaque groupe, et les comparer. L'âge est une donnée *quantitative continue*. Sa description peut être faite sous la forme de ***moyenne*** (somme de toutes les valeurs divisée par le nombre de sujets), et de ***variance*** (somme de l'ensemble des carrés des écarts entre la moyenne de l'échantillon et l'âge de chaque sujet, rapportée au nombre de sujets : comme pour le chi-2, les écarts sont élevés au carré de telle sorte à ne pas s'annuler).

La variance peut donc s'écrire : $V = \Sigma \frac{(\mu - a)^2}{n}$,

où μ désigne la moyenne des âges pour l'ensemble du groupe, a l'âge de chaque sujet du groupe, et n le nombre total de personnes dans le groupe.

Cette formule exprime la **variance idéale**, d'une grande population. Nous travaillons le plus souvent en médecine sur des échantillons de patients bien plus réduits, même dans les études multicentriques.

Un échantillon ne peut fournir qu'une *estimation, qu'une valeur approchée,* de la moyenne et de la variance de la population totale. Par sécurité, il vaudra mieux sur un échantillon définir cette variance de façon un peu plus large (elle aura ainsi plus de chances de recouvrir la variance de la population totale). Pour ce faire, on corrige le dénominateur en remplaçant n par n-1: la valeur numérique de la variance augmente légèrement, *et sa formule pour un échantillon* devient : $V = \frac{\Sigma (\mu - a)^2}{n - 1}$

Cette correction est d'autant plus importante que n, la taille d'échantillon, est petit, et d'autant plus insignifiante que n est grand : la variance d'une variable sur un échantillon de 1000 personnes a plus de chances d'être proche de celle de la population totale, que la variance de la même variable sur un échantillon de 15 personnes.

De même, on souhaite approcher la moyenne de la population globale à partir de celle de l'échantillon, en sachant cependant que si l'échantillon avait été sélectionné différemment, sa moyenne serait sans doute un peu différente : il se peut que la moyenne d'hémoglobine glyquée d'un groupe de 40 patients diabétiques de type 2 soit *exactement la même* que celle d'un autre groupe de 40 patients. Il est cependant probable qu'elle n'en soit pas trop éloignée. On définit la notion d'*intervalle de confiance à 95 %* pour exprimer le fait que la moyenne se trouverait comprise dans cet intervalle chez 95 % des 100 échantillons potentiels de taille égale que l'on pourrait tirer au sort au sein de la population globale. Autrement dit, la moyenne du taux d'hémoglobine glyquée dans la population globale de diabétiques de type 2 - que l'on cherche à approcher- a sans doute 95 % de chances d'être comprise dans l'intervalle ainsi défini à partir de l'échantillon. On peut définir, de la même façon, un intervalle de confiance à 90 ou 99 ou 99,9 %... en fonction de la précision que l'on souhaite obtenir. L'intervalle de confiance à 95 % est celui le plus souvent retenu en médecine.

La variance est très souvent exprimée sous la forme de sa racine carrée, appelée ***écart-type, ou déviation standard.***

Ecart-type = $\sqrt{v} = \sigma$

L'intérêt de l'écart-type est qu'il permet d'apprécier assez rapidement la distribution de la variable : pour une taille d'échantillon supérieure à 30, 95 % de l'échantillon sera compris entre la moyenne $\pm$ 1,96σ. La valeur correspondant à 1,96 augmente lorsque la taille d'échantillon diminue : là encore, cette augmentation traduit le fait que la précision diminue avec la taille d'échantillon, et donc que la dispersion probable des valeurs, augmente.

Attention :

- Cette façon de décrire une variable quantitative n'est valable que lorsque la distribution des valeurs suit une courbe gaussienne (distribution normale).
- La moyenne n'a de sens que si
    - les valeurs se distribuent de façon symétrique autour d'elle, et si
    - elle correspond à la valeur la plus fréquemment rencontrée (dans notre exemple, la tranche d'âge la plus importante). En effet, d'établir une moyenne d'âge à 40 ans pour une série comportant 20 enfants de 10 ans et 20 adultes de 70 ans ne permettrait pas d'appréhender correctement la réalité du groupe de patients : il n'y a aucun adulte de 40 ans dans ce groupe composé de deux sous-groupes très différents, et calculer une moyenne arithmétique à 40 ans amènerait à une description fausse de la réalité.
- Pour que la moyenne et son écart-type décrivent correctement le groupe considéré, il faut donc que la distribution de la variable examinée soit *symétrique et unimodale*,

autrement dit, que la distribution des valeurs ne reflète pas l'existence de deux groupes différents de patients, d'un taux biologique, ou d'une valeur numérique quelconque. La maladie de Hodgkin a deux pics d'incidence, autour de 20-25 ans, puis autour de 60 à 65 ans. De dire que la moyenne d'âge des patients est de 40 ans et traiter les patients âgés de 75 ans de la même façon que ceux âgés de 20 ans sous prétexte d'une tolérance moyenne égale à celle des 40 ans serait un non-sens médical.

- Il est toujours utile de réaliser un graphe de la distribution des données qui permette d'en apprécier la forme (symétrique, unimodale). La plupart des logiciels statistiques permettent également de réaliser un test de normalité de la distribution, qui pourra guider l'analyse statistique ultérieure.

Lorsque la distribution de la variable quantitative n'est pas *normale (gaussienne)*, il faut privilégier d'autres modes de description de données et d'autres modes d'analyse. Il arrive assez souvent en médecine, que des variables quantitatives aient une distribution gaussienne (*statistiquement normale = gaussienne, à différencier de biologiquement normale = dans les normes biologiques)* chez le sujet sain (exemple : le taux d'hémoglobine, des leucocytes, des plaquettes...). Cette normalité statistique disparaît très souvent chez le sujet malade : les leucocytes peuvent varier de 10 000/mm^3 à plus de 100 000 dans une leucémie myéloïde chronique ou une leucémie aiguë, les plaquettes de 400 000 à plus de 1 500 000 dans une thrombocytémie essentielle ou réactionnelle, et l'on ne peut pas en général extrapoler la distribution *normale* de la variable chez le sujet *sain* à la distribution dissymétrique, parfois logarithmique, parfois difficile à décrire, de la variable chez le sujet *malade.*
Si la distribution n'est pas statistiquement normale à l'appréciation visuelle ou sur le test de normalité, recourir à la *médiane et aux percentiles* donnera une idée plus précise de la population considérée : la médiane définit le seuil en valeur absolue, en-dessous duquel se trouve 50% de l'échantillon et au-dessus duquel, se trouvent les 50% restant. Une leucocytose médiane à 20 000/mm^3 signifie que 50% des patients ont un taux de leucocytes inférieur à 20 000, et 50% d'entre eux, supérieur. Les 5ème, 10ème, ou 75ème percentiles correspondent au taux de leucocytes *en-dessous desquels* se trouvent 5%, 10%, ou 75% des patients.

Le *mode,* enfin, est le troisième descriptif d'une variable quantitative : il correspond à la valeur la plus souvent rencontrée dans l'échantillon. Il est relativement peu utilisé.
Dans une distribution *statistiquement normale,* moyenne, médiane et mode sont confondus. Dans une distribution non normale, ils sont habituellement distincts. Une distribution à plusieurs pics de hauteur égale peut comprendre plusieurs modes, mais ne comprendra qu'une seule médiane... et qu'une seule moyenne, non représentative.

Application :
Dans notre exemple, l'étude de la distribution des âges donne les résultats suivants :

	Moyenne	Ecart-type	Médiane	Extrêmes
Femmes	62,8	19,8	67	18,7-102
Hommes	55,5	17,5	55,5	18,5-96,5

A première vue, les valeurs données peuvent être compatibles avec une distribution normale : retrancher ou additionner deux écarts-types à la moyenne ne conduit pas à des âges aberrants (par exemple : négatifs), mais à des âges relativement voisins des extrêmes.

A deuxième vue, si médiane et moyenne sont confondues chez les hommes, elles s'écartent de plus de 4 ans chez les femmes, ce qui fait suspecter compte tenu de la taille d'échantillon, une distribution non normale. La question est de savoir si cet écart est significatif, ou non.

2. Comparaison de moyennes, ou analyse de variance :
Les hommes et les femmes de notre étude ont-ils le même âge ?
Le principe reste le même que pour le test du chi-2 : l'hypothèse nulle, simplement, devient :

Les deux moyennes sont semblables, ou $\mu_f - \mu_h = 0$ (μ_f représentant la moyenne d'âge des femmes, et μ_h la moyenne d'âge des hommes).
Cette simple soustraction cependant ne suffit pas à assurer une comparaison correcte : on serait prêt à reconnaître que les deux moyennes sont différentes si l'écart-type était très petit :

Par exemple :

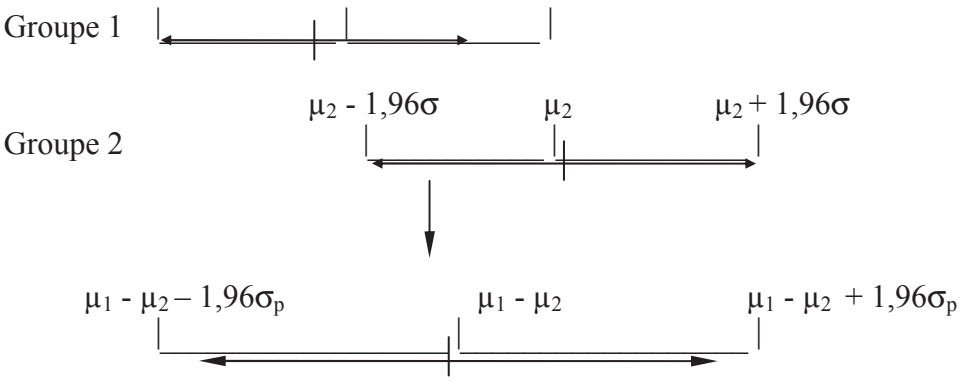
- $62,8 \pm 0,2$ conduirait à un échantillon comprenant 95% des patientes entre 62,4 et 63,2 années, et $55,5 \pm 0,2$ à un échantillon comprenant 95% des patients entre 55,1 et 55,9 ans (moyenne $\pm$ 1,96 σ). Les distributions des âges de ces deux échantillons ne se recouvrent pas, et l'on peut donc admettre que quoique proches en moyenne, les deux échantillons soient différents.
- Dans notre étude, l'écart-type est beaucoup plus important, et conduit à une distribution des âges pour 95% des patients entre 24 et 101,6 années pour les femmes, et 21,2 et 99,8 années pour les hommes : le recouvrement de ces deux distributions est considérable, et quoique les moyennes soient les mêmes que pour l'exemple précédent, l'on ne serait pas enclin spontanément à reconnaître comme différentes ces deux distributions.

La prise en compte de la variance est donc indispensable : le test de comparaison des moyennes est en fait, une *analyse de variance.* L'analyse de variance prend en compte d'une part, la différence entre les moyennes, et d'autre part, la variance combinée des deux distributions comparées (= variance poolée).
Cette variance combinée des deux distributions comparées, permet d'estimer la variance de la distribution de la différence de moyenne : la variance poolée traduit la variance de l'ensemble des échantillons comparés, et la variance de la différence des moyennes, celle de l'écart moyen entre les deux populations comparées.

La variance de la différence des moyennes permet ensuite de calculer un intervalle de confiance autour de la différence de moyennes. Si cet intervalle contient 0, la différence est alors considérée comme similaire à 0, et les deux moyennes semblables. Si l'intervalle de confiance ne contient pas 0, la différence des moyennes est considérée comme éloignée de 0, et les moyennes comme différentes. Selon l'exigence, l'intervalle de confiance peut être calculé à 95% (correspondant à un p de 0,05), à 99% (correspondant à un p de 0,01), à 99,9%... la *significativité de la différence des moyennes peut être donnée pour n'importe quelle valeur de p. p* donne la probabilité pour la différence des moyennes d'être égale à 0. On admet, comme plus haut, que les moyennes peuvent être considérées comme différentes si cette probabilité est inférieure à 5%, *si $p < 0,05$.*

En résumé, une représentation graphique de nos groupes pourrait être :

$\mu_1 - 1,96\sigma$ μ_1 $\mu_1 + 1,96\sigma$



où σ_p représente l'erreur-type de la différence des moyennes, calculée à partir de la variance *poolée* des deux échantillons initiaux, groupe 1 et groupe 2.

Si l'analyse de variance porte sur une population complète (rarissime en médecine), le test utilisé est le *test Z*, qui prend en compte la variance (avec n en dénominateur). Si l'analyse porte sur des échantillons de patients (cas habituel), le test à utiliser est le *test t,* qui prend en compte la formule de la variance corrigée (avec n-1 en dénominateur). On peut ensuite se référer aux tables du test t, à la ligne correspondant au nombre de patients totaux -2 (nombre de degrés de liberté du test t), et trouver la probabilité pour que la valeur de t soit différente de 0.
Bien sûr, n-1 sera très proche de n lorsque n, la taille d'échantillon, est grande. Dans ces situations, le test Z (prenant n au dénominateur), et le test t (prenant n-1 au dénominateur) donneront des résultats similaires.

Dans notre exemple, la valeur du test t est de 6,7. Nous ne serions pas surpris que sa probabilité d'être proche de 0 soit très faible, compte tenu de la taille d'échantillon : elle est, en effet, de 0,0001. Autrement dit, la probabilité pour que la différence d'âge entre le groupe des hommes et le groupe des femmes soit égale à 0 est de 1/10 000, ou *la différence des moyennes est significativement différente de 0 avec p = 0,0001. L'hypothèse nulle est refusée, et l'on accepte qu'il existe une différence d'âge significative entre les hommes et les femmes :* $\mu_f - \mu_h$ est significativement différente de 0.

3. Test non paramétrique : test de rang de Wilcoxon :

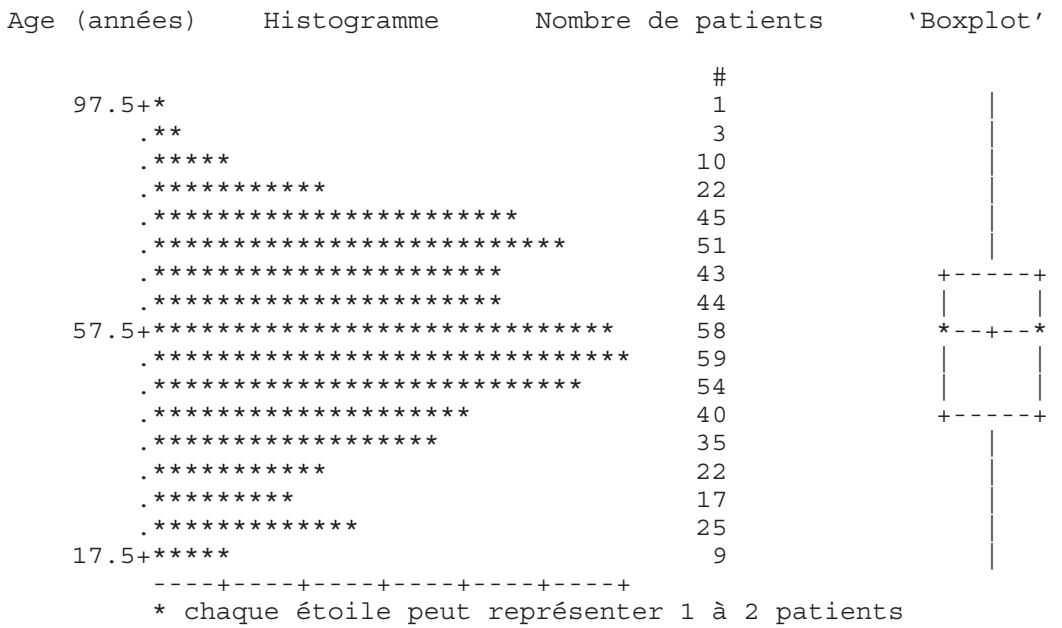
En regardant grossièrement la moyenne, l'écart-type, et les extrêmes des âges des femmes et des hommes, nous avons vu que ces valeurs étaient *compatibles* avec une *distribution normale.* Cet examen sommaire n'est cependant pas suffisant. Regardons les données de plus près (

```
Fig X : Pyramide des âges en années: groupe des femmes.
      Age (années)    Histogramme      Nombre de patientes     'Boxplot'
                                                #
        102.5+*                                 1                 |
              .***                              6                 |
              .*******                         14                 |
              .******************              35                 |
              .****************************    57                 |
              .******************************  62               +-----+
              .**************************      56               |     |
```

L'allure de la pyramide des âges ne paraît pas très gaussienne : il existe trois pics, le premier correspondant à 23 patientes sous la barre des 40 ans, le second à 40 patientes sous la barre des 60 ans, le troisième à 62 patientes au-dessus. Il est possible qu'existent dans cette population 3 groupes de patientes différentes, et il peut être intéressant d'examiner ces trois groupes séparément dans une étude exploratoire. Le schéma dit en 'boxplot' confirme cette impression : la moyenne (+) est différente de la médiane (barre transversale au milieu de la boîte), et la médiane n'est pas située à équidistance du 25ème et 75ème percentile (barres transversales inférieure et supérieure de la 'boîte'.

Le test de normalité effectué par notre programme statistique confirme cette impression : l'hypothèse nulle (celle de la normalité) est rejetée avec $p \leq 0{,}0001$ (la probabilité pour que cette distribution soit normale est inférieure à 1 chance sur 10 000).

Fig. 2 : Pyramide des âges chez les hommes



L'allure de la pyramide des âges des hommes est un peu différente de celle des femmes, mais il existe aussi différents pics. L'hypothèse nulle de normalité est également refusée avec p = 0,0001.

L'analyse de variance repose sur l'hypothèse de la normalité des distributions et de l'égalité des variances entre les groupes. Il existe des tests statistiques ne requérant pas la normalité des distributions ou l'égalité des variances, et permettant de comparer les groupes ainsi inhomogènes. Cela suppose de considérer les variables autrement et de prendre en compte, à la place de la valeur numérique, absolue, de la variable quantitative, *le rang qu'elle occupe dans la distribution.*
Dans notre exemple, les patients seraient ainsi classés dans chaque groupe du plus jeune au plus âgé, et le test consiste à évaluer la tendance générale des âges dans un groupe par rapport à l'autre : la question 'la moyenne d'âge compte tenu de la variance est-elle différente dans les deux groupes ?' devient 'un groupe est-il globalement plus jeune, ou plus âgé, que l'autre ?'

Application :

âge	18	19	20	21	.....	98	99	100	101	102
Femmes : n	2	3	4	2		2	1	3	1	1

âge	18	19	20	21	.....	95	96	97	98	99	100	101	102
Hommes : n	1	3	2	4		3	2	1	1	0	0	0	0

L'âge des femmes est ensuite comparé à l'âge des hommes par la constitution de 3 catégories de paires comprenant chaque fois une seule femme et un seul homme :

1- Les paires pour lesquelles les femmes sont plus jeunes que les hommes
2- Les paires pour lesquelles les femmes et les hommes sont d'âge égal
3- Les paires pour lesquelles les femmes sont plus âgées que les hommes

Dans notre exemple, les deux patientes âgées de 18 ans sont plus jeunes que les 537 hommes sur 538 âgés de plus de 18 ans : il y a donc 2 x 537 paires (= 1074 paires) pour lesquelles une patiente du groupe 1 est plus jeune qu'un patient du groupe 2.
Les 3 patientes de 19 ans sont plus jeunes que les (538 – 4) hommes âgés de plus de 19 ans. Il y a donc 3 x 534 (= 1602) paires supplémentaires pour lesquelles une patiente est plus jeune qu'un patient.
On continue ainsi à compter toutes les paires pour lesquelles une patiente est plus jeune qu'un patient, et on les additionne (nombre total : x).
Les deux patientes de 18 ans ont le même âge que le patient de 18 ans : cela fait donc deux paires d'âge égal. On dénombre ainsi toutes les paires d'âge égal (nombre total : y).
La patiente âgée de 102 est plus âgée que tous les 538 hommes, de même que la patiente âgée de 101 ans, celle âgée de 100 ans et celle âgée de 99 ans : cela donne 4 x 538 paires pour lequelles les patientes sont plus âgées que les patients, et l'on comptabilise toutes les paires 'plus âgées' pour l'ensemble de la série.
Le nombre de paires égales est ensuite partagé à parts égales dans le groupe « paires plus jeunes » et le groupe « paires plus âgées », qui comprennent alors x + y/2 paires pour le premier, et z + y/2 paires pour le second (nombre total : z). Nos deux groupes de patientes et patients sont ainsi transformés en deux groupes de paires, 'plus âgée' et 'moins âgée', que l'on va comparer.

L'hypothèse nulle devient : le nombre de paires est égal dans les groupes 'Plus âgée' et 'Moins âgée', ou (x + y/2) - (z + y/2) = 0, et la comparaison revient à une comparaison de proportions. La nombre de paires 'plus jeunes' rapporté au nombre total de paires est-il similaire, ou différent, du nombre de paires 'plus âgées' rapporté au nombre *total* de paires ? Comparaison de proportions renvoie au type d'analyse effectué par le test de Chi-2 décrit plus haut.

Ce *test de comparaison par rangs* dit de *Wilcoxon ignore* donc la distance existant entre les âges, et ne donne pas de poids supplémentaire aux âges extrêmes, contrairement à l'analyse de variance. Il n'est pas dépendant du caractère normal ou non de la distribution. Il permet l'analyse de petits échantillons : une moyenne calculée sur 10 patients a des chances de ne pas être très précise, la variance importante, et la comparaison avec un autre groupe de 10 patients peu puissante. Le test de comparaison par rang portera sur 10 x 10 = 100 paires, plutôt que sur 20 individus : dans le cas de petits échantillons, le test par rang de Wilcoxon sera probablement plus performant, mais également plus rigoureux qu'une comparaison de moyennes car il est rare qu'une distribution soit normale dans ces conditions, et l'analyse de variance pourrait conduire à une estimation fausse de p.

Le test de *U-Mann-Whitney* est basé sur le même principe et arrive, par une technique de calcul différente, aux mêmes résultats que le test de Wilcoxon. Ils sont interchangeables et les logiciels de statistiques fournissent indifféremment l'une ou l'autre procédure.

Attention :

- Comme tout test, le test de Wilcoxon rencontre des limites. Chaque groupe doit comprendre au moins 10 patients, sinon le test perd de sa précision et de sa valeur… mais vouloir comparer de trop petits groupes nous fait quitter de fait le champ de l'analyse statistique !
- En pratique, l'analyse de variance donnera les mêmes résultats que le test de Wilcoxon en terme de p lorsque la taille de chaque groupe dépasse 30, même si les distributions ne sont pas gaussiennes. Elle aurait été possible dans l'exemple ci-dessus, mais seule l'observation de la distribution des variables permet de se rendre compte qu'il y a peut-être trois groupes d'âge de patients.

4. Variables ordinales :

On utilise souvent en médecine des variables semi-quantitatives d'un type particulier : le cancer du colon est classé en stade A, B, ou C de Duke en fonction du degré d'envahissement de la muqueuse, et la gravité va croissante du stade A vers B puis vers C, mais C n'est pas *trois fois* plus grave que A, ou *deux fois* plus grave que B. Il en est de même des stades I à IV de la dyspnée ou de l'artériopathie des membres inférieurs. Ces stades ne peuvent pas s'additionner, se multiplier, ou se diviser : simplement, ils traduisent une hiérarchie dans la gravité de la maladie plus qu'une quantité strictement mesurable.
Ces variables semi-quantitatives sont dites *ordinales*.
On peut souhaiter comparer, cependant, des données de ce type en conservant leur caractère hiérarchisé, ou *ordonné* : l'administration d'un diurétique chez ce groupe de patients atteints d'insuffisance cardiaque améliore-t-il le stade de la dyspnée, autrement dit, la fait-il passer d'un ordre supérieur (III ou IV par exemple), à un ordre inférieur (II) ?

La comparaison du nombre de patients atteints d'une dyspnée de stade I, II, III, ou IV dans un groupe traité et non traité pourra faire appel au test de Wilcoxon : chacun des stades peut être considéré comme un rang, et le test de Wilcoxon (ou le test de U-Man-Whitney) pourra permettre de dégager la tendance vers le stade moins - ou plus - important de la dyspnée.

5. Généralisation de l'analyse de variance à plusieurs groupes :

Il peut être intéressant de comparer les moyennes et variances d'une variable quantitative entre plusieurs groupes. Dans notre exemple, on peut vouloir comparer l'importance de l'amaigrissement entre les groupes 'origine psychogène', 'origine cancéreuse', 'origine nutritionnelle ou endocrinienne', et 'origine autre'.

Il existe plusieurs façons de considérer le problème :
La première consiste à faire une seule analyse, et à examiner s'il existe, au moyen d'un seul test, une différence *quelque part* au sein de l'échantillon global des 4 sous-groupes. Le principe de base en est le même que pour l'analyse de variance entre deux groupes, et le test employé est le *F-test* (Généralisation du t-test à plusieurs groupes). Si le F-test ne met pas en évidence de différence significative au sein des x échantillons, on peut en conclure qu'il n'existe vraisemblablement pas de différence entre le groupe à plus faible, et le groupe à plus forte, moyenne. Par conséquent, il n'existe sans doute pas de différence entre les groupes dont les moyennes sont comprises entre la plus petite, et la plus grande, moyenne, et notre analyse pourra s'arrêter là.

Si par contre, le F-test objective une différence significative, c'est qu'il existe sans doute une différence entre le groupe à la plus petite, et le groupe à la plus forte, moyenne... sous réserve que ces deux groupes aient une taille d'échantillon permettant d'arriver à une différence significative. Il est possible aussi que la différence, en fait, siège entre deux groupes intermédiaires à forte taille d'échantillon, voire entre plusieurs groupes, voire de façon diffuse entre chaque groupe de patients. Comme un test de chi-2 à multiples cellules, le F-test permet de détecter une différence *quelque part, mais ne permet pas de la localiser.* Il peut pourtant être important de savoir si devant un amaigrissement conséquent, il vaille mieux s'orienter en premier lieu vers une étiologie ou une autre...

On peut envisager, pour répondre à cette question, de comparer les groupes deux à deux au moyen d'un test t. Ainsi, de comparer le groupe 'origine psychogène' au groupe 'origine cancéreuse', puis au groupe 'origine endocrinienne', puis au groupe 'origine somatique autre'... et de continuer par les comparaisons 'cancer'-'endocrinien', 'cancer'-'autre', puis 'endocrinien'-'autre' et d'épuiser toutes les combinaisons logiques possibles. Pour quatre groupes, il existe ainsi six possibilités logiques. Pour 5 groupes, dix. Pour 6, 15. La comparaison systématique de tous les groupes deux à deux pourrait paraître plus rigoureuse, ou plus porteuse d'information intéressante, que l'analyse globale du F-test.
Le problème cependant est qu'elle multiplie les chances de montrer une différence *significative par hasard :* nous acceptons une différence comme significative si elle a moins de 5 chances sur 100, d'être liée au hasard. De ce fait, si 100 comparaisons au hasard sont réalisées, il est très probable que 5 au moins d'entre elles, s'avèrent significatives... par hasard, puisqu'il s'agit de la limitation même du test statistique. Sur 20 comparaisons, une au moins peut être statistiquement significative par hasard, et cette *significativité* serait dépourvue de toute *signification* clinique ou biologique.

Il est donc nécessaire, pour éviter de tomber dans le piège du hasard et de tirer des conclusions fausses de données correctement recueillies, mais mal analysées, d'établir un garde-fou. Une première solution serait de diviser le p exigible par le nombre de comparaisons effectuées : si 6 comparaisons sont réalisées, le risque d'obtenir une différence à p = 0,05 est de 5% x 6, soit de 30 %. Pour ramener ce risque à 5%, le plus simple est de diviser le p exigible par 6, ce qui ramène à 0,05/6 = 0,0083 (6 comparaisons effectuées, 4 sous-groupes comparés deux à deux), le p exigible pour avoir moins de 5% de chances de se tromper en affirmant une différence significative. Cet ajustement s'appelle *ajustement de Bonferroni.*
D'aucuns prétendront cependant que cet ajustement est trop sévère, et ne permettra pas de reconnaître, dans cette démarche exploratoire sans hypothèse a priori sur la localisation de la différence, une différence existant dans la réalité à niveau de p = 0,05 entre deux sous-groupes donnés. D'autres types d'ajustement de p, moins sévères, ont été proposés : l'ajustement selon Tukey (qui exige des groupes de taille identique, ce qui est rare en médecine), selon Scheffé, que l'on peut appliquer même si les groupes sont de taille différentes, selon Neuman-Peul, qui consiste à diminuer le nombre de comparaisons effectuées (on commence par comparer les deux groupes extrêmes : s'il n'existe pas de différence, les calculs s'arrêtent là et une seule comparaison aura été faite. S'il existe une différence, on compare ensuite un des groupes extrêmes (à moyenne la plus basse par exemple) avec le deuxième groupe extrême opposé (à moyenne immédiatement inférieure au groupe ayant la moyenne la plus élevée). S'il n'existe pas de différence, les tests s'arrêtent là et on n'aura effectué que deux comparaisons : on déclare qu'a priori il ne doit pas y avoir d'autres différences significatives entre les sous-groupes. S'il existe une différence, on poursuit les comparaisons entre le groupe à moyenne inférieure et le groupe à moyenne directement inférieure au dernier groupe testé, et ainsi de suite). Ce type de procédure permet de ne pas comparer de façon systématique, tous les sous-groupes formés et d'arrêter les comparaisons de moyennes dès lors que la dernière différence testée n'est plus significative : moins de tests sont effectués, et l'ajustement de p pourra être moins sévère.
Il existe toujours un certain *trade-off*, un certain équilibre, entre rigueur du test et obtention de résultats *statistiquement significatifs :* les chances de déceler une différence significative sont moindres avec un ajustement de type Bonferroni qu'avec un ajustement de type Neuman-Peuls, mais une différence significative en Bonferroni aura plus de chances d'être réellement significative qu'une différence observée d'après Neuman-Peuls...

En tout état de cause, les comparaisons de moyennes multiples ne doivent se faire qu'après ajustement de p, et l'interprétation de ces différences doit tenir compte de la rigueur de l'ajustement : le p < 0,05 n'établit pas à lui seul une vérité médicale ou biologique, mais représente une donnée d'analyse devant faciliter la lecture des résultats.

6. Généralisation du test de Wilcoxon à plusieurs sous-groupes :

La comparaison d'une variable quantitative à distribution non normale, ou d'une variable ordinale peut être intéressante entre plusieurs sous-groupes : on peut vouloir comparer l'action d'un diurétique (groupe 1), d'un inhibiteur de l'enzyme de conversion (groupe 2) et d'un beta-bloqueur (groupe 3) dans le traitement de l'insuffisance cardiaque, et évaluer combien de patients dans chacun des trois groupes passent dans le ou les stades de dyspnée NYHA (cotée de I à IV) inférieure.
Une comparaison de moyenne ne sera pas possible, car on ne peut pas calculer de *moyenne* de stade de dyspnée. Un test de Wilcoxon cependant pourrait être réalisé pour des comparaisons deux à deux.

Le test de Kruskall-Wallis représente la généralisation du test de Wilcoxon pour comparaison de variables ordinales ou quantitatives non gaussiennes entre multiples sous-groupes. Il est par ailleurs équivalent au test de Wilcoxon pour la comparaison entre deux groupes.

7. Comparaison entre deux variables quantitatives :

Il ne s'agit plus maintenant de comparer deux moyennes, ou la variance d'une variable quantitative entre deux groupes (l'importance de l'amaigrissement chez les patients à étiologie somatique ou psychogène), mais d'examiner s'il existe une relation entre deux variables quantitatives : le taux de créatinine est-il fonction de la masse corporelle ? Les taux sériques d'un médicament potentiellement néphro- ou myélotoxique sont-ils fonction de la clearance de la créatinine ? Le taux de beta2-microglobuline est-il fonction de la masse tumorale du lymphome ? Existe-t-il une relation entre l'importance de la virémie HIV et le nombre de lymphocytes CD4 circulants ? Le taux d'hémoglobine glycosylée est-il un bon reflet des moyennes glycémiques ?

Ces questions relèvent toutes de la notion de *corrélation,* et la relation peut tout d'abord s'appréhender sur un graphe : le nombre de copies virales est porté sur l'axe des x, le nombre de lymphocytes CD4 circulants sur l'axe des y, et la forme du nuage des points peut être appréciée. Si les points se trouvent tous sur une droite parfaite, il existe très certainement une corrélation nette entre les deux variables, à deux exceptions près : si les points se trouvent tous sur un nuage vertical, cela veut dire que les valeurs des y (le nombre de CD4), ne varient pas en fonction de la virémie, mais peuvent prendre toutes les valeurs pour une virémie constante donnée. *Il n'y a donc pas de corrélation, et la pente de la droite (verticale) est égale à + l'infini.* Si les points se trouvent tous sur une droite horizontale, le nombre de CD4 reste constant quelle que soit la valeur de la virémie : *il n'y a aucune corrélation, là non plus, et la pente de la droite horizontale est égale à 0.*
Pour qu'il y ait corrélation, il faut tout d'abord que la pente de la droite soit significativement différente de 0 et de l'infini : le nombre de CD4 circulants doit varier avec le nombre de copies virales, le taux d'hémoglobine glycosylée doit varier avec la moyenne des glycémies. La pente de la droite sera positive (supérieure à 0), si l'augmentation d'un taux est associée à l'augmentation de l'autre. Elle sera négative (inférieure à 0) si l'augmentation d'un taux est associée à la diminution de l'autre.
Exemples : il y a une corrélation positive entre le taux d'hémoglobine glycosylée et les taux de glycémie, mais la corrélation est négative entre le nombre de copies virales HIV et le nombre de lymphocytes CD4 circulants.

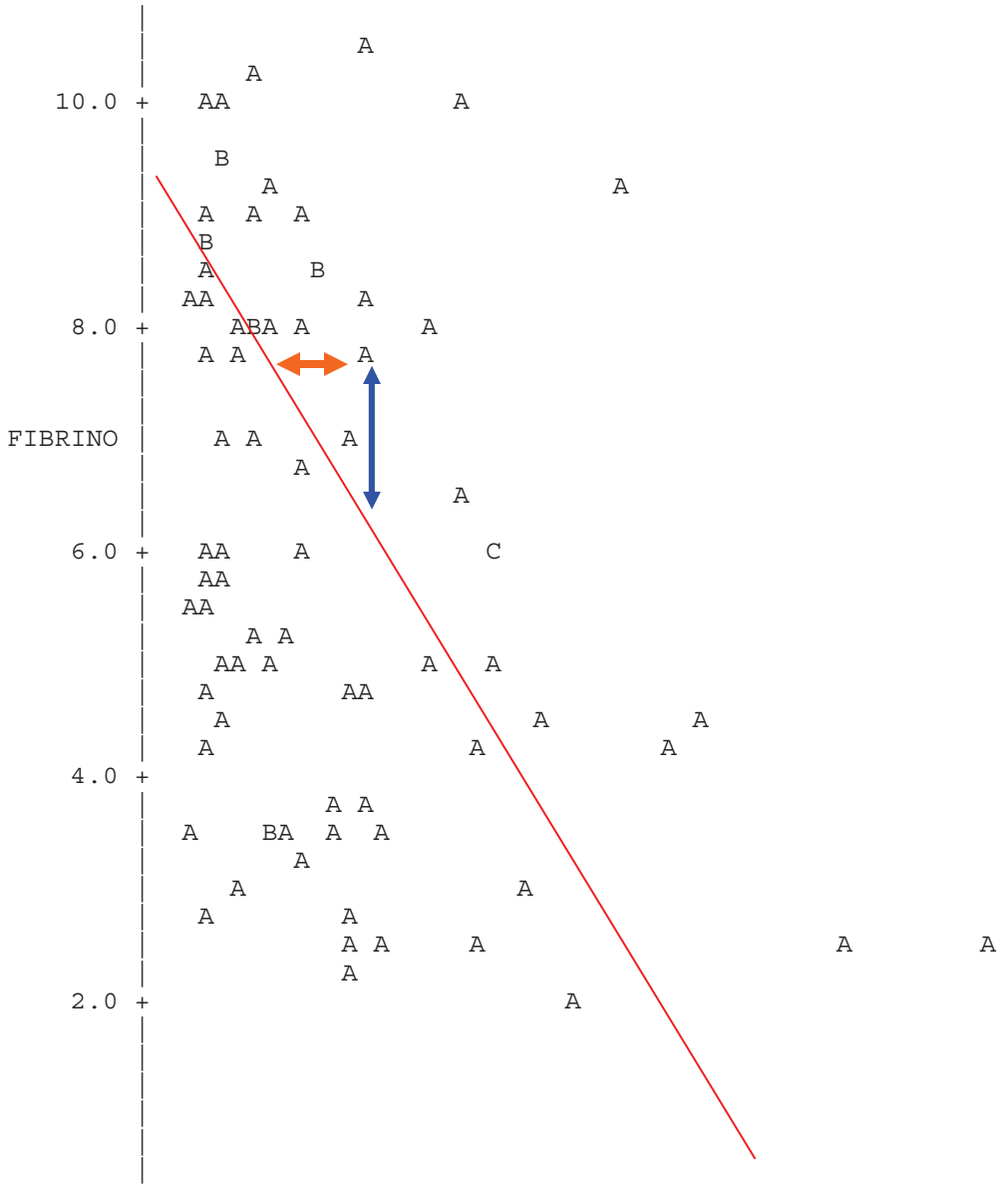
Le modèle de base de l'équation d'une corrélation est donc l'équation d'une droite :
$y = a.x + b$, où a représente la pente de la droite, et b l'intercept, ou la valeur de y lorsque x est égal à 0. Les valeurs biologiques égales à 0 sont rares en médecine, du moins pour les paramètres biologiques de base (NF, ionogramme sanguin, paramètres de la coagulation...) ou lorsque les techniques de dosages sont suffisamment sensibles pour détecter des taux faibles (une TSH rigoureusement égale à 0 est rare avec la technique de dosage ultra-sensible). L'intercept est donc souvent une valeur extrapolée par l'équation, et il n'est pas certain qu'elle corresponde à une valeur *biologiquement observée.*

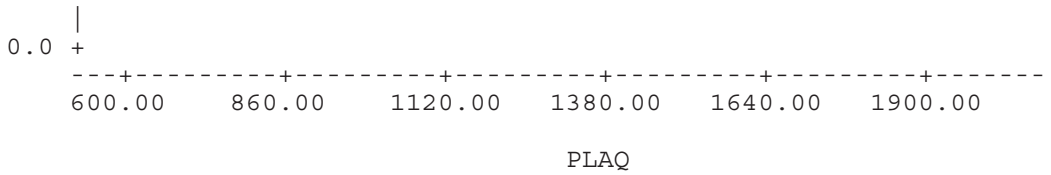
Il est rare cependant qu'en biologie ou en médecine, les points de corrélation entre deux variables puissent être reportés de façon *exacte* sur une droite. Ils forment le plus souvent, un nuage dont la droite de corrélation est la bissectrice. L'équation de la droite ne suffit pas à

décrire de façon satisfaisante le phénomène observé. Comme dans la comparaison de moyennes, où la variance décrit la dispersion des valeurs autour de la moyenne, il existe une variance de chacune des deux variables autour de la droite de régression. Pour un x donné (un nombre de copies virales), l'y correspondant (le nombre de lymphocytes CD4 circulants) peut se trouver à plus ou moins grande distance en-dessous ou au-dessus de la droite. La meilleure droite, celle qui représentera le mieux la corrélation, est celle pour laquelle la somme de l'ensemble de ces distances (la distance de chaque y par rapport à la droite) est la plus petite possible. Certaines de ces distances négatives (lorsque l'y observé est situé sous la droite), diminueraient artificiellement la somme des distances entre l'y observé et la droite (figurant l'y prédit par le modèle) : la parade à ce problème consiste à additionner non pas les distances brutes négatives et positives, mais le carré de ces distances. La meilleure droite décrivant le phénomène est celle pour laquelle la somme de ces carrés sera minimale : la technique de *fabrication* de la droite est appelée technique de la ***somme des moindres carrés (least square sum)*** (Fig. 1).

Figure 1 : CORRELATION ENTRE FIBRINOGENE ET PLAQUETTES

Légende: A = 1 observation, B = 2 observations, etc.





Dans cet exemple, nous avons réalisé un graphe de corrélation entre les valeurs de fibrinogène et celles de la numération de plaquettes dans une population d'hyperthrombocytose réactionnelle (essentiellement inflammatoire), et avons voulu tester l'hypothèse de mécanismes de régulation physiologique du risque de thrombose chez ces patients.
On se rend compte tout d'abord qu'il existe une corrélation négative (plus la thrombocytose est importante, plus le fibrinogène est bas), avec un coefficient de corrélation à – 0,34690. De plus, cette corrélation est significative puisque la valeur de p (égale à 0,0016) est largement inférieure au seuil classiquement admis de 0,05.
La flèche en bleu représente la distance entre la valeur observée de y (le point A sur notre graphe) et sa valeur attendue (calculée) sur la droite de régression linéaire : il s'agit de la distance (y – y'). La flèche en orange représente la distance entre la valeur observée de x et sa valeur calculée sur la droite de régression linéaire : il s'agit de la distance (x- x'). On comprend que si toutes les valeurs observées de y et de x se trouvaient sur la droite, alors la corrélation entre les deux valeurs serait parfaite : y serait toujours exactement prédictible en fonction de x, et vice-versa. Le coefficient de corrélation serait égal à 1 en valeur absolue. Plus les valeurs observées y et x se 'promènent' à distance de la droite, plus la corrélation est lâche : plus les distances (y-y') et (x-x') sont grandes, plus le nuage de points est épars, plus la corrélation sera faible, et plus le coefficient de corrélation s'approchera de 0.

Le ***coefficient de corrélation r (ou rho)*** prend en compte l'ensemble de ces distance y-y' et x-x' en les rapportant au nombre d'observations. Il prend également en compte, par son signe, la pente de la droite de régression : il sera négatif si la pente est descendante (plus les plaquettes sont hautes, plus le fibrinogène est bas) et positif si la pente est montante (plus les plaquettes sont hautes, plus le fibrinogène est haut). *Il traduit donc la somme de l'ensemble des écarts des points obervés, par rapport aux points calculés situés sur la droite.*

La *pente de la droite a (ou alpha)* traduit, elle l'augmentation ou la diminution de la valeur de y lorsque x varie : si alpha = 2, la valeur de y augmente deux fois plus vite que la valeur de x.

L'ensemble de la procédure (calcul de l'équation de la droite, de la variance de x, de la variance de y, du coefficient de corrélation) s'appelle une *régression linéaire.* A nouveau, une corrélation sera dite significative si la pente de la droite a suffisamment de chance de différer de 0 ou de l'infini, et si le coefficient de régression a suffisamment de chances de différer du 0. Autrement dit, si la droite de régression est suffisamment loin d'une droite horizontale ou verticale, et si les points observés sont suffisamment proches de la droite estimée. Le p de la corrélation prend en compte ces deux ingrédients, et l'on pourra voir ainsi des corrélations significatives à moins de 5 % de chances de se tromper, avec une pente faible mais un coefficient de corrélation proche de 1, ou à l'inverse avec une pente significative mais un coefficient de corrélation faible. Le plus souvent, une corrélation significative signe une tendance, mais il est rare qu'en médecine on puisse, à l'instar de ce qui se fait en physique ou en chimie, prédire connaissant x une valeur y à partir de l'équation de régression.

IV. Analyse multivariée : notion de régression logistique

L'ensemble des tests exposés jusqu'à présent constituent les outils de l'analyse univariée : analyse d'une variable en fonction d'un paramètre (groupe ou sous-groupes, autre variable dans la régression linéaire simple). Il existe de nombreuses situations en médecine dans lesquelles l'analyse univariée marque rapidement ses limites : on sait que l'hypertension artérielle, l'hypercholestérolémie, le diabète, le tabagisme, le stress, sont des facteurs de risque de maladie cardiovasculaire ; mais quel est le poids respectif de chacun de ces facteurs de risque dans la survenue d'un infarctus du myocarde ? Ces facteurs de risque sont-ils indépendants les uns des autres, ou certains d'entre eux ne font-ils que 'traduire' les autres (le tabagisme et l'hypertension artérielle par exemple ne sont-ils que des expressions, partiellement ou totalement, du stress) ? Certains de ces facteurs de risque sont-ils synergiques, ou au contraire antagonistes ? Les mêmes questions peuvent se poser pour les facteurs de risque connus du cancer du sein, du cancer du poumon, de la survenue d'une maladie thrombo-embolique, voire des maladies dont la cause objective est connue : le BK est à l'origine de la tuberculose, mais le contact avec le BK seul ne suffit pas à déclencher la maladie... avant la découverte du BK, les facteurs de risque de la tuberculose maladie multifactorielle auraient compté la malnutrition, la promiscuité, les conditions socio-économiques défavorables... on rajouterait à l'heure actuelle toutes les causes d'immuno-dépression, et plus récemment encore, la façon de réagir du système immunitaire, et notamment de l'immunité innée macrophagique, au contact du BK. La maladie multifactorielle avant la découverte du BK, devenue monofactorielle avec sa découverte, redevient de fait multifactorielle avec les progrès de l'immunologie et de la génétique…

Il est possible, pour évaluer le rôle de chacun de ces facteurs de risque, de les 'peser' de façon indépendante, de procéder à une succession d'analyses univariées en stratifiant par chacun d'entre eux : on pourrait comparer, dans une étude de cohorte, l'incidence de l'infarctus du myocarde parmi les fumeurs et les non-fumeurs ; puis, chez les fumeurs d'une part, et les non-fumeurs de l'autre, l'incidence de l'infarctus chez les hypertendus et les non-hypertendus ; puis, chez les fumeur hypertendus, les fumeurs normo-tendus, les non-fumeurs hypertendus, et les non-fumeurs normotendus, l'incidence de l'infarctus chez les diabétiques d'une part, et les non-diabétiques de l'autre ; puis, chez.... et ainsi de suite.

Cette succession d'analyses univariées permet bien sûr, de déterminer si le stress rajoute un risque supplémentaire d'infarctus dans chacune des sous-catégories, et permet de mesurer l'importance de ce risque : s'il est plus élevé dans la sous-catégorie fumeur-hypertendu-diabétique que dans la sous-catégorie fumeur-hypertendu-non-diabétique, c'est que peut-être le stress agit de façon synergique avec l'un des trois premiers facteurs de risque... mais on ne saura pas lequel.

Il va de soi que la puissance des tests diminue avec la taille d'échantillon : la stratification en sous-groupes demande une taille d'échantillon initiale très importante si les derniers sous-groupes doivent encore comprendre un nombre suffisant de patients... en pratique, cette stratégie bien que théoriquement satisfaisante, est rarement possible. Elle est également très lourde.

L'alternative est représentée par la *régression logistique* : sans entrer dans les détails mathématiques, le principe de son équation pourrait s'écrire ainsi (*attention, cette expression est mathématiquement fausse, mais sa signification globale est juste*).

Maladie = intercept + OR1. FR1 + OR2.FR2 + OR3.FR3 + OR4.FR4 + ...ORn.FRn.

Dans cette équation, la maladie s'exprime généralement de façon binaire : elle existe, ou non. Les différents facteurs de risque (FR) peuvent s'exprimer de façon binaire (1 ou 0), de façon ordinale (1, 2, 3, 4...), voire sous la forme d'une variable quantitative continue. Lorsqu'il s'agit d'une variable binaire, on peut extraire du coefficient qui lui est attribué l'odds ratio (OR), représentant le risque relatif lié au facteur de risque considéré *compte tenu du risque correspondant aux autres facteurs de risque.* Autrement dit, OR1, OR2, OR3, OR4, quantifient le risque associé à chaque facteur de risque, sachant que la maladie est aussi expliquée par les autres facteurs de risque gardés dans l'équation.

En pratique, on introduit dans le modèle de régression logistique les facteurs de risque significatifs à 0,1 en analyse univariée (pour lesquels $p < 0,1$). On peut forcer dans le modèle des facteurs de risque non significatifs en analyse univariée si cela paraît justifié sur le plan biologique ou médical. Il existe plusieurs types de procédures en régression logistique, mais le principe consiste à introduire d'abord dans le modèle le facteur de risque le plus significatif ; s'il n'explique pas à lui seul toute la maladie, on introduit dans le modèle le deuxième facteur de risque ; s'il est trouvé significatif, il reste dans le modèle. Si non, il en sort et le troisième est alors introduit et testé. L'introduction du xième facteur de risque s'arrête lorsque plus aucun facteur de risque n'est trouvé significatif, autrement dit, lorsque l'introduction d'un facteur de risque supplémentaire n'apporte plus d'explication supplémentaire à la survenue de la maladie.

Exemple : si dans les maladies cardiovasculaires, l'hypertension ou le tabagisme n'étaient qu'une expression (qu'une traduction) du stress, mais n'expliquaient pas par eux-mêmes une partie de l'incidence de l'infarctus de myocarde, ils seraient éliminés du modèle au profit de la variable stress. Si au contraire le stress ne jouait pas de rôle, et que l'hypertension ou le tabagisme agissaient comme facteurs *confondants* permettant au stress d'apparaître comme significatif en analyse univariée, le stress ne sortirait plus significatif du modèle de régression logistique qui ne garderait que l'hypertension et le tabagisme comme facteurs de risque vrais.

On peut calculer à partir des paramètres de chaque facteur de risque intégré dans l'équation de régression logistique les odds ratio pour chaque facteur de risque testé et leur intervalle de confiance. On pourra également tester l'interaction entre deux facteurs de risque. Imaginons, dans une étude des facteurs de risque du mésothéliome, que le tabagisme soit codé 1 si présent, 0 si absent. L'exposition à l'amiante sera codée de la même façon. Il est facile de créer une variable traduisant l'exposition simultanée aux deux facteurs de risque : tabac.amiante prendra la valeur 1 (1x1) lorsque les deux facteurs de risque seront présents, et 0 (1x0, 0x1, ou 0x0) lorsqu'il n'y aura pas présence simultanée du tabagisme et de l'exposition à l'amiante. On pourra écrire l'équation de régression logistique suivante :

mésothéliome = intercept + OR1.Tabac + OR2. Amiante + OR3. Tabac.amiante

Si la variable tabac.amiante est retenue dans le modèle comme significative avec un odds ratio supérieur à 1, c'est que l'association tabac-amiante rajoute un risque significatif par rapport à l'exposition au tabac d'une part, et à l'amiante d'autre part. C'est donc qu'il y *synergie* entre le tabac et l'amiante. Si l'association tabac-amiante est retenue comme significative, mais avec un odds ratio inférieur à 1, c'est que l'association est *antagoniste* : la présence conjuguée des deux facteurs de risque diminue le risque global de survenue de la maladie (les

antagonistes sont plutôt rares en médecine !). Si l'association des deux facteurs de risque n'est pas retenue dans le modèle, c'est qu'elle ne modifie pas le risque déjà exprimé par la présence du tabac d'une part, et de l'amiante d'autre part : les deux facteurs de risque ne sont ni synergiques, ni antagonistes, mais s'additionnent.

La régression logistique représente donc un outil extrêmement puissant et utile en médecine ou en biologie. Elle permet de réaliser une analyse fine en évitant l'écueil des tailles d'échantillon gigantesques nécessitées par l'analyse univariée stratifiée. Elle permet de *peser chaque facteur de risque, d'en mesurer l'indépendance par rapport aux autres, de tester les interactions, de contrôler les éléments confondants.*

V. Analyse du pronostic. Courbes de survie

Les courbes de survie représentent l'outil indispensable à l'analyse du pronostic : l'événement 'vivant/décédé' peut bien sûr constituer le facteur descriptif utilisé, mais tout autre événement traduisant le pronostic peut l'être également : survenue ou non d'une complication, survenue ou non de la guérison, survenue ou non d'une maladie intercurrente : leur caractéristique commune est qu'à chaque fois, l'événement peut être déjà survenu au moment de l'analyse des données ou peut ne pas l'être (parce qu'il n'est pas encore survenu, ou parce qu'il ne surviendra pas, et l'on parle alors de *donnée censurée*). Les courbes de survie incluent donc des patients pour lesquels on ne sait pas si l'événement étudié va survenir, ou non, un jour. Il peut sembler présomptueux d'analyser des données dont on ne saisit pas la réalité dans le futur : médecins, biologistes, statisticiens sont tous des humains, et l'avenir ne leur appartient pas... c'est vrai. Une courbe de survie ne permet donc jamais de décrire *ce qui va se passer chez un patient : elle permet tout au plus de décrire ce qui s'est passé chez les patients précédemment connus, et une probabilité de survie à un moment de l'évolution de la maladie. Il ne faut jamais se servir d'une courbe de survie pour affirmer à un patient ou à sa famille que son risque de décès, sa chance de guérison, son risque de survenue d'une complication est de x % à un an : pour un patient donné, le risque de décès est soit de 100 %, soit de 0 %. Il n'y a pas d'alternative entre le fait de vivre, ou de ne plus vivre. A l'heure actuelle, aucune science ne permet de prédire le futur, et d'utiliser un vocabulaire scientifique, voire une démarche calculée, pour essayer de l'approcher ne permet jamais d'être affirmatif au niveau d'un individu : le domaine du pronostic est sans doute celui dans lequel la médecine offre le plus d'incertitudes non circonscrites par les techniques d'analyse.*

Comment construit-on une courbe de survie ? Faiblesses et forces :

1- Le temps 0

Les seuls temps fixes dans une vie sont ceux de la naissance, habituellement datée avec précision, et de la mort, dont on peut connaître l'heure... une fois survenue. La seule courbe de survie *fournissant une information indubitable* serait donc celle décomptant le temps entre la naissance et la mort. Il faut, pour analyser une durée, partir si possible d'un temps 0 identique pour tous les patients.

Lorsque l'on s'intéresse au pronostic d'une maladie, le temps 0 est plus difficile à définir. On considère très souvent comme temps 0 le moment du diagnostic de la maladie, et l'on dit, un peu légèrement, que le pronostic du cancer du poumon est de x% de survie *à deux ans*. Les limites apparaissent de façon évidente : le temps 0, celui du diagnostic, n'est pas le même pour le patient dépisté en médecine du travail (petite tache ronde asymptomatique) que pour

celui diagnostiqué devant une altération massive de l'état général avec métastases osseuses et cérébrales. On pourrait, pour plus de précision, stratifier le temps 0 en fonction du stade du cancer. Les études de dépistage ont cependant montré que le temps 0 n'était pas le même, à taille d'image ronde pulmonaire asymptomatique sur une radiographie effectuée de façon systématique, pour une lésion diagnostiquée l'année du premier dépistage effectué dans une entreprise, année au cours de laquelle on dépiste des tumeurs ayant pu être présentes depuis 2, 3, 4... années, et l'année suivante, celle du deuxième dépistage, où l'on ne dépiste plus que les tumeurs apparues dans les douze derniers mois. A 1 centimètre de diamètre, une tumeur développée en 3 mois est sans doute plus agressive qu'une tumeur de même diamètre évoluant depuis 3 ans. Le temps 0 du diagnostic de la tumeur n'est pas le même : une tumeur agressive à trois mois peut déjà être évoluée, une tumeur indolente à trois mois peut encore être dans sa phase pré-clinique…

Le temps 0, préalable à la construction de toute courbe de survie, est donc défini par les moyens d'observation dont nous disposons. Nous retiendrons avant de construire une courbe de survie, que le soleil existe probablement avant l'heure de son lever, que l'heure de son lever change en tout point de la surface du globe, et que, si midi au sein d'un fuseau horaire sonnera au même moment, le vrai midi sera différent pour chaque individu en fonction de sa localisation dans le fuseau horaire. Une approximation de même nature, mais d'amplitude sans doute plus grande en fonction de la pathologie, préside à la construction d'une courbe de survie.

2- Notion de survie conditionnelle

L'idéal serait bien sûr, de disposer de la même période d'observation pour tous les patients inclus dans l'étude : on pourrait ainsi affirmer, en retenant cependant l'incertitude du temps 0, qu'à dix ans du diagnostic la survie de la cohorte atteint, par exemple, 50% (ce qui ne veut pas dire, encore une fois, que les chances de survie de Mr P. Dupont à 10 ans sont de 50%). Les patients cependant ne débutent pas tous leur maladie au même moment : certains seront suivis depuis 10 ans au moment de l'analyse, d'autres seulement depuis 1 an. Les premiers auront eu le temps de guérir, d'entrer en rémission, de décéder, les autres, non. On parle de *données censurées,* lorsque l'événement mesuré n'est pas encore survenu, et de *données non censurées,* lorsqu'il est survenu. Si le décès est l'événement mesuré, les données censurées correspondront aux patients vivants au moment de l'analyse (leur date de décès n'est pas connue).

La courbe de probabilité de survie cumulée basée sur le produit des probabilités conditionnelles (courbe de Kaplan-Meier) est donc bâtie de la façon suivante : tous les patients (100% d'entre eux), sont vivants au moment du diagnostic. La courbe de survie part, en ordonnée, de la valeur 100. Au premier décès (mettons-le à 3 mois du temps 0), 1% de la population initiale disparaît : la courbe descend d'une marche de 1% au niveau de l'abscisse 3 mois. Imaginons que deux décès surviennent à 6 mois du temps 0 : la courbe enregistrera alors une marche descendante de 2% sur la base des 99 survivants, à l'abscisse 6 mois. Si seuls 5 patients ont été suivis plus de 5 ans (soit parce que tous les autres sont morts, soit parce qu'ils ont été diagnostiqués durant les 4 dernières années...) et que deux patients parmi ces 5 meurent à 5 ans, la marche descendante sera de 2/5, soit 40% de la population résiduelle.

Ceci explique que sur une courbe de survie, les marches descendantes deviennent de plus en plus marquées vers les temps importants. Il apparaît également qu'elles deviennent de moins

en moins précises, car elles portent sur une taille d'échantillon de plus en plus réduite : *l'intervalle de confiance autour de la valeur, donc l'incertitude sur la valeur, augmente lorsque la taille d'échantillon diminue, et elle diminue de façon inéluctable au fil de la courbe de survie.*

La prudence s'impose donc lorsque l'on parle de probabilité de survie en médecine, et certains points doivent toujours rester présents à l'esprit :

- Une probabilité de survie s'applique à un groupe de patients, mais ne s'applique pas à chaque patient de façon individuelle. On ne sait pas, car l'outil statistique ne répond pas à cette question, si le patient diagnostiqué tel jour vivra, ou ne vivra pas, dans 5 ans : sa survie sera de 100%, ou ne sera pas. Sa probabilité de survie ne sera pas égale à celle du groupe.

- La probabilité de survie à x années ne vaut pour le groupe, qu'au moment du temps 0 : la probabilité conditionnelle de survie varie en fonction du temps, et pour l'exemple des 5 patients survivants à 5 années parmi 100 patients initialement inclus, elle sera, à ce moment de la courbe, de 40 % pour le temps à venir.

- L'avenir, même entouré de statistiques, reste une donnée très difficile à approcher.

3- Comparaisons de courbes de survie en mode univarié : le test de log-rank.

Deux courbes de survie finissent toujours par se rejoindre, il ne s'agit que d'une question de temps… à terme, la différence ne peut pas être significative. Par conséquent, leur comparaison dépendra de la rapidité avec laquelle chaque événement (le décès, la survenue de telle complication), surviendra en cours de suivi, et du nombre d'événements survenant à un temps donné. A un instant t, x% de patients dans un groupe, y% dans l'autre, survivront. La comparaison de proportions, à cet instant t, pourra donc être réalisée par un test de chi-2. A l'instant suivant, une des proportions pourra avoir varié : un second test de chi-2 devra alors être réalisé. Il en sera de même pour chaque changement de proportion sur l'une, ou l'autre courbe, autrement dit, pour chaque nouvelle marche d'escalier sur l'une, ou l'autre courbe.

La comparaison globale de l'ensemble des courbes de survie dépendra donc de la résultante de l'ensemble des comparaisons de proportions réalisées à chaque changement dans une des courbes analysées, soit, en termes statistiques, de la somme de l'ensemble des chi-2 effectués ajustés pour la taille d'échantillon variant à chaque étape. Le log-rank test réalise cette analyse et résume l'ensemble des différences additionnées : il s'agit d'une forme particulière de Mantel-Haenszel.

En pratique, ce test statistique mesure les écarts entre les deux courbes de survie comparées, à chaque marche d'escalier survenant sur l'une, ou l'autre, des courbes de survie. Il teste ensuite l'hypothèse nulle : la somme de ces distances est égale à 0. Si cela est vrai, les deux courbes ne sont pas très éloignées l'une de l'autre, et elles traduisent une survie similaire. Si cela n'est pas vrai, alors les deux courbes sont éloignées l'une de l'autre, et traduisent des survies différentes.

4- Comparaison de courbes de survie en mode multivarié : le modèle de Cox

Plusieurs éléments peuvent entrer en ligne de compte dans la survie : la présence ou l'absence de maladie, bien sûr, mais aussi le type de traitement reçu, la compliance au traitement, le respect du protocole initial, la présence de co-morbidités, l'existence d'autres facteurs de risque, le stade ou grade de la maladie, l'état clinique du patient mesuré par un score au moment du diagnostic, etc.

Tous ces éléments ne peuvent pas être représentés sur une courbe de survie, mais l'on peut imaginer que certains d'entre eux peuvent être plus importants, pour la survie du patient, que le type de traitement reçu. Si l'on compare deux traitements A et B, deux formes de la maladie X et Y, il sera important d'égaliser les facteurs pronostiques éventuels entre les deux groupes. Cette égalisation est le but de la randomisation dans un essai contrôlé, mais elle peut ne pas toujours être atteinte. Elle est rarement atteinte lorsqu'il n'y a pas de randomisation, et la prise en compte des divers facteurs intervenant dans la survie devrait alors faire appel à une stratification étagée, et la comparaison effectuée dans des sous-groupes de patients homogènes pour chaque facteur pronostique. Cela aboutirait rapidement à une multiplication de sous-groupes à taille d'échantillon réduite, et à une perte importante de puissance pour l'étude. Les résultats deviendraient difficiles à interpréter.

Une autre possibilité consiste à intégrer l'ensemble des variables pouvant jouer un rôle sur le pronostic dans un modèle d'analyse de type régression logistique, adapté à l'analyse de courbes de survie (capable de prendre en compte des données censurées, et non censurées). Ce modèle particulier de régression logistique est le modèle publié par Cox, qui pourra donner une estimation du risque relatif associé à chaque facteur pronostique. Si cet ce risque relatif est significativement différent de 1, le facteur considéré jouera un rôle significatif dans le pronostic. Dans le cas contraire, il pourra être considéré comme non déterminant. Dans tous les cas, le modèle donnera une estimation du poids relatif de chaque facteur pronostique dans la survenue de l'événement mesuré (décès, survenue d'une complication iatrogène ou naturelle,…), sans perte de puissance secondaire à une stratification quelconque. Par contre, seules les observations sans donnée manquante pour toutes les variables analysées seront prises en compte, ce qui peut résulter en une importante diminution de la taille d'échantillon si ces variables n'ont pas été correctement renseignées : comme pour toute étude clinique, la rigueur dans le recueil des données conditionne la fiabilité des résultats, quel que soit le degré de sophistication de l'analyse employée.

Conclusion

L'analyse statistique est indispensable en médecine et en biologie pour tester des tendances au niveau d'échantillons (qu'ils soient des groupes de patients, de colonies cellulaires, des groupes de gènes, des familles de protéines…) ou de population. Les tendances testées seront valables pour les populations testées, mais ne peuvent pas s'appliquer telles quelles à un individu donné, notamment pour les études de survie intégrant, en plus des données connues (non censurées) des données de valeur inconnue au moment de l'analyse (censurées). L'analyse n'est valable que si les données ont été correctement recueillies, sont complètes, et si elles correspondent à une hypothèse de travail *a priori*. L'analyse 'en pêche à la ligne' dans une base de données, à la recherche d'associations ou de corrélations statistiquement significatives sans fondement biologique ou clinique pressenti par l'investigateur, a de grandes chances de ne donner de résultats significatifs *que par hasard et doit donc être proscrite*. Il est possible que ce type d'analyse non planifiée soit à l'origine d'une grande partie des résultats contradictoires rapportés dans la littérature médicale, et les controverses alors engagées pourraient ne reposer, au moins partiellement, que sur des effets de hasard… et

non pas sur des bases d'acquisition de connaissances dites 'scientifiques'. Comme pour toute méthode d'observation, les conditions d'application, les indications et les contrindications des différents tests statistiques doivent être connues par les médecins qui seront de facto, les utilisateurs des résultats. Tout résultat ensuite *doit être interprété* en fonction de l'ensemble de la méthodologie de l'étude en amont du test statistique, et des conditions dans lesquelles le test a été appliqué.

Un test statistique prouve rarement. Il essaie de circonscrire le fait du hasard autour des événements observés, à condition que le modèle sous-jacent au test décrive assez bien la réalité multiforme de la vie (ceci peut parfois être assez difficile à affirmer). Il permet, en revanche, plus facilement d'éviter des affirmations indues (tel traitement est supérieur, ou différent, de tel autre) lorsque la différence de fréquence des événements observés est trop proche de 0.

Pour en savoir plus :

Bouyer J. Méthodes statistiques. Médecine-Biologie. ESTEM, Editions INSERM, 2000.

Bailer JC III, Mosteller F. Medical uses of Statistics, 2nd Edition, 1992, NEJM Books, Boston, Massachusetts.

Hill C, Com-Nougué C, Kramar A, Moreau T, O'Quigley J, Senoussi R, Chastang C. Analyse statistique des données de survie, 2ème Edition, 1996. INSERM Médecine-Sciences, Editions Flammarion.

Tableau I : Récapitulatif des tests statistiques courants, de leurs indications et limites.

Type de variable	Test indiqué	Limites d'applicabilité	Alternative nécessaire si limite du test indiqué atteinte
Dichotomique (proportions), 2 groupes	Chi-2 à 1 degré de liberté Alternatives : Yates (Chi-2 corrigé) Mantel-Haenszel (Chi-2 pour données stratifiées)	Une des cellules au moins de la table contient moins de 5 événements attendus	Test de Fisher exact
Dichotomique (proportions), plusieurs groupes	Chi-2 avec degrés de liberté = (nombre de colonnes -1)(nombre de lignes -1)	Ne permet pas de savoir où se trouve une différence si le test est significatif, mais indique simplement qu'il existe une notion de différence au sein de la table des données Donnera des résultats faussés si une des cellules de la table contient moins de 5 événements attendus	Si l'une de cellules de la table contient moins de 5 événements attendus, un test de Fisher est théoriquement possible, mais peu d'ordinateurs auront une mémoire suffisante pour le réaliser.... savoir que les résultats sont faussés, difficilement interprétables, et envisager un autre type d'analyse des données.
Ordinale, 2 groupes	Wilcoxon rang sum test	Chaque groupe doit contenir au moins 10 observations	
Ordinale, groupes multiples	Kruskall-Wallis	Indique s'il existe une différence quelque part entre les différents groupes. N'indique pas où se trouve la différence.	
Quantitative, 2 groupes	Analyse de variance (comparaison de moyennes)	La distribution des variables doit être normale (gaussienne) dans les deux	Si la distribution n'est pas normale, employer le test de Wilcoxon comme

	Z-test si population très importante t-test si échantillon : situation clinique habituelle	groupes.	pour les variables ordinales. Le test de Wilcoxon et l'analyse de variance donneront des résultats similaires si n > 30
Quantitative, plusieurs groupes	Analyse de variance F-test : analyse globale Analyse en sous-groupes avec ajustement selon Bonferroni, Scheffe, Tukey,	La distribution des variables doit être normale dans les différents sous-groupes. Indique qu'il existe une différence quelque part entre les groupes. N'indique pas où se trouve la différence. Indique quels sous-groupes diffèrent entre eux. La sévérité du test dépend du type d'ajustement (Bonferroni étant le plus rigoureux)	Si la distribution n'est pas normale, utiliser le test de Kruskall-Wallis comme pour les variables ordinales.
Relation entre deux variables quantitatives	Test de corrélation selon Pearson	La distribution des deux variables doit être normale. Test très sensible aux valeurs extrêmes (outsiders)	Si la distribution n'est pas normale, utiliser le test de corrélation de rang (selon Spearman).
Relation entre une variable quantitative et plusieurs variables quantitative	Régression linéaire multiple	La distribution de toutes les variables doit être normale.	Si la distribution n'est pas normale, ou pour des variables de distribution particulière, il existe des modèles de régression non linéaire, avec ou sans

			transformation des variables
Relation entre une variable dichotomique, et plusieurs variables dichotomiques, ordinales ou quantitatives	Régression logistique	Permet l'estimation d'interactions entre les variables, le contrôle des éléments confondants et modificateurs d'effet. Ne requiert pas la distribution normale pour les variables quantitatives.	
Construction d'une courbe de survie	Modèle de Kaplan-Meier	Etablit la probabilité conditionnelle de survie. Sa précision diminue au fur et à mesure du suivi et de la diminution du nombre de patients suivis	
Comparaison de deux courbes de survie	log-rank test (équivalent du test de Mantel-Haenszel)	Ne permet pas la prise en compte des co-facteurs intervenant dans la survie	
Comparaison de deux courbes de survie, avec prise en compte de co-facteurs, d'éléments confondants, ou de modificateurs d'effet	Modèle de Cox	Permet de déterminer le poids respectifs des différents facteurs intervenant dans la survie. Permet d'estimer le risque relatif pour les facteurs dichotomiques. Permet de repérer les facteurs confondants ou les facteurs modificateurs d'effet.	

Chapitre XVII
Budget et recherche de financements

Pierre Duhaut, Jean Schmidt, Claire Andréjak, Jean-Pierre Ducroix

L'établissement du budget représente une partie importante de l'élaboration d'un projet de recherche :

- Sur le fond, il doit prendre en compte tous les éléments du projet comportant un coût, tant de fonctionnement que d'équipement (matériel). Il comprend les coûts salariaux (emploi de technicien ou d'assistant de recherche). Il traduit la faisabilité du projet.
- Sur la forme, il doit respecter les indications fournies par chaque appel d'offres : le non-respect entraînera souvent le rejet du projet sans que le fond en soit examiné.

Il existe de nombreux appels d'offres chaque année. Ce chapitre se propose de les passer en revue par grandes catégories, mais il conviendra également, pour un projet ou un sujet donné, d'effectuer une recherche systématique des sources de financement possibles, actuellement exposées sur l'Internet.

La recherche clinique pendant très longtemps s'est passée de financement individualisé, car ses coûts étaient pris en charge de façon globale par les institutions au sein desquelles elle était réalisée (hôpitaux et universités). Ces coûts cependant ont toujours été réels et se divisent, comme toujours, entre fonctionnement et équipement. Les demandes de crédit doivent détailler de façon explicite ces deux grands postes dans le budget, et reprendre de façon précise les points spécifiques à chaque projet de recherche pour en préciser les montants de façon crédible.

I- Le budget :

Le plan est particulier à chaque demande, et il convient de vérifier avant de déposer la demande les conditions posées par l'appel d'offres : certains appels d'offres ne concernent que le fonctionnement, d'autres (plus rares), l'équipement.

1- Le fonctionnement :

Il comprend les salaires (charges comprises) et toutes les dépenses directes ou indirectes en relation avec la collection, l'analyse, et la présentation des données du projet de recherche.

a- Les salaires :

Il est difficile à l'heure actuelle de concevoir un projet de recherche clinique sans technicien (ou assistant) de recherche clinique dont le rôle sera d'assurer la coordination pratique entre les différents centres pour les projets multicentriques (ouverture des centres sur place, installation des outils informatiques), l'entrée des données, la gestion des données manquantes (partie importante du travail), l'envoi régulier des questionnaires de suivi, etc. Les compétences requises peuvent être celles d'une secrétaire maîtrisant les outils informatiques ou d'un technicien (BTS ou plus) maîtrisant les techniques de laboratoire, d'imagerie, ou d'informatique nécessaires au projet. Le budget devra mentionner le nombre d'équivalents temps plein et le salaire charges comprises (employé et employeur) par an et pour la durée totale du projet. Le montant des salaires et des charges est habituellement calqué sur la grille salariale de l'institution dans laquelle travaillera la personne engagée. Les salaires constituent habituellement la partie la plus importante du budget demandé. Il faudra également prévoir, en cas d'emploi à durée déterminée, les primes de licenciement en fonction de la durée prévisible de l'emploi.

b- Le petit équipement : il entre dans les frais de fonctionnement, et comprend tout le matériel de maintenance informatique, les fournitures de bureau, le petit matériel de laboratoire (pipettes…), les réactifs le cas échéant.
c- Les frais afférents à la communication entre centres : courriers, téléphone, fax…
d- Les surcoûts induits par le projet sur les patients à inclure : ils comprennent notamment le coût de tous les examens complémentaires non justifiés par la prise en charge médicale proprement dite des patients inclus. Ils peuvent être importants en cas d'imagerie (IRM, PET-scan) ou de dosages biologiques (radio-immuno-assays…). Chaque acte cependant est affecté de coûts différents : le prix coûtant est très largement inférieur le plus souvent, à celui indiqué par la nomenclature de la sécurité sociale (B, Z…), et ces prix peuvent se négocier avec les institutions parties prenantes dans le projet (laboratoires de biologie, services d'imagerie…).
e- Les coûts de conservation d'échantillons biologiques (Biobanque).
f- Les coûts afférents aux transports nécessaires entre les différents centres participants, hébergements éventuels : il existe dans les entreprises publiques et privées, des normes en fonction du kilométrage effectué et du type de transport.
g- Les coûts afférents à l'entrée des données si celle-ci est sous-traitée à une entreprise spécialisée.
h- Les coûts afférents à l'analyse des données si elle n'est pas faite par l'un des investigateurs du projet.
i- Les coûts relatifs à la communication des résultats, notamment dans les congrès ou dans des revues à facteur d'impact élevé, dont certaines imposent un tarif par page voire par illustration.
j- Les frais généraux prélevés par certaines institutions (locaux, gestion du compte)…
k- Les coûts de passage au comité de protection des personnes (anciens comités d'éthique)…
l- Tout autre coût identifiable…

2- L'équipement :

Il doit souvent faire l'objet d'une procédure à part, dans des appels d'offres différents. Les coûts d'équipement sont par exemple l'achat de matériel informatique dédié à l'étude envisagée (micro-ordinateur, moyens physiques de sauvegarde, moyens de protection des données…), un outil de mesure indispensable à l'étude (matériel de laboratoire, hors consommables qui entrent dans les frais de fonctionnement)…

Il est important pour la crédibilité de la demande que chaque ligne budgétaire renvoie à une étape précise du projet de recherche, et que chaque étape nécessitant un financement soit représentée dans le budget. Un projet peut faire l'objet de plusieurs demandes de financement : il faudra alors préciser quelle part du budget est demandée à chaque source de financement.

Un budget ne respectant pas la forme de l'appel d'offres auquel il répond aura toutes les chances d'être rejeté sans être examiné sur le fond.

II- Les sources de financement :

Elles sont multiples et peuvent varier d'une année à l'autre. Nous en considérerons quelques unes parmi les plus constantes à l'heure actuelle, en sachant que de multiples facteurs prévus (élections…) ou imprévus (crises…) peuvent en modifier les montants et qu'une source de financement ne peut que rarement (jamais ?) être considérée comme pérenne. L'information passe actuellement par Internet, et les sites des différents organismes sont à consulter de façon

régulière. L'information par voie institutionnelle est beaucoup plus lente et arrive parfois après… l'expiration des délais de soumission.

A- Les grandes sources de financement public :

Elles ont l'immense avantage de garantir une certaine liberté de recherche à l'heure actuelle. Elles correspondent à des thématiques particulières, définies par période (années, ou durée d'un plan), par les pouvoirs publics correspondants.

1- Le ministère de la santé :

a- A l'échelon national : le programme hospitalier de recherche clinique (PHRC national). Il existe en France depuis 1993 et a été lancé presque chaque année depuis. Il couvre 5 à 6 grandes thématiques particulières définies chaque année (maladies rares, Alzheimer, évaluation de nouvelles technologies, etc.) et obéit à un calendrier précis de soumission de l'idée du projet (habituellement en début d'année civile), puis du projet complet (habituellement au cours du premier trimestre de chaque année). Il est doublé d'un programme 'Cancer' particulier, aux modalités similaires. Il favorise les projets multicentriques et multidisciplinaires, et peut délivrer des financements allant d'une à plusieurs centaines de milliers d'euros par projet pour une période de 3 années.

b- A l'échelon inter-régional ou régional : un deuxième programme hospitalier de recherche clinique est coordonné au niveau régional ou inter-régional, et favorise les projets de recherche stimulant la coopération entre établissements hospitaliers d'une région ou inter-région. L'appel d'offres est actuellement lancé chaque année, il précède l'appel d'offres du PHRC national, et permet des financements de l'ordre d'une centaine de milliers d'Euros pour une période de 3 ans maximum. Un même projet ne peut pas être déposé au niveau national et au niveau régional ou inter-régional simultanément.

c- A l'échelon local :

i. Les hôpitaux disposent également de budgets au titre d'un PHRC local, pouvant habituellement financer des projets de l'ordre d'une dizaine ou de plusieurs dizaines de milliers d'euros. Le calendrier et les thématiques potentielles sont fixés par chaque hôpital.

ii. Système d'Interrogation, de Gestion et d'Analyse des Publications Scientifiques (SIGAPS) : depuis 2007, tous les CHU, à la demande du ministère de la Santé, se sont équipés de ce logiciel permettant de recenser l'activité de publication d'un service. En gros, le logiciel reprend les publications signalées dans PubMed (Medline) sur la base du nom des auteurs et les ré-attribue ensuite aux services et aux pôles. La part du budget annuel du CHU théoriquement consacrée à la recherche est calculée sur la base de l'index SIGAPS, et peut atteindre 13 % du budget global. Ces budgets, très importants, sont comptabilisés dans l'effort de recherche de la France… mais il revient à la CME (Commission Médicale d'Etablissement) et à chaque chef de service ou de pôle d'en négocier avec la direction de l'hôpital l'attribution effective au moins partielle aux activités propres à la recherche. On rappelle que le temps consacré à la recherche par les hospitalo-universitaires ne saurait être décompté de ce budget et qu'il doit être considéré comme faisant partie du mi-temps universitaire.

2- Les grands organismes publics de recherche :

Ils sont actuellement regroupés sous l'égide de l'Agence Nationale de la Recherche (ANR), qui centralise les appels d'offres et les attributions de budget auparavant répartis entre

l'INSERM, le CNRS et les autres organismes. Le site Internet *(http://www.agence-nationale-recherche.fr/Accueil)* doit être consulté régulièrement.

3- Les Régions, départements et collectivités locales :

a- Les Groupements Régionaux de Santé Publique (GRSP) :
Ces groupements sont présidés par le Préfet. Ils sont composés de l'Etat [services de l'Etat : DRTEFP (Direction Régionale du Travail, le l'Emploi et de la Formation Professionnelle), DRPJJ (Direction Régionale de la Protection Judiciaire de la Jeunesse), Rectorat], de ses établissements publics [INPES (Institut National de Prévention et d'Education pour la Santé) et INVS (Institut National de Veille Sanitaire)], de l'ARS (Agence Régionale de Santé) qui remplace depuis 2009 l'Agence Régionale de l'Hospitalisation, de l'Assurance Maladie [URCAM (Union Régionale des Caisses d'Assurance Maladie) et CRAM (Caisse Régionale d'Assurance Maladie)] et des collectivités territoriales volontaires. Les appels d'offres sont habituellement publiés deux fois par an et concernent avant tout des problèmes de santé publique (au sens large du terme). Les associations peuvent être financées à condition de disposer d'un numéro SIRET (indispensable pour pouvoir bénéficier de financements publics). Les sites Internet à consulter sont spécifiques pour chaque région.

b- Les Conseils Régionaux :
Ils disposent de moyens financiers importants gérés de façon spécifique suivant les régions. Il existe, dans la plupart des régions, un budget dévolu à la recherche. Selon les régions, les financements peuvent être obtenus sur appels d'offres émis soit directement par le conseil, soit par l'intermédiaire des institutions régionales telles que les CHU ou les universités.

c- Les départements, les villes, peuvent également soutenir des actions de recherche, et ces possibilités sont à explorer au cas par cas.

4- Les financements européens :
Une part importante du budget de l'Union Européenne est consacrée au développement et à la recherche. Si l'industrie (aéronautique notamment) se taille la part du lion, la santé et les sciences du vivant ne sont pas négligées. La France est moins présente dans l'attribution des budgets que les pays nordiques ou anglophones, et a sans doute négligé pendant trop longtemps cette source de financement.
Les priorités sont définies après une consultation internationale dans des programmes quinquennaux, appelés 'Programmes-Cadres' (PC) ou 'Framework Programs' (FP). Le 7ème PC a débuté en 2007 et sera étendu sur une période de 7 (et non plus 5) années. Des appels d'offres dans différents domaines sont lancés chaque année, dont les modalités sont consultables sur le site CORDIS (http://cordis.europa.eu/home_fr.html**).** Les financements concernent des projets plurinationaux impliquant le plus souvent au moins trois pays de l'Union Européenne et/ou les pays alliés (Suisse, Israël), souvent multidisciplinaires. Les dossiers doivent impérativement correspondre aux thématiques fixées. Une année de préparation de dossier est à prévoir si l'on veut correspondre aux exigences du programme, et le calendrier des appels d'offres est donc à consulter régulièrement. Des programmes spécifiques (maladies rares, pathologies particulières), peuvent faire l'objet d'appels d'offres indépendants à modalités particulières, mais peuvent être recensés sur le site CORDIS.
Des budgets particuliers sont également prévus pour l'échange de chercheurs (Actions 'Marie Curie').
Pour faciliter la constitution des dossiers, des 'Points de Contact Nationaux' spécifiques pour les différents domaines de la recherche ont été mis en place dans les différents pays. Leurs

coordonnées (adresse e-mail et localisation géographique) sont spécifiées sur le site CORDIS. Le contact national plus spécifiquement en charge de la recherche clinique en France fait habituellement partie de la Délégation à la Recherche de l'Assistance Publique des Hôpitaux de Paris, et il est vivement conseillé de le consulter avant de s'engager dans la procédure.

B- Les sources de financements privées, ou publiques/privées :

Un grand nombre de sources couvre un domaine extrêmement large de la pathologie et des recherches possibles en rapport. Les budgets alloués par ces sources sont en règle générale plus modestes que ceux proposés par les grands appels d'offres publics. Ils ne sont cependant pas à négliger, car un projet peut avoir plus de chances d'être financé lorsqu'il bénéficie de soutiens multiples. De plus, ces sources peuvent financer des aspects particuliers, et très souvent les jeunes chercheurs ou leur mobilité.

1- Les sociétés savantes :
Leur site doit être systématiquement consulté, notamment pour les financements des internes et assistants se lançant dans un projet de recherche, ou à la recherche d'un financement pour une année de master ou de mobilité, notamment dans le cadre d'une thèse de sciences : diverses bourses peuvent être offertes, cumulables à d'autres pour pouvoir couvrir l'ensemble des frais engagés.

2- Les fondations :
De multiples fondations, le plus souvent en rapport avec une pathologie précise, un groupe de pathologies (affections neuro-dégénératives…) ou un symptôme particulier (troubles de la vue, handicap moteur) peuvent financer des projets de recherche clinique visant soit à l'amélioration directe des conditions de vie des patients, soit à l'approfondissement des connaissances fondamentales. Certaines de ces associations-fondations sont très célèbres et répandues sur tout le territoire (Ligue Nationale contre le Cancer, Association de Recherche contre le Cancer, Téléthon, Sidaction…) et peuvent financer des projets de tous ordres au niveau régional ou national. La plupart sont moins connues et méritent d'être recherchées systématiquement sur le web en fonction du sujet de la recherche. Certains appels d'offres peuvent également être publiés sur le site de l'Agence Nationale pour la Recherche (ANR).

L'Académie de Médecine à cet égard fonctionne comme une fondation et attribue chaque année des prix d'un montant très variable, provenant de legs pour lesquels le donateur a émis une orientation particulière. Les jeunes médecins, en particulier, peuvent bénéficier de bourses ou financements dans le cadre de masters ou de thèses.

3- L'industrie et les entreprises :
Les financements proposés correspondent aux intérêts des entreprises et sont donc tout particulièrement orientés (exemple : nutrition du diabétique, des enfants, des nourrissons…). Le cas particulier de l'industrie pharmaceutique est à considérer : de plus en plus, elle construit, coordonne, analyse, et publie ses propres projets, et les médecins 'investigateurs' en fait… incluent les patients sans autre participation à la recherche (de la façon de poser la question jusqu'à la rédaction des articles !). La question de l'indépendance de la recherche, de l'objectivité de l'analyse , de l'interprétation et publication des résultats doit donc se poser, et, dans la mesure ou les financeurs de cette recherche d'entreprise sont *in fine* les systèmes d'assurance maladie, un partenariat plus équilibré entre producteurs de traitements et recherche clinique est à inventer pour retrouver la liberté de la recherche universitaire ou hospitalière, éviter le mélange des genres (recherche fondamentale et production/évaluation et

application clinique/formation des jeunes médecins et formation médicale continue/décision de mise sur le marché/recommandations voire recommandations opposables), et mieux œuvrer pour le mieux-être des patients…

III- Les stratégies de financement :

D'une manière générale, la constitution d'un bon dossier est longue. Plusieurs règles sont donc à observer :

1- Le calendrier des appels d'offres :
Il doit être anticipé en se référant aux dates de l'année précédente. Ces dates peuvent parfois changer, mais le délai entre le lancement de l'appel d'offres et sa clôture est le plus souvent trop court pour la rédaction d'une première demande d'un clinicien non expérimenté dans ce type de dossier, et souvent submergé par son activité clinique. La construction correcte d'un premier projet (type master par exemple) peut prendre 6 à 9 mois, même si l'histoire abonde de projets mal construits et bien financés. L'expérience prouve que raccourcir ce temps mène trop souvent à des projets dont l'apparence peut sembler correcte, qui éventuellement pourront être financés, mais qui se heurteront à de nombreuses difficultés de réalisation pratique ou d'analyse en limitant considérablement la valeur. *Un architecte ne construit pas une tour sur un brouillon ou des approximations… et le vivant est plus complexe que l'inerte.*
La possibilité du financement d'un projet de recherche clinique est à prendre en compte dès le début (c'est la question « mon idée de recherche est-elle réalisable »). Cependant il ne faut pas confondre remplir un dossier de demande de financements, et rédiger le protocole de votre étude. Il faut d'abord que votre protocole soit bien construit et réfléchi, avant d'en espérer un financement.

2- Les thématiques des appels d'offres :
Elles doivent être respectées… au même titre que la forme imposée par l'organisme financeur : cet aspect peut paraître secondaire mais peut être un élément clé dans la décision. Un dossier incomplet ou ne correspondant pas à la présentation demandée est souvent rejeté sans même avoir été évalué sur le fond.

3- La taille du projet :
Les structures nationales, régionales ou locales le plus souvent financent des projets à leur échelle : il faut donc choisir l'appel d'offres correspondant à la dimension du projet, à la fois sur le plan budgétaire (une étude sur une pathologie fréquente peut être réalisée sur quelques sites seulement, mais être de taille d'échantillon nécessitant un financement de type PHRC national) et sur le plan de sites de recrutement des patients et de la collaboration inter-disciplinaire (allant du local à l'international).

4- L'âge et les qualifications de l'investigateur principal :
Un interne même thésé ne peut pas, dans l'état actuel des choses, être investigateur principal d'un projet soumis à PHRC national ou inter-régional. Le projet devra donc être porté par un senior même s'il a été conçu par un collègue plus jeune. A l'inverse, un certain nombre de bourses et prix comportent une limite d'âge, vite atteinte dans la profession médicale : il faut donc encourager les internes et jeunes assistants à y postuler, et encadrer leur travaux pouvant par ailleurs s'intégrer dans un projet plus vaste de service ou d'unité. Ceci est d'autant plus nécessaire que l'obtention de budgets plus importants de type PHRC est conditionnée en partie par les publications antérieures de l'investigateur principal et des investigateurs associés : *là comme ailleurs, on prête plus facilement aux riches…* et il faut donc se soucier

d'établir la richesse de ceux qui pourront postuler dans les quelques années à venir, sous peine d'effondrement rapide de l'activité de recherche d'une unité. Les Directions de la Recherche et de l'Innovation des CHU peuvent également vous aider dans plusieurs étapes. Ces structures sont informées des appels d'offres et de leur calendrier. Elles peuvent fournir un « conseil » dans l'établissement de vos dossiers de demande de financement dont les exigences et la mise en forme sont parfois déroutant à première vue.

5- Nombre de projets par service ou unité :
Un seul clinicien ne peut pas pour des raisons évidentes de disponibilité et de compétence, mener de front plusieurs projets de recherche tant soit peu solides : un même projet doit permettre d'approfondir différents aspects si l'on veut rester à la pointe dans son domaine. Par contre, il peut être risqué pour un service de n'investir que sur une thématique unique : beaucoup de moyens (technicien/ne de recherche, moyens logistiques, informatiques) peuvent être mutualisés sur quelques projets de recherche clinique différents et auront donc plus de chances d'être pérennisés si une thématique passe de mode ou si d'aventure un projet bien commencé venait à manquer de financement. *Ici comme ailleurs, une entraide bien comprise peut bénéficier à chacun.*
Les Centres d'Investigations Cliniques peuvent répondre à cet objectif en mettant à disposition une structure (locaux, moyens informatiques) et des personnels (attachés de recherche clinique), à condition d'apporter le financement spécifique au projet..

6- Les échecs de financement :
Ils font partie intégrante de toute recherche de fond, et peuvent être motivés par des raisons plus ou moins objectives. Ils peuvent servir à améliorer un projet soumis lorsque les critiques des experts sont constructives ; mais toute activité humaine est soumise aux aléas de l'humain, et les échecs ne doivent jamais servir de prétexte à abandon lorsque les critiques ne sont pas à la hauteur de l'idée ou du projet soumis… l'histoire des sciences et de la médecine est remplie d'idées rejetées puis passées à la postérité ! De façon plus modeste, un projet valable doit souvent être soumis à plusieurs appels pour pouvoir être financé, et il faut sans doute le débuter au moins de façon partielle dès les premiers financements obtenus pour avoir une chance de pouvoir le compléter ultérieurement.

Points importants :

- *Anticiper les appels d'offres, pour prendre le temps de les construire correctement en en respectant les modalités et la forme.*
- *Intégrer dans le budget tous les éléments, mais uniquement les éléments, du projet décrits précisément dans les procédures.*
- *Visiter régulièrement les sites des sources de financement.*
- *Encourager les plus jeunes à soumettre leurs projets aux structures adéquates, pour leur permettre plus tard l'accès aux appels plus importants*

Mais aussi :

- *Avoir plus d'un projet dans une même équipe, avec un coordinateur par projet*

Et encore :

- *L'indépendance –de conception, d'analyse, de publication- est sans doute un des ingrédients de la qualité en recherche clinique appliquée.*
- *Toutes qualités requises égales par ailleurs, un peu d'obstination est nécessaire à la construction et réalisation d'un projet.*

Quatrième Partie : L'Exécution du Protocole et la Présentation des Résultats

QUATRIEME PARTIE: L'EXECUTION DU PROTOCOLE ET LA PRESENTATION DES RESULTATS

Il s'est passé du temps entre l'idée initiale et la soumission du protocole écrit aux différents partenaires impliqués dans le projet.

La phase de réalisation proprement dite de l'étude peut également être longue, surtout s'il s'agit d'une étude prospective nécessitant un suivi de longue durée du ou des groupes étudiés. Mais ce suivi sera d'autant plus facile, la mise en route de l'étude et la gestion d'autant plus simples que le protocole aura pu identifier et résoudre par anticipation toutes les difficultés potentielles. C'est redire l'importance de ne pas être avare de son temps lors de la conception d'un protocole.

Que sont les résultats de la recherche ?

- positifs: il y a un effet ou une différence. Celle-ci est-elle intéressante cliniquement ?

- négatifs: l'information ne mérite-t-elle pas cependant d'être soumise à la communauté médicale, pour le bénéfice des patients ?

Une étude de recherche clinique, aussi intéressante soit-elle, n'a de sens que si ses résultats sont rendus publiques. La diffusion de l'information, par le canal de la publication médicale, est l'étape ultime et indispensable de la recherche. Celle-ci ne dépend plus exclusivement du chercheur. Elle aura une bonne chance d'exister si les résultats de l'étude sont pertinents, mais aussi si leur présentation est bonne. C'est là l'aspect délicat de la rédaction médicale, qui s'attache à la forme, et que le chercheur doit maîtriser pour faire passer le fond.

Chapitre XVIII

LA REDACTION MEDICALE

H Maisonneuve

"Ce qui se conçoit bien s'énonce clairement". Le fond assure la validité d'une communication scientifique; la forme assure la lisibilité et la transmission du message. Le fond et la forme doivent être rigoureux.

Que désire un chercheur ? Mener une recherche pour contribuer aux connaissances de la communauté scientifique. Il faut une bonne idée, un protocole précis et détaillé, et mener la recherche avec rigueur. Il veut en communiquer les résultats à la communauté scientifique. Il désire être lu: il doit être clair, précis, et concis.

La rédaction scientifique procède de la science et non de la littérature. Elle est guidée par des principes, qui relèvent de la rigueur scientifique elle-même. Le bon usage de la langue et le respect des règles grammaticales sont indispensables !

Les journaux ont changé rapidement avec l'Internet depuis 1994 et avec le web 2.0 depuis 2005. En 2009, plus de 500 journaux ont un format électronique sans support papier, et revendiquent un accès gratuit pour les lecteurs. Les réseaux sociaux jouent un rôle dans la diffusion de la science et pourraient mettre en péril des journaux dont le fonctionnement est très critiqué.

www.h2mw.eu pour l'actualité en rédaction médicale.

Plan du chapitre

I - LES DIFFERENTS TEXTES MEDICAUX

- **L'article original** rend compte d'un travail de recherche. Sa structure dite IMRAD résulte de la logique scientifique: Introduction (pourquoi le travail a été fait), Méthodes (comment il a été fait), Résultats (ce qui a été observé) et (And) Discussion (ce que je pense de mon travail). Sont également incorporées des références, et si besoin des figures et des tableaux. La thèse de doctorat a la même structure que l'article original, ses buts étant les mêmes.

- **L'éditorial** est habituellement demandé par le comité de rédaction d'une revue à un auteur faisant autorité sur le sujet. L'auteur y exprime librement son opinion en analysant la littérature et en formulant des hypothèses ou des projets de recherche.

- **Le fait clinique** a pour but de rapporter une observation et de la commenter brièvement. S'en rapproche la discussion anatomo-clinique. La structure IMRAD doit être respectée.

- **La lettre à la rédaction**, brève, concerne un cas clinique court, les résultats préliminaires d'un travail de recherche, ou le commentaire d'un article publié dans la même revue.

- **La revue générale** est une revue aussi complète que possible des connaissances sur un sujet à partir d'une analyse exhaustive des travaux publiés. La mise au point est plus courte, de type "actualisation" d'un sujet. Les revues systématiques doivent être encouragées car plus rigoureuses que certaines revues générales dites ‘autoritaires’. La structure IMRAD doit être respectée. L'auteur doit expliquer comment il a sélectionné les articles référencés.

- **L'analyse commentée** consiste à analyser, puis à commenter, des articles parus au cours des derniers mois dans d'autres revues.

- **L'article didactique** a pour but d'enseigner les lecteurs; il correspond à un enseignement; bien fait, il permet la vulgarisation et la dissémination des connaissances.

Avant d’écrire un article, vous devez choisir le journal, étudier les instructions aux auteurs et formaliser par écrit quels sont les auteurs et dans quel ordre.

II - LE TITRE

Le titre annonce le contenu de l'article avec le maximum de précision et de concision. C'est le premier élément d'attraction du lecteur. Sa rédaction, faite après avoir terminé l'article, doit être attentive. Il doit être court (10 à 15 mots) et précis. Les mots informatifs doivent être placés au début, en position forte. Les expressions inutiles (à propos de, contribution à l'étude de, …) sont à proscrire. Un sous-titre, du type "méthode utilisée", est utile.

III - LES AUTEURS

En théorie l'auteur est celui qui a rédigé le manuscrit. En pratique, un auteur travaille très rarement seul, et ses collègues de travail souhaitent une reconnaissance. Le premier auteur est habituellement le rédacteur de l'article. C'est celui qui a réalisé la plus grande

partie du travail ou qui l'a dirigé. Les auteurs doivent s'être entendus entre co-auteurs avant de commencer la rédaction, et il est prudent de formaliser par écrit ce pacte d'auteurs.

Le nom du chef de service ou de laboratoire peut apparaître puisqu'il a été initiateur du travail, a réuni l'équipe, et obtenu les crédits le rendant possible. Mais le nombre des auteurs doit être limité. Tout auteur doit connaître le travail et être capable de remplacer le premier auteur à une présentation orale du travail. D'après le groupe des rédacteurs de journaux biomédicaux (groupe de Vancouver), pour être auteur, trois conditions sont indispensables: 1) avoir conçu et organisé le travail qui a conduit au texte écrit, ou avoir interprété les résultats, ou avoir participé à ces deux étapes; 2) avoir participé à la rédaction des versions successives du manuscrit, 3) avoir approuvé la version finale. Dans cette conception, le nombre d'auteurs est réduit, et le nombre de personnes citées dans les remerciements est augmenté.

Les controverses sont nombreuses dans les journaux à propos de deux phénomènes connus et dont l'ampleur est mal évaluée. Les auteurs cadeaux (concept de gift ou guest authors) représentent entre 20 et 40 % des auteurs. Il s'agit d'ajouter un auteur qui n'a pas légitimité pour être auteur, et d'avoir un 'renvoi d'ascenseur' en retour. Ces petits services entre amis sont fréquents. Par contre des auteurs sont absents de la liste des auteurs, soit parce qu'ils ont été volontairement enlevés en raison de conflits entre auteurs, soit parce qu'ils sont des auteurs dits fantômes (concept de ghost authors). Principalement dans des domaines industriels, des sociétés de service rédigent les articles qui seront signés par des leaders d'opinion qui n'ont pas le temps de rédiger eux-mêmes des articles. Les leaders d'opinion acceptent cette pratique (contre dédommagement), alors que les journaux voudraient rendre cette pratique transparente en publiant tous les noms des rédacteurs, car ils ne peuvent pas l'éliminer.

IV - L'INTRODUCTION

L'introduction de l'article original est un pont entre les connaissances de l'auteur et celles du lecteur. Elle donne au lecteur une idée concise et claire du sujet afin qu'il comprenne pourquoi le travail a été fait. L'intérêt du travail est mis en valeur, afin que le lecteur ait envie de poursuivre sa lecture.

L'introduction est écrite en connaissant la revue à laquelle l'article sera proposé. Les éléments nécessaires et suffisants à la compréhension du travail diffèrent selon le public auquel on s'adresse. La quantité d'information dans une introduction est inversement proportionnelle au niveau supposé des connaissances des lecteurs. Il convient également de moduler l'introduction en fonction des usages de la revue, qu'on appréhende en lisant la revue.

Une bonne introduction devrait avoir 3 paragraphes : dire ce qui est connu sur le sujet en exposant des points précis. Il n'y a pas d'historique ou de revue pédagogique. La seconde partie doit préciser ce qui n'est pas connu, donc un aspect particulier du problème qui sera abordé dans le travail original. La troisième partie est, en une ou deux phrases, la question qui indique le but du travail. Le temps des verbes doit être le passé si on cite un autre auteur, et le présent pour l'exposé des faits admis et/ou prouvés. Toute affirmation doit se fonder sur une ou plusieurs références, mais celles-ci ne doivent pas être trop nombreuses.

Des revues soutiennent qu'il faut dès l'introduction donner un aperçu des résultats. Ceci est controversé, et la plupart considèrent que les résultats ne doivent être exposés que dans la section correspondante.

V - LES METHODES

Ce chapitre d'un article original comporte l'exposé des méthodes de travail. Il doit être suffisamment précis pour que le lecteur puisse reproduire ou vérifier le travail. Dans les descriptions, il faut suivre un ordre logique, qui est habituellement l'ordre chronologique : caractères cliniques avant les examens radiologiques ou biologiques ; critères de jugement précoces avant les critères tardifs.

Le premier objectif (méthodes de sélection) du chapitre est d'indiquer sur qui ou sur quoi a porté le travail: population de malades, animaux, souche cellulaire. Il convient de donner tous les détails nécessaires à l'interprétation des résultats. La description doit indiquer les critères d'inclusion et d'exclusion de la population étudiée, et la période pendant laquelle les malades ont été observés. Les résultats en dépendent. Le chapitre doit préciser s'il s'agit d'une série consécutive ou non, d'un travail prospectif ou rétrospectif, randomisé ou non, ouvert ou en insu. Pour toute étude prospective expérimentale chez l'homme il est nécessaire de faire état de l'accord d'un comité d'éthique. Une étude sur une lignée cellulaire doit préciser son mode d'obtention et de conservation. Dans les études sur l'animal, les lots d'animaux doivent être décrits.

Le deuxième objectif (méthodes d'intervention) est de préciser ce qu'on teste: action d'un médicament, résultats d'une intervention chirurgicale, valeur d'un examen biologique ou radiologique, … Pour un médicament, la dose, le mode et les horaires d'administration sont précisés. Pour des techniques chirurgicales, des examens biologiques, ou des méthodes expérimentales connues, on peut seulement donner la référence de la description initiale. Quand la technique ou la méthode est nouvelle, elle doit être décrite avec précision. Des réactifs ou des substrats doivent être désignés par leur nom chimique, leur origine doit être spécifiée. Pour des appareils, le type, l'origine, et le nom du fabricant doivent être indiqués.

Le troisième objectif (méthodes d'évaluation) est de décrire les critères de jugement, et les méthodes utilisées pour valider les résultats, tests d'inférence statistique par exemple. La description des critères de jugement doit être précise: un amaigrissement doit être chiffré, une diarrhée quantifiée. Si le critère de jugement est un résultat éloigné, il faut indiquer comment seront pris en compte les sujets exclus de l'analyse, en donnant les raisons sans commentaires, et de même pour les perdus de vue. Dans la survie des malades opérés, il faut dire si la mortalité opératoire a été incluse ou non. Si le critère de jugement est une évaluation d'ordre biologique, il faut préciser la nature du prélèvement (sang, plasma, …) et les unités de mesure. Les méthodes statistiques doivent être décrites, et en dehors des tests très usités, une référence doit être citée.

Il existe une terminologie biomédicale internationale. Pour citer un médicament, l'emploi de la dénomination commune internationale est préférable, sans majuscule. Si le nom commercial est utilisé, il doit être suivi du sigle ® (pour registered). Un colibacille est en réalité un *Escherichia coli*, le maxillaire inférieur est la mandibule. Le nom d'une bactérie ou d'un animal comportant deux mots latins est écrit en italique. Pour les microorganismes, seul le premier mot a une majuscule. Pour les animaux, on

met une majuscule si on désigne l'espèce (le Rat, le Hamster), mais pas si on désigne les animaux utilisés (les rats ont été anesthésiés).

Il ne faut pas dans le chapitre des méthodes introduire des commentaires ou des résultats. Toutes les données marginales sans rapport direct avec le travail doivent être éliminées. Tous les verbes doivent être au passé.

VI - LES RESULTATS

Ce chapitre est le coeur de l'article original. Les résultats exposés sont l'aboutissement de la recherche décrite dans l'introduction, et des méthodes employées pour y parvenir; ils sont la base de la discussion. Il convient de rapporter tous les résultats dans ce chapitre et seulement lui. Une faute est de faire figurer des résultats dans la discussion. Il ne faut pas donner des résultats qui ne sont pas en rapport avec le but du travail. Il convient de ne rapporter que des résultats: ce chapitre ne doit comporter aucun commentaire, explication, comparaison avec d'autres travaux. Il ne doit donc comporter aucune référence, car seuls les résultats des auteurs sont exposés.

Une difficulté créée par la structure des comptes rendus de recherche est le risque de faire des répétitions en rappelant des résultats dans le chapitre Discussion. Ce risque est limité par l'utilisation de figures et de tableaux. Dans la Discussion, les résultats sont commentés en faisant référence aux figures et tableaux, et donc sans répéter ces résultats. L'autre avantage des figures et des tableaux est de fournir le maximum d'informations dans un minimum de place, sous une forme synthétique et claire. Le texte ne doit pas répéter les données fournies par les figures et les tableaux.

Comme dans le chapitre précédent, le temps des verbes doit être le passé. La précision doit se traduire dans la cohérence des nombres, en s'assurant notamment que les totaux sont bien égaux à la somme des parties, aussi bien dans le texte que dans les figures et tableaux. Le même nombre de décimales, et les mêmes unités sont utilisés pour un paramètre. La clarté impose de suivre un ordre rationnel dans l'exposé des résultats: immédiats puis tardifs, simples puis compliqués, normaux puis anormaux.

VII - LES ILLUSTRATIONS

Les termes de « figure » et de « tableau » ne sont pas synonymes. Le tableau est composé en caractères d'imprimerie. La figure est faite de tous matériaux qui ne peuvent être transcrits en caractères d'imprimerie. Mais les illustrations ont évolué très vite avec l’apparition des outils électroniques. Les clips vidéos avec son, les podcasts sans images sont acceptés, voire demandés par des journaux. La technologie liée à ces outils n’est pas exposée dans ce chapitre.

Entre les deux modalités, que choisir ? Le choix entre figure et tableau dépend en partie de l'objectif visé: il répond à une intention précise. Chacun a des avantages et des inconvénients. Le tableau a l'avantage de la précision mathématique, au prix de l'aridité. Le lecteur peut refaire les tests statistiques, ou comparer les données à celles d'autres auteurs mais le message est moins facilement appréhendé que sur une figure. La perte d'information d'une figure est acceptable si elle est compensée par l'utilisation d'indices statistiques, tel l'écart-type, qui indiquent la dispersion des valeurs individuelles. Le message est plus facile à transmettre dans une figure. Un tableau est rarement trompeur; une figure peut tromper: il faut toujours bien identifier les échelles.

L'auteur doit se conformer à la présentation adoptée par la revue à laquelle il destine son article, en lisant les recommandations aux auteurs. Il existe des principes généraux. Les figures et tableaux doivent avoir une autonomie d'information: des légendes, titres, et notes permettent de les lire sans l'aide du texte. Si une abréviation est utilisée, elle doit être expliquée dans une note. Les figures et tableaux doivent être appelés dans le texte et numérotés dans leur ordre d'apparition. Ils doivent être préparés avant de rédiger le texte. Le maximum de données est exprimé de façon précise et claire, le texte apporte des informations complémentaires.

La reproduction d'une figure ou d'un tableau issu d'une autre publication nécessite l'autorisation du titulaire du "droit d'auteur", l'auteur ou plus souvent l'éditeur de la publication. La légende précise: "reproduit avec l'autorisation de …".

A - Les figures

L'excellente qualité des figures rend un article plus attractif. Chaque figure doit avoir une légende, imprimée immédiatement au-dessous d'elle. Cette légende est dactylographiée sur une feuille à part, placée à la fin du manuscrit, sur laquelle toutes les légendes de toutes les figures de l'article sont regroupées. Les figures doivent être adressées à la revue sous forme de photographies sur papier glacé, ou plutôt sous format électronique type jpeg. Pour identifier une figure non électronique, il est conseillé d'écrire sur une étiquette adhésive son numéro, son orientation, et les deux premiers mots du titre. Les documents ne doivent être ni pliés, ni rayés, ni agrafés.

- **La présentation "en camembert" e**st adaptée aux pourcentages. Elle permet d'avoir une information précise sur l'importance respective des différentes parties d'un ensemble. Pour une meilleure clarté, il est conseillé de ne pas dépasser sept secteurs, et d'éviter de représenter des secteurs de moins de 5 %, qu'on peut aussi décaler pour mieux les mettre en évidence.

- **L'histogramme** est constitué de barres ou de rectangles verticaux ou horizontaux. Il permet une comparaison statistique de différents nombres. Le nombre de rectangles ou de barres ne doit pas dépasser sept. L'histogramme gagne en précision si on adjoint une échelle. Les barres peuvent être juxtaposées, superposées, ou complètement dissociées. Des traits au sommet des barres permettent de faire figurer l'écart-type des données.

- Dans les **graphiques de type "nuages de points" et les tracés de courbes**, la variable en abscisse est habituellement la variable contrôlée, et la variable en ordonnée la variable "expliquée". Les nuages de points, non reliés par des courbes, sont très utiles lorsqu'on souhaite faire apparaître concrètement des mesures individuelles, en particulier leur distribution ou l'existence ou non d'une corrélation. La droite de régression ne doit être représentée que si la corrélation est statistiquement significative; elle ne doit pas déborder le nuage de points. Le minimum et le maximum des échelles doivent être choisis en fonction des valeurs des deux variables, pour utiliser au mieux l'espace du graphique. Une rupture d'échelle ou une transformation mathématique (par exemple logarithmique) sont parfois utiles. Une rupture d'échelle doit être indiquée sur l'axe lui-même par un double trait incliné //. Il en est de même lorsque le zéro n'est pas l'origine d'un axe. L'extrémité des axes ne doit pas se terminer par une pointe de flèche. Le trait des axes doit être plus fin que celui des courbes.

• **Les courbes** représentent de façon dynamique, par opposition au caractère statique de l'histogramme, l'évolution d'une variable en fonction d'une autre. Les points qui résultent d'une mesure sont graphiquement plus importants que la ligne qui les relie. Divers symboles peuvent être utilisés pour différencier plusieurs courbes, que ce soit les points ou les lignes. Il ne faut pas mélanger les deux systèmes dans une même figure; il faut veiller à l'homogénéité de la présentation sur toutes les figures. Les points doivent être accompagnés de leur indice de dispersion statistique. On peut porter sur une même figure l'évolution de deux variables en fonction d'une troisième; la figure comporte alors deux axes d'ordonnées.

• **Les illustrations** sont des photographies de radiographies, de coupes histologiques, d'enregistrements (électrocardiographiques, …) transmises sous format électronique le plus souvent. On peut utiliser des symboles en surimpression, qui doivent nettement contraster avec le fond. Un schéma d'accompagnement peut être utile. La légende précise les caractéristiques techniques du type facteur multiplicateur (oculaire, objectif du microscope, et objectif de l'appareil photographique). Elle précise les colorations pour un cliché histologique, les paramètres techniques pour une technique d'imagerie. Les photos sur lesquelles un patient peut-être reconnu doivent être accompagnées d'une autorisation. Les journaux n'acceptent plus de masquer les yeux ou voiler les signes de reconnaissance, car cela n'a jamais eu l'effet attendu.

B - Les tableaux

Chaque tableau doit être dactylographié sur une page séparée.

Les différentes parties d'un tableau sont: le titre, la souche - en haut et à gauche du tableau, qui doit rester libre et ne pas être utilisée comme en-tête -, les têtes de colonnes, les têtes de lignes, le corps ou le champ du tableau, et les notes en bas de tableau.

Un tableau ne nécessite pas plus de trois traits horizontaux pour en délimiter les différentes parties: un sépare le titre et les têtes de colonnes, un est au-dessous de celles-ci, un troisième est au-dessous du corps de tableau, le séparant des notes en bas de tableau. Le recours à des traits verticaux est déconseillé.

Les instructions aux auteurs des revues médicales ne précisent pas toujours le format des tableaux. La consultation de quelques exemplaires de la revue permet de savoir si les tableaux sont édités sur une demi-page (il ne faut pas dépasser 60 caractères ou intervalles par ligne) ou une page entière (représentant 120 caractères). Si le nombre de colonnes est supérieur au double du nombre de lignes, il est conseillé d'inverser lignes et colonnes, bien que ceci ne soit pas toujours souhaitable, puisque les têtes de colonnes correspondent aux variables mesurées, et les têtes de lignes aux variables contrôlées.

La composition d'un tableau doit être logique, c'est-à-dire respecter le mode général de lecture, de gauche à droite et de haut en bas. Chaque tableau a un titre, dactylographié et imprimé au-dessus de lui. Ce titre doit être informatif et respecter le principe de la position forte, c'est-à-dire mettre en début de phrase ou de titre les mots informatifs. Il ne faut pas commencer le titre par des mots inutiles ou redondants. Il faut éviter de répéter les informations présentes dans les têtes de colonnes et de lignes. Des tableaux montrant des données comparables doivent être cohérents et utiliser les mêmes mots, dans le même ordre, et les mêmes unités. Souvent 2 ou 3 tableaux peuvent être regroupés en un seul, plus informatif et plus lisible.

Chaque tête de colonne ou tête de ligne désignant des valeurs numériques doit indiquer l'unité de mesure appliquée aux données situées dans la colonne ou la ligne. Les unités ne doivent pas figurer dans le champ du tableau. Si l'unité est la même pour toutes les colonnes, il peut être judicieux de la faire figurer dans le titre. Lorsque plusieurs têtes de colonnes appartiennent à un même ensemble d'informations, il est souhaitable de les regrouper par un trait horizontal situé au-dessus d'elles, et surmonté d'une tête de colonne désignant l'ensemble.

Le corps d'un tableau ne doit comporter que des nombres, ou des sigles simples (M pour masculin, F pour féminin par exemple). Les nombres dans les colonnes doivent être alignés sur la virgule s'il y a une décimale (ou le point dans les articles en anglais), et comporter le même nombre de décimales pour des variables identiques. Les nombres inférieurs à 1 doivent comporter un zéro avant la virgule ou le point. Il doit y avoir un deuxième alignement sur le sigle ±, et un troisième alignement sur la virgule de l'écart type.

Lorsqu'une donnée manque, un symbole, défini dans une note en bas de tableau, vient remplir la case. Les notes en bas de tableau renseignent le lecteur sur les abréviations utilisées dans le tableau. Lorsque deux tableaux comportent les mêmes abréviations, on peut se contenter d'indiquer dans un des tableaux: "mêmes abréviations que dans le tableau …". Les notes en bas de tableau doivent être appelées par des symboles situés dans le tableau lui-même, l'ordre d'appel se faisant de gauche à droite et de haut en bas. On peut utiliser des lettres minuscules entre parenthèses, mais il est préférable de se servir des symboles, qui par convention doivent être appelés dans l'ordre suivant: *, †, ‡, §, ‖, ¶. Si plus de symboles sont nécessaires, on peut utiliser les mêmes en les doublant.

VIII - LA DISCUSSION

Avec le résumé, ce chapitre est celui dans lequel les auteurs français ont le plus de mal à faire abstraction de leurs pulsions pédagogiques. L'erreur la plus fréquente est de discuter l'ensemble du sujet et non pas le travail lui-même.

Le but de la discussion est d'interpréter le travail qui a été réalisé, et lui seul, c'est-à-dire les moyens mis en oeuvre, la méthode de travail, et les résultats. La discussion diffère des chapitres précédents dans sa conception: il convient d'exprimer personnellement ce qu'on pense. La qualité et l'intérêt de la discussion reflètent la culture scientifique et l'intelligence des auteurs. Ce chapitre répond à trois objectifs.

Le premier objectif est de dire si le but du travail exposé à la fin de l'introduction a été atteint ou non. Cela implique de résumer les principaux résultats, seule redite acceptable dans un compte rendu de recherche. Aucun résultat nouveau ne doit apparaître.

Le second objectif est de juger la qualité et la validité des résultats. La discussion critique et objective du travail porte sur chacun des chapitres de l'article, en identifiant notamment les biais qui ont pu intervenir. Le nombre de sujets a-t-il été suffisant ? Y-a-t-il eu un biais dans la sélection des sujets ? La méthode de travail était-elle la mieux adaptée à la question ? Cette partie ne doit pas être une autocritique trop sévère, qui conduirait au refus de l'article, son but est d'aller au devant des critiques en expliquant les choix que l'on a faits.

Le troisième objectif est de comparer les résultats à ceux d'autres auteurs. S'il y a des différences, il faut essayer de les expliquer. Les auteurs ont ici à faire part de leur apport personnel dans la manière dont ils ont abordé le problème. Il ne s'agit pas de faire une revue de la littérature.

Un paragraphe peut aller au devant des interrogations des lecteurs : avez-vous observé les résultats attendus ? Avez-vous des hypothèses ou mécanismes expliquant les observations ? Avez-vous changé des pratiques depuis la fin de ce travail ? Y-a-t-il des risques de santé publique nécessitant des précautions ? Chaque sujet doit vous faire évoquer des questions spécifiques pour lesquelles le lecteur aimerait connaître votre opinion.

Comment construire une discussion ? L'usage se développe de commencer par son premier objectif. Aucune règle ne propose un ordre dans lequel les éléments d'une discussion doivent être présentés. Des essais successifs de rangement des paragraphes sont indispensables pour un exposé logique et clair. Certaines revues tolèrent que l'article se termine par une conclusion, mais ce n'est pas recommandé: une conclusion risque d'être une redite, ou une tentative de sauvetage d'une discussion mal élaborée. La discussion ne doit pas non plus se terminer par un résumé. Des hypothèses pour un travail futur peuvent être formulées.

Une autre erreur est de répéter dans la discussion ce qui a été dit dans l'introduction. Le principe est de rappeler dans l'introduction l'état de nos connaissances, et dans la discussion de confronter ses résultats avec ceux d'autres auteurs. Une autre erreur est l'inexactitude des citations, dans la transcription des résultats des autres auteurs, ou dans ce qu'on leur fait dire. Pour éviter cette erreur, il ne faut jamais citer des auteurs sans avoir lu l'article original. Il ne faut pas non plus citer un auteur sans donner de référence. Les expressions émotionnelles doivent être bannies. Le temps présent ne doit être utilisé que pour les notions bien établies. Si la discussion dépasse la moitié de la longueur totale de l'article, elle est trop longue et probablement mal conduite.

IX - LE RESUME

Son but est de présenter au lecteur, dans un espace réduit, la substance des informations de l'article. C'est la partie d'un article la plus lue. C'est, avec le titre, ce qui incitera le lecteur à lire tout l'article. Il est susceptible d'être reproduit dans de nombreux documents sans l'article. Il doit donc être compréhensible en lui-même. Le résumé doit être informatif. Sa construction reprend la structure IMRAD, et répond aux quatre questions: pourquoi, comment ce travail a-t-il été fait, quels ont été les résultats, quelles conclusions ou généralisations peut-on en tirer ?

Le résumé d'un article original n'est pas une épreuve de synthèse comme dans le cas des études de sciences politiques. C'est un exercice de copier / coller à partir de 4 éléments de votre article : 1) la dernière phrase de l'introduction ; 2) la première phrase de chaque paragraphe des méthodes ; 3) une phrase par illustration pour les résultats ; 4) la première phrase de la discussion. Le résumé ne demande aucune réflexion ou synthèse afin d'exposer exactement les données de l'article.

Le résumé ne doit pas contenir d'appel à des références, de figures, tableaux, ou notes, voire des abréviations non expliquées dans le résumé lui-même. Sa longueur est souvent indiquée dans les instructions aux auteurs, souvent 250 à 300 mots, soit environ une

page dactylographiée en double interligne. La plupart des revues ont adopté des résumés dits structurés. Six à huit paragraphes sont identifiés: buts de l'étude, protocole, méthode de l'étude, lieu de l'étude, méthodes de sélection, méthodes d'intervention, méthodes d'évaluation, critère de jugement principal, résultats, conclusions. Les instructions aux auteurs du JAMA expliquent en une page comment écrire un tel résumé. Ces résumés sont très longs, très détaillés.

X - LES REFERENCES

Leur but est de justifier tout fait énoncé. C'est un principe fondamental de la recherche scientifique. La référence est appelée le plus tôt possible après l'énoncé du fait, elle n'est pas obligatoirement à la fin d'une phrase. Une même référence peut être appelée plusieurs fois dans l'article. Il ne doit pas y avoir de référence dans le chapitre des résultats. Les références sont données à la fin de l'article, et doivent être distinguées de la bibliographie. Le chapitre des références contient la liste des articles cités dans le texte. Une bibliographie concerne l'ensemble des articles et livres écrits sur un sujet précis. Le terme "références bibliographiques" est impropre. Beaucoup de revues limitent le nombre de références, sauf pour les articles de revue générale. Les rédacteurs des revues contrôlent l'appel des références: les erreurs y sont extrêmement fréquentes, de deux ordres: erreurs de transcription du libellé de la référence, erreurs dans la citation du contenu de la référence.

Les références qui ne sont pas accessibles au lecteur doivent être évitées: thèse, résumé de congrès lorsqu'il n'est pas publié dans un supplément de périodique, communication personnelle, article sous presse ou soumis pour publication, communication orale non publiée, référence de seconde main.

Les recommandations aux auteurs des revues indiquent le système qu'elles utilisent. Trois sont principalement utilisés: le système auteur-année, appelé système Harvard; le système numérique séquentiel, dont la variante connue est le système de Vancouver; le système alphabétique-numérique, qui est hybride.

A - Le système Harvard

Dans le corps du texte, le ou les auteurs sont cités, avec l'année de la publication, une lettre suivant l'année si plusieurs articles des mêmes auteurs cités ont été publiés la même année. L'auteur et l'année peuvent être cités entre parenthèses dans le corps ou à la fin de la phrase. Dans la liste des références, les références sont classées alphabétiquement sans numéro d'ordre. La dactylographie des références est précise pour l'utilisation de l'italique, des graisses de caractères, et des sigles (&, et al., …). Ce système adopté par de nombreuses revues, essentiellement britanniques, diminue la lisibilité du texte, mais il est apprécié par bon nombre d'auteurs et de lecteurs. Il est pratique lors de la préparation d'un manuscrit, car une référence oubliée peut être facilement introduite sans devoir décaler le numéro des références suivantes.

B - Le système numérique séquentiel

Dans le corps du texte, les références sont numérotées par un nombre arabe selon leur ordre d'apparition. Si une référence est citée plusieurs fois, elle garde son premier numéro d'appel. Les numéros sont cités entre parenthèses. Si plusieurs références sont citées dans la même parenthèse, elles sont classées par ordre croissant et séparées par des virgules. Si plusieurs références successives sont citées dans la même parenthèse,

seules la première et la dernière figurent, séparées par un trait d'union. Dans la liste des références, les références sont dans leur ordre d'appel dans le texte. La dactylographie est simplifiée, avec seulement six auteurs au maximum, avec des abréviations pour les titres des revues et une ponctuation spécifique. Ce système facilite la lecture. Il est recommandé par de nombreux rédacteurs de revues internationales. L'inconvénient majeur est pour les auteurs la renumérotation nécessaire, donc le risque d'erreur, en cas d'introduction d'une référence nouvelle.

C - Le système alphabétique-numérique

Dans le corps du texte, les références sont citées par leur numéro d'ordre, indiqué entre parenthèses. Dans la liste des références, les références sont classées par ordre alphabétique, et le numéro d'ordre est attribué selon ce classement. Ce système est en voie d'abandon.

Quel que soit le système, il convient de transcrire les références selon les recommandations de la revue. Certains ouvrages précisent comment transcrire une référence. Il est conseillé d'envoyer un article à une revue avec les références dactylographiées selon les recommandations de cette revue.

Pour référencer un article de périodique, il faut respecter l'ordre suivant: les auteurs sont tous cités jusqu'à six auteurs, suivis de "et al" s'il y en a plus de six. Le nom des auteurs, dont seule l'initiale est majuscule, est suivi des initiales du prénom, en majuscules, contiguës, et d'une virgule; le dernier nom est suivi d'un point. Le titre de l'article est transcrit dans sa langue originale (précisée entre crochets si elle est différente de la langue de l'article envoyé à la revue), et est suivi d'un point. Le nom de la revue est indiqué en abrégé selon l'Index Medicus, et est suivi, sans point, de l'année de publication, suivie d'un point virgule, puis du numéro du tome ou volume, suivi de deux points, puis de la première page de l'article, séparée de la dernière page par un tiret, dont seuls les chiffres différents de ceux de la première page sont transcrits. Il n'y a aucun espace entre les divers nombres contrairement aux règles habituelles de dactylographie. Si la référence provient d'un supplément, la mention "[suppl …]" suit le numéro de volume. Il n'y a pas de retour à la ligne après chaque élément des références.

La référence d'un livre doit comporter dans cet ordre le nom et les initiales du prénom des auteurs, le titre du livre, le numéro de l'édition (à partir de la seconde), la ville et le nom de la maison d'édition, l'année de l'édition, et le nombre de pages ou les pages exactes (première et dernière) à consulter. Pour référencer un chapitre dans un livre, si les auteurs de chaque chapitre sont identifiés, la référence comporte les noms et les initiales des prénoms des auteurs suivis d'un point, puis le titre du chapitre suivi d'un point, puis la mention "Dans" ou "In" suivie de deux points. Viennent ensuite les noms et initiales des prénoms des rédacteurs du livre, suivis de "ed" ou "eds", suivi d'un point. Le titre du livre est ensuite transcrit en entier dans sa langue originale, suivi d'un point. La ville puis le nom de la maison d'édition sont cités, puis l'année de l'édition, et la première et la dernière page du chapitre.

Les références à des liens électroniques sont acceptées et beaucoup de règles ont été proposées. Il est établi que leur demi-vie serait de 4 ans environ dans les grands journaux. Il est souhaitable qu'elles soient un complément et ne représentent pas la totalité des références. Il est d'usage de préciser la dernière date de consultation, et surtout de toutes les vérifier sur la dernière version d'un article.

La qualité des références est mauvaise en général car les auteurs ne font pas suffisamment attention à leur sélection et ne les lisent pas complètement. On évoque le rôle social des références qui doivent plaire aux collègues que l'on cite et au journal qui veut augmenter son facteur d'impact en s'auto-citant. Il a été montré que ces citations inadéquates étaient à l'origine de biais et aussi de transformation d'hypothèses en faits acquis. C'est regrettable et cela nuit à la qualité de la science et aux malades.

Exemples de mode de dactylographie des références dans les 2 systèmes les plus répandus.

Système numérique (Vancouver)

1. Greenberg SA. How citation distortions create unfounded authority : analysis of a citation network. BMJ 2009 ;339 :b2680. doi 10.1136/bmj.b2680

2. Sox HC, Rennie D. Research misconduct, retraction, cleaning the medical literature: lessons from the Poehlman case. Annals of Internal Medicine. 2006;144:609-613.

3. Goetzsche PC, Kassirer JP, Woolley KL, Wager E, Jacobs A, Gertel A, et al. What should be done to tackle ghostwriting in the medical literature? PLoS Med 2009;6(2): e1000023. doi:10.1371/journal.pmed.1000023

Système alphabétique (Harvard)

Goetzsche PC, & al. (2009) What should be done to tackle ghostwriting in the medical literature? PLoS Med 6(2): e1000023. doi:10.1371/journal.pmed.1000023

Greenberg SA (2009). How citation distortions create unfounded authority : analysis of a citation network. BMJ 339 :b2680. doi 10.1136/bmj.b2680

Sox HC & Rennie D. (2006) Research misconduct, retraction, cleaning the medical literature: lessons from the Poehlman case. Annals of Internal Medicine 144:609-613.

XI - LE STYLE

Le style scientifique diffère du style littéraire. La logique scientifique impose pour les verbes d'utiliser les temps du passé pour tous les événements survenus dans le passé, et de n'utiliser le présent que pour les notions bien établies. Le temps du futur n'est pas utilisé en rédaction scientifique. Le passif de modestie amène l'ambiguïté et doit de même être proscrit. Il faut utiliser 'je' ou 'nous' pour éviter ces passifs de modestie qui sont imprécis.

Même si nous avons appris à éviter l'emploi du même mot à des intervalles rapprochés, et donc à rechercher des variations élégantes, la rigueur scientifique implique d'utiliser

le même mot pour désigner la même chose. Les expressions émotionnelles ou de courtoisie doivent être évitées.

- **La précision,** toujours présente pendant le déroulement de la recherche, doit aussi guider la rédaction de l'article. Une méthode expérimentale doit être décrite assez précisément pour être reproductible par le lecteur. Un amaigrissement de 10 kg n'a pas la même signification chez un sujet de 100 kg ou de 50 kg, ou s'il s'est produit en un mois ou un an. La précision impose de vérifier tous les nombres, et leur cohérence entre le texte et les tableaux. Les totaux doivent être exacts. Adjectifs et adverbes imprécis ou inutiles (examen attentif, grosse tumeur, étude récente, souvent, beaucoup, …) doivent être supprimés.

- **La clarté** est la deuxième vertu d'un article scientifique. Elle implique d'utiliser des mots et une syntaxe simple. Elle est améliorée par la mise en position forte - au début d'une phrase, d'un paragraphe, d'un titre - des mots les plus informatifs. La virgule clarifie le sens d'une phrase: que déduire de la phrase suivante: "les trois malades avaient respectivement douleurs, nausées et vomissements et diarrhée" ? La virgule, placée après "nausées" ou après "vomissements" lèverait l'ambiguïté. Contrairement à l'usage littéraire, il convient de mettre une virgule avant le "et"; cette politique a été adoptée par les plus grands journaux biomédicaux. "*Et caetera*", comme "tel que" et "par exemple" sont imprécis, sauf si logiquement le lecteur peut déduire tout ce qui est sous-entendu.

Les abréviations internationales d'unités sont licites, voire recommandées, quand elles suivent un nombre, mais pas dans les autres cas. L'orthographe d'une abréviation doit être vérifiée si elle est incertaine. Une abréviation d'unité est invariable. L'intérêt des abréviations est de raccourcir un texte, et de le rendre plus lisible en se substituant à une expression, ou un mot trop long, employés souvent. Leur usage ne doit pas être abusif. Toute abréviation doit être annoncée au premier usage, même si elle paraît évidente.

- **La concision** est la troisième qualité d'un article scientifique. Noms, adverbes, adjectifs et expressions creux (il va sans dire que, il est opportun de signaler que, …) doivent être supprimés, comme les données marginales. La répétition va à l'encontre du principe de concision, et doit être évitée, hormis la seule autorisée, et même nécessaire, du résumé par rapport au texte. Il faut éviter la répétition du titre dans le résumé, celle des résultats ou des parties de l'introduction dans la discussion. Mais il faut éviter aussi un excès de concision: l'ellipse nuit à la clarté.

XII - DE LA PREPARATION D'UN MANUSCRIT A LA CORRECTION DES EPREUVES

Le choix de la revue est guidé par l'adéquation entre le sujet de l'article et les objectifs de la revue. Il est indispensable de consulter et de connaître une revue avant de lui soumettre un article. Un autre élément du choix est la diffusion du journal. Un auteur a intérêt à envoyer son article à une revue de diffusion large, mais qui risque plus de le refuser. Avant d'écrire, il est nécessaire de disposer des éléments d'information qui sont la substance de l'article, que ce soit le matériel et les méthodes du travail, les résultats, et la bibliographie dans laquelle sont choisies les références. Il faut avoir à portée de la main un exemplaire de la revue et ses recommandations aux auteurs.

Dans quel ordre écrire ? Il faut terminer par la rédaction du titre et du résumé. La discussion est écrite après les chapitres des méthodes et des résultats. Ces deux derniers chapitres et l'introduction sont écrits dans un ordre variable selon les préférences de chacun.

Une fois écrite, une première version n'est jamais la dernière. Lors de la préparation du manuscrit, les références sont mises entre parenthèses en utilisant le système Harvard (noms des auteurs dans le texte). Lorsqu'on est certain de ne plus modifier le texte, les noms des auteurs sont remplacés par des numéros selon les souhaits de la revue.

Il existe des règles de rédaction concernant les nombres. Un nombre est écrit en lettres s'il débute une phrase, en chiffres dans le corps de la phrase. Les nombres inférieurs à 10 sont écrits en lettres, les autres en chiffres. La règle vaut aussi pour les nombres ordinaux. Dans une série comportant à la fois des nombres inférieurs et supérieurs à 10, tous les nombres sont écrits en chiffres. Les chiffres sont utilisés pour écrire les dates et devant une unité de mesure. Selon les revues, les nombres de 4 chiffres sont écrits avec ou sans espace (virgule en anglais) entre le troisième et le quatrième chiffre; à partir de cinq chiffres, des espaces (ou des virgules) sont systématiques. Il existe de nombreuses recommandations pour l'utilisation de la ponctuation, tirets, espaces, traits d'union, … et des ouvrages exhaustifs dans ce domaine. La dactylographie doit être adaptée à celle de la revue. Lorsque l'auteur a rédigé sa version "définitive", qui peut être la sixième, la huitième, ou même plus, il la soumet à ses co-auteurs qui la revoient sur le fond et sur la forme.

Les soumissions par voie électronique feront une partie du travail décrit dans ce paragraphe. Sur la page de titre figurent le titre de l'article et les noms des auteurs, les prénoms en entier ou leurs initiales précédant les noms (à la différence de la section des références). Si les auteurs ont des adresses différentes, il faut après chaque nom un astérisque ou un numéro, pour indiquer l'adresse de chacun. Le nom, l'adresse, et le numéro de téléphone de l'auteur qui assure la correspondance, et de celui qui envoie les tirés à part, sont précisés. Les institutions qui ont apporté un support financier doivent aussi être mentionnées, avec les références du contrat.

Avant d'adresser l'article à la revue, sept derniers contrôles sont conseillés:

- le temps des verbes des chapitres Méthodes et Résultats est au passé; le temps des verbes des chapitres Introduction et Discussion est au passé et parfois au présent (généralisation);

- tous les mots, adjectifs et adverbes creux sont supprimés;

- les références, les tableaux, et les figures sont tous appelés dans le texte;

- les chapitres Méthodes et Résultats ne contiennent aucun commentaire;

- les figures et les tableaux sont compréhensibles en eux-mêmes;

- les totaux sont cohérents;

- pour un article en français, les mots utilisés sont susceptibles d'être compris par un lecteur étranger francophone.

Après l'envoi, la revue accuse réception de l'article, qu'en général elle envoie à un ou plusieurs experts scientifiques. La réponse parvient à l'auteur après 6 à 10 semaines. Si l'article est accepté avec modifications, l'auteur doit faire ces modifications et accompagner l'article corrigé d'une lettre répondant point par point aux critiques. Plus tard, il ne faut pas négliger la correction des épreuves, lecture comparative avec le manuscrit.

Ainsi l'auteur verra dans un périodique le fruit d'un long labeur, dont la rédaction de l'article aura été une étape pénible mais très enrichissante. Le respect de quelques règles, augmente grandement ses chances d'être publié.

XIII – LES INSTRUCTIONS AUX AUTEURS

Depuis 1978, le groupe dit de Vancouver actualise régulièrement les recommandations standards pour les manuscrits soumis à des journaux biomédicaux. Ce groupe qui s'appelle le Comité International des Rédacteurs de Journaux Biomédicaux comprend les grands journaux internationaux, et surtout les 'big five' (deux anglais, BMJ et Lancet ; trois américains Annals of Internal Medicine, JAMA, NEJM). Leurs recommandations contiennent surtout des principes d'éthique pour tous les acteurs : auteurs, rédacteurs des revues, relecteurs, propriétaires des revues. Elles sont actualisées avec des liens sur un site www.icmje.org.

Les instructions pour soumettre les manuscrits sont un cours de rédaction, avec le format IMRAD comme base. Les autres instructions de ce texte ne peuvent être toutes citées : critères pour être auteur ; liste des contributeurs et personnes à remercier ; rôle du rédacteur et liberté éditoriale ; peer-review (revue par les pairs) ; conflits d'intérêts ; confidentialité et protection des personnes, et animaux ; obligation de publier les études dites négatives ; corrections, rétractations et expressions de doutes ; propriété des articles ; mauvaises pratiques pour les doubles publications, et principes pour accepter des publications secondaires ; correspondance ; suppléments ; publications électroniques ; publicité ; relations avec les médias grand public ; enregistrement des essais cliniques.

D'autres instructions ont été préparées au cours du temps, en commençant par CONSORT qui a aussi un site dédié www.consort-statement.org. Une traduction française par des équipes de Paris 5 est une excellente initiative http://eb.medecine.univ-paris5.fr/moodle/course/view.php?id=2. Nous citons certaines de ces instructions et vous proposons de passer par un moteur de recherche ou le site www.icmje.org pour les localiser. Ce sont CONSORT pour Consolidated Standards of Reporting Trials, STARD pour STAndards for the Reporting of Diagnostic accuracy studies, TREND pour Transparent Reporting of Evaluations with Nonrandomized Designs, MOOSE pour Meta-analysis of observational studies in epidemiology, SQUIRE pour Standards for QUality Improvement Reporting Excellence, EQUATOR pour Enhancing the QUAlity and Transparency Of health Research, et PRISMA pour Preferred Reporting Items for Systematic reviews and Meta-Analyses.

XIV – LES JOURNAUX ELECTRONIQUES

Ces dix dernières années ont modifié l'économie du système sans avoir engendré de grandes restructurations, mais celles-ci pourraient survenir. Les faits marquants à l'origine de ces bouleversements sont : 1) la très forte montée en puissance de l'Internet : les revues dites électroniques ont d'importants avantages compétitifs sur les revues dites papier (par exemple, la possibilité d'inclure des vidéos dans les illustrations ou des podcasts commentant le numéro, etc.) ; le web 2.0 accentue ces différences ; 2) la forte pression pour un accès gratuit au contenu des revues : si pour certains ce principe semble acquis, rien n'est encore définitif. Si ce principe devenait réalité, les ressources des maisons d'édition ne proviendraient plus des lecteurs (particuliers, services hospitaliers, bibliothèques), mais d'autres acteurs qui devront supporter les coûts de l'édition (auteurs, institutions, laboratoires pharmaceutiques, …?) ; 3) la possibilité pour tout internaute (l'e-patient) d'accéder aux mêmes informations que les professionnels de santé. Le public, avec les scientifiques, a contribué à l'augmentation des consultations de la base de données Medline après sa mise en accès gratuit sur internet : 163 000 recherches en janvier 1997, 12,5 millions en janvier 1999, 64,8 millions en janvier 2005, et 82,3 millions en mars 2007 ; depuis 2007, la croissance a cessé ; 4) la possibilité de quantifier de plus en plus finement la notoriété de chaque auteur ; le calcul de la notoriété d'une revue (le fameux facteur d'impact) devrait céder le pas au calcul de la notoriété de chaque article (et donc de leurs auteurs) via le nombre de « hits » (consultation) par exemple pour un article sur un site internet donné. Des listes des articles les plus consultés ou téléchargés sont déjà disponibles sur bon nombre de sites ; la montée en puissance d'outils innovants de communication avec les réseaux sociaux ont été rapidement adoptés par les journaux et les jeunes générations ; beaucoup de journaux donnent accès à des réseaux type Facebook pour discuter des articles.

Il existe trois types de journaux électroniques : 1) les publications électroniques d'un périodique édité sur papier avec toutes ses rubriques, accessibles sur un site internet ; il s'agit en général de journaux qui continuent leur activité éditoriale et ont ajouté la fonctionnalité électronique en mettant sous format pdf téléchargeable leurs articles ; le contenu électronique est identique à celui du journal déjà existant ; ce modèle a été adopté par de nombreux journaux ; 2) les journaux hybrides qui correspondent au format précédent mais avec des rubriques et/ou des données complémentaires qui sont ajoutées sur le site internet. La notion dite ELPS *(Electronic Long / Paper Short,)* dont le BMJ s'est fait le promoteur depuis 1999, se développe : un article long est publié sous forme électronique, un article plus court (préparé par la rédaction) dans le journal imprimé. Le format électronique, plus long que le texte imprimé dans le journal, comporte plus de références, des données complémentaires (fichiers Excel, documents annexes - protocole opératoire, questionnaires, clips vidéos d'une technique chirurgicale, etc..). En 2009, le BMJ a introduit le modèle 'Pico' (Pico ; car c'est une unité petite (10^{-12}) et aussi pour l'acronyme 'Population, Intervention, Comparison, Outcomes'). Il s'agit d'une évolution du format ELPS. L'article de recherche complet, avec ses compléments multimédia, est gratuit et exclusivement électronique ; un format très standardisé de type résumé long est préparé par l'auteur pour la publication papier. Le format papier devient une compilation d'informations et de résumés longs pour les articles originaux ; 3) les "e-journaux" qui n'ont pas de publication sous forme papier et n'existent que sur un site internet ; les premiers journaux de ce type ont été créés vers le milieu des années 1990 (plusieurs dominent comme PLoS et BioMed Central, d'autres arrivent comme Bentham Open). La publication en ligne est faite en temps réel sans attendre les délais nécessaires à une impression papier. Le pdf du manuscrit accepté est parfois mis en ligne avant que l'article ne soit révisé, préparé et mis en page. Les mises en ligne n'ont plus de périodicité régulière (hebdomadaire, mensuelle, etc.) puisque

l'actualisation du site se fait en continu. Lors de la publication, les avis des relecteurs et les versions successives de l'article (de la version soumise jusqu'à la version définitive) peuvent être mises en ligne (sur www.biomedcentral.com par exemple).

XV – UN SYSTEME DE PUBLICATION A L'AVENIR INCERTAIN

Avec Richard Smith, ancien rédacteur du BMJ qui a publié un pamphlet intitulé 'The trouble with medical journals', je crois que les objectifs principaux d'un journal ne sont pas toujours de diffuser de la bonne science qui est d'ailleurs très rare, et que les journaux biomédicaux sont faits pour les auteurs et pas pour les lecteurs. Posséder un journal s'accompagne d'une stratégie propre à des actionnaires ou des leaders d'opinions. Les journaux appartiennent à des sociétés commerciales ou à des sociétés professionnelles (dites sociétés savantes).

Le propriétaire est le décideur et il devrait accorder une liberté éditoriale totale au comité de rédaction, mais ce comité reçoit toujours quelques orientations. C'est l'intérêt du comité de rédaction de faire plaisir au propriétaire. Pour améliorer le fameux facteur d'impact, les stratégies sont nombreuses, et l'objet de discussions peu scientifiques au sein des comités de rédaction. Un journal préfère accepter un article médiocre qui sera diffusé ensuite dans la presse grand public plutôt qu'une étude de qualité, austère, sans retentissement médiatique. Un journal a des compétiteurs : *« acceptons cet article car sinon, il sera publié par un concurrent…. »*. La profitabilité peut mener à accepter des articles qui génèreront des tirés à part vendus à un industriel.

Les 5 acteurs (auteurs, rédacteur en chef, comité de rédaction, relecteurs, propriétaires) ont chacun leurs intérêts particuliers, parfois contradictoires avec ceux des autres acteurs. La plupart ont une position universitaire qui génère la compétition. Je n'ai pas mentionné le plus important des acteurs car son rôle est minime et il n'a aucun moyen de s'exprimer : c'est le lecteur. On ne sait pas s'il lit, on connait peu ses besoins, et d'ailleurs ce n'est pas lui qui paye le système. Ce sont des institutions, voire de plus en plus souvent les auteurs qui payent le système.

Richard Smith estime que les journaux publient de mauvais articles, car la science de qualité est rare, et que tout le système est fait pour les auteurs. Il trouve anormal que des recherches publiques donnent lieu à publication d'articles dans des journaux payants par abonnement ou pay-per-view (à la carte) et milite pour le libre accès. Citons, sans les commenter certains de ses griefs, pour vous laisser juge : le système de peer-review ne marche pas ; les journaux s'intéressent à leur facteur d'impact ; il y a conflit entre la médecine (ne pas nuire) et le journalisme qui consiste à diffuser toutes opinions pour engendrer des dialogues ; l'industrie et autres financeurs influencent trop les journaux ; les conflits d'intérêts ne sont pas gérés ; la fabrication, la falsification de données et le plagiat sont trop fréquents ; les doubles publications sont du gaspillage ; les citations inadéquates arrivent à transformer des hypothèses en faits ; publier est un des critères utilisés pour progresser dans une carrière et les auteurs acceptent tout des journaux ;…

Les journaux biomédicaux ont pour vocation de remplir cinq missions : 1) s'assurer de la meilleure qualité possible des textes publiés grâce au système de *peer review* ; 2) communiquer les résultats originaux des recherches ; 3) réunir des communautés scientifiques et médicales ; 4) archiver le savoir ; 5) œuvrer à la mise en place d'un consensus scientifique autour d'une problématique indépendamment des pressions commerciales, politiques, professionnelles et économiques du moment. Il nous paraît essentiel que ces missions, qui fondent la légitimité de l'existence des revues, perdurent et que les nouvelles technologies, en dépit des changements qu'elles induisent, demeurent au service de la diffusion de l'information, sans en modifier le sens.

Connaissez le système et vous saurez saisir des opportunités pour bien publier vos travaux, en ne perdant jamais à l'esprit que l'acceptation par un journal est autant émotionnelle que scientifique.

Chapitre XIX
Applicabilité des résultats et valorisation de la recherche

P Duhaut, R Chapurlat, J-P Ducroix

Le but de la recherche clinique est, *in fine* , l'amélioration de l'état de santé des patients en passant par la prévention, un meilleur diagnostic, une meilleure thérapeutique permettant de ralentir l'évolution, de préserver l'existant, d'améliorer le présent, voire, parfois, de guérir le patient. Ce chapitre passe en revue les conditions d'applicabilité de la recherche clinique aux patients effectivement pris en charge en pratique médicale quotidienne.
La recherche clinique permet aussi de promouvoir dans la carrière médicale les jeunes collègues s'y engageant et de maintenir chez les médecins praticiens une exigence de progrès, conditions *sine qua non* au maintien de soins de bonne qualité. Ce chapitre revoit également les modes de valorisation de la recherche dans les différents types d'exercice médical.

Le but final de la recherche clinique est d'améliorer la prise en charge des patients dans le domaine diagnostique ou thérapeutique, d'éventuellement prévenir certaines maladies par une meilleure compréhension des facteurs de risque ou l'application de mesures en découlant, de mieux répartir les ressources disponibles en matière de santé pour le bénéfice du plus grand nombre… il s'agit donc essentiellement d'une recherche appliquée se situant à une extrémité du spectre de la recherche, alors que la recherche fondamentale, visant à mieux comprendre le monde environnant sans action immédiate sur celui-ci, se situe à l'autre extrémité.
La recherche clinique se doit d'être applicable, puisqu'il s'agit de son ambition première… qu'en est-il dans la réalité, et quelles peuvent être les conditions de son applicabilité ?

A- Applicabilité et application :

I- Applicabilité, population atteinte et critères d'inclusion ou d'exclusion des patients dans les études :

L'une des premières conditions serait que la recherche clinique soit faite sur des patients similaires à ceux de la pratique quotidienne, puisque ses résultats s'appliqueront, justement, à ces patients. Il s'agit aussi d'une des difficultés majeures : les patients inclus doivent donner leur consentement éclairé, beaucoup d'essais randomisés notamment comportent une limite d'âge d'inclusion, des comorbidités représentent souvent autant de critères d'exclusion pour atteindre une population homogène… alors que les patients soignés quotidiennement sont plus souvent âgés, polypathologiques, et lorsqu'ils sont gravement atteints, peuvent ne pas être en mesure de donner un consentement éclairé. Par contre, lorsque les résultats d'un essai randomisé sont publiés, le message tend à se centrer sur l'efficacité de la thérapeutique, et les critères d'inclusion ou d'exclusion tendent à être oubliés.

La plupart des essais randomisés en cancérologie sont construits sur la base d'une espérance de vie des patients supérieure à 6 mois. De ce fait, les patients avec métastases cérébrales, altération de l'état général, comorbidités, antécédents d'autres cancers, sont le plus souvent exclus. Or, ils sont nombreux dans certains cancers fréquents comme les cancers pulmonaires, pour lesquels ils peuvent représenter jusqu'à 65 % des patients vus en première intention[1].

Les essais randomisés portant sur la thrombolyse dans l'accident vasculaire cérébral (AVC) aigu ont montré un bénéfice -modeste au demeurant- à condition qu'elle soit réalisée dans les trois premières heures suivant le début des symptômes, et que le patient ne présente pas de risque hémorragique évident : ceci ne représente pas plus de 10 % des patients en pratique courante dans de nombreuses régions [2] Si l'on peut espérer une amélioration du pronostic chez une minorité de patients parmi ces 10 %, il sera difficile de mettre en évidence l'impact de la thrombolyse de l'AVC en termes de santé publique.

Les statines représentent en France l'unique famille médicamenteuse classée en Service Médical Rendu 1 (SMR 1), dans une échelle allant de 1 à 5 (1 : service maximal prouvé, 5 : service médical inexistant ou non prouvé). Les anti-tuberculeux, les antibiotiques, les diurétiques de l'anse sont classés en 2, les anti-épileptiques en 3… Il est cependant troublant de constater que toutes les études concernant les statines ont été réalisées dans les pays à forte mortalité et morbidité cardiovasculaire, essentiellement les pays du Nord de l'Europe, alors qu'aucune étude n'a été réalisée dans les pays à faible mortalité cardiovasculaire comme la France ou d'une manière générale les pays du pourtour Nord de la Méditerranée. Si la mortalité cardiovasculaire peut atteindre 60 % en Finlande ou en Ecosse, elle ne dépasse pas 40 % en France : compte tenu de ce différentiel, quel est le bénéfice réel de la prescription des statines en France ? Peut-on justifier leur poste budgétaire, très important à la Sécurité Sociale, sans avoir étudié leur impact dans les conditions de leur utilisation ? Que penser de la généralisation de leur prescription chez les personnes de plus de 80 ans ayant vécu en bonne harmonie toute leur vie avec un taux de cholestérol à 2,4 g/l, sans autre facteur de risque cardiovasculaire évident ? (l'étude chez les personnes âgées n'a été faite, elle aussi, que dans les pays à forte mortalité cardiovasculaire)[3] .

Un des principes de base de l'épidémiologie clinique est que les résultats d'une étude s'appliquent à la population dont sont issus les cas : ceci est particulièrement vrai en thérapeutique des facteurs de risque, par opposition aux thérapeutiques à visée étiologique (antibiotiques essentiellement), dans la mesure où les facteurs de risque varient beaucoup d'une région du monde à l'autre.

II- Applicabilité en pratique clinique et taille d'échantillon des études :

a- Dans le domaine de la thérapeutique :
On entend souvent dire qu'un essai est d'autant plus valable que sa taille d'échantillon est importante : il s'agit souvent d'un argument des représentants de l'industrie pharmaceutique. Taille d'échantillon et validité n'ont bien entendu que peu de relations, et l'on oublie trop souvent qu'une taille d'échantillon importante est nécessaire pour mettre en évidence *une efficacité marginale :*

Le premier essai randomisé sur acyclovir et encéphalite herpétique a été interrompu après 20 inclusions : 8 patients sur 10 dans le bras traité par acyclovir évoluaient favorablement, alors que 8 sur 10 dans le bras placebo décédaient : l'action était telle qu'*a posteriori*, la randomisation n'aurait pas été nécessaire. La randomisation, *nec plus ultra* des études cliniques car elle égalise toutes les conditions au départ entre les groupes comparés, n'est utile que lorsque l'efficacité n'est pas visible à l'œil nu. Les anti-tuberculeux, les antibiotiques dans la méningite à méningocoque n'ont pas eu besoin de randomisation.

Lorsque l'efficacité n'est pas immédiatement visible, c'est-à-dire, lorsque l'action du médicament testé est marginale dans l'immédiat, ou l'amplitude de l'effet modérée, voire

modeste à moyen ou long terme, une taille d'échantillon importante permet de rendre une différence statistiquement significative entre *deux grands groupes de patients.* Dès lors, le bénéfice individuel devient plus aléatoire (et en règle générale très difficile à définir), et se pose la question de la différence entre le *statistiquement significatif* et le *cliniquement significatif.*

Exemples :
Personne ne songerait à contester l'intérêt de la prise en charge en soins intensifs de l'infarctus du myocarde au stade aigu, avec réalisation d'une coronarographie et thrombolyse ou mise en place d'un stent. La mortalité en est améliorée de 20 %, ce qui est présenté comme spectaculaire… et est retrouvé de façon constante dans toutes les études. *Il s'agit d'une réduction du risque relatif de mortalité de 20%.*
Cependant, la mortalité spontanée de l'infarctus aigu du myocarde tout venant est de 12 %. La prise en charge intensive la ramène (en supprimant les décimales) à 9 %, et 20 % (ou 25 %) représente le différentiel (12 – 9)/12. Autrement dit, la réduction absolue, effective du risque de mortalité est de 3%… [4]
Quelles en sont les conséquences en pratique lors de la prise en charge d'un patient aux urgences ?

- On sauve 23 vies en traitant 1000 patients : il faut donc appliquer le traitement.
- On provoque 9 accidents vasculaires hémorragiques graves, pour le prix des 23 vies sauvées.
- On a sans doute traité 973 patients sans utilité notable pour eux.
- Le problème est que l'on ne sait pas au moment de la prescription, et que l'on ne saura jamais ensuite, quel patient, quel individu, aura bénéficié du traitement. On saura rapidement par contre pour qui le traitement aura été délétère, et il n'est pas certain –il est même peu probable- qu'il s'agisse des patients qui seraient décédés de toute façon.

Un moyen pratique de déterminer l'applicabilité des résultats est de s'interroger sur la fréquence de l'événement, que l'on peut juger par l'intermédiaire de l'incidence observée dans le groupe placebo. Si celle-ci est importante (par exemple 30% de femmes présentant au moins une nouvelle fracture vertébrale en 3 ans de suivi), avec une réduction du risque relatif notable, cela signifie que le produit est intéressant dans une population à haut risque. On identifiera facilement ces individus à haut risque dans la pratique clinique en observant les caractéristiques des individus inclus dans l'essai (par exemple des femmes ménopausées de plus de 70 ans en moyenne ayant déjà au moins une fracture vertébrale.

Point important : l'application généralisée en pratique clinique des résultats des essais randomisés à très grande taille d'échantillon permet sans doute de sauver quelques vies dans le meilleur des cas, mais traduit avant tout notre incapacité à définir les patients bénéficiant réellement du traitement et à affiner les contre-indications : pour être mieux applicable, la recherche clinique de demain devra s'intéresser aux facteurs prédictifs de réponse. C'est une démarche parfois envisagée dans les essais cliniques actuels, avec la publication d'analyses permettant de définir les groupes à meilleure réponse. Trop souvent ces analyses sont néanmoins réalisées a posteriori, pour servir des intérêts commerciaux. Mais de plus en plus, ces analyses sont prévues dans les protocoles a priori, et facilitent l'orientation de la prise en charge. Par exemple, l'analyse d'un certain nombre de catégories de patientes qui bénéficient le plus du traitement de l'ostéoporose par l'acide zolédronique a été prévue dans le protocole, ce qui a permis de mettre en évidence les meilleures cibles.

Par ailleurs, les essais thérapeutiques contrôlés contre placebo sont de moins en moins réalisés car il existe des traitements de référence dans nombre de pathologies. De ce fait, la différence attendue entre un traitement de référence et un nouveau produit est généralement moindre qu'entre un placebo et un nouveau médicament, si bien que les tailles d'échantillon s'accroissent. Pour échapper à ce problème, des critères de jugement intermédiaires sont souvent utilisés, mais leur pertinence clinique est souvent inférieure à celle des critères cliniques dits durs. Certains polymorphismes génétiques pourraient être associés à une meilleure réponse au traitement, et les détecter pourrait permettre dans l'avenir de mieux individualiser les traitements.

b- Dans le domaine des études d'observation:
Il n'y a pas là d'action directe et voulue sur le patient comme dans un essai randomisé. Un facteur d'exposition peut être rare mais très toxique, un événement peut être rare mais grave, et la rareté même de l'exposition ou de l'événement nécessitera des études de population à grande taille d'échantillon pour être mise en évidence.
La plupart des cancers ne sont pas des pathologies fréquentes, mais mettre en évidence leurs facteurs de risques environnementaux de telle sorte à en prévenir l'action peut être utile :
Exemples :

- L'exposition à l'amiante est plutôt rare comparée aux facteurs de risque cardiovasculaire, et le mésothéliome reste à l'heure actuelle une tumeur rare. La reconnaissance de la toxicité de l'amiante a cependant permis de diminuer l'exposition en modifiant un certain nombre de procédés de fabrication.
- L'hépatocarcinome n'est pas le cancer le plus fréquent, mais de larges études de cohorte ont permis de confirmer le rôle étiologique du virus de l'hépatite B : le vaccin devrait permettre d'en diminuer l'incidence.
- L'observation de dizaines de milliers de nouveaux-nés a permis d'objectiver le rôle du décubitus ventral dans la mort subite du nourrisson, événement heureusement 'rare' si on le rapporte à l'ensemble de la cohorte. L'applicabilité en est évidente.
- Les registres de cancer basés sur la population permettent d'objectiver les différences régionales en incidence, et de formuler des hypothèses quant aux facteurs favorisant tel ou tel cancer : les registres nationaux des pays scandinaves, exhaustifs sur l'ensemble de la population, en illustrent tout l'intérêt.

Point important : une taille d'échantillon importante peut être nécessaire dans une étude d'observation pour mettre en évidence l'effet d'un facteur de risque corrigeable ou modifiable, avec conséquences non 'marginales' et des résultats et conséquences (élimination ou modification du facteur de risque) plus souvent justifiées en terme de santé publique et individuelle.

III- Applicabilité, comparabilité des bras d'un essai, et éléments confondants :

De fait de sa fréquence, la pathologie cardiovasculaire est riche de multiples essais randomisés, et bêta-bloqueurs, calci-bloqueurs, anti-hypertenseurs, hypocholestérolémiants, anti-agrégants ont été testés sans relâche depuis plus de 20 ans. Pour certains médicaments, un bénéfice au long terme en survie a pu être montré. Pour beaucoup, l'efficacité a porté sur des données moins 'dures' que la survie. Pour tous, les essais ont toujours été de grande taille, car l'efficacité de ces différentes classes thérapeutiques ne se voit pas à l'œil nu. La randomisation, lorsqu'elle a fonctionné, a permis d'égaliser les conditions pathologiques initiales et les facteurs de risque entre les groupes comparés. *Son but est d'éliminer de la*

comparaison les éléments confondants, ceux-ci pouvant expliquer par eux-mêmes tout ou partie du pronostic indépendamment de la thérapeutique testée.
Cependant, la mortalité ou la morbidité à long terme ne dépend pas uniquement des conditions initiales, mais également de la persistance, ou non, de l'action des facteurs de risque pendant le suivi. Le tabagisme est un des facteurs de risque cardiovasculaire majeurs, qui plus est corrigeable. Il est fort probable que sa poursuite au long cours aggrave le pronostic, et l'on pourrait penser que son arrêt puisse l'améliorer.
Aucun essai randomisé dans le domaine cardiovasculaire ne prend en compte ce facteur confondant majeur *pendant le suivi, et non pas seulement comme condition initiale,* dans l'analyse des résultats. Il n'est pourtant pas impossible de le mesurer et de le prendre en compte en analyse multivariée, *ce qui permettrait de mesurer avec plus d'exactitude l'ampleur de l'action thérapeutique une fois le puissant facteur de risque 'tabac' éliminé.* Il serait intéressant de savoir ce qui resterait alors de l'efficacité des multiples traitements proposés (augmentée ? ou diminuée car marginale par rapport à l'arrêt de l'intoxication ?). Faudrait-il continuer à prescrire des hypocholestérolémiants, inhibiteurs de l'enzyme de conversion ou sartans coûteux, ou se lancer dans des campagnes anti-tabac énergiques et soutenues dans le temps alors que des centaines de millions d'euros ont été dépensés chaque année dans le cadre de la politique agricole commune pour… soutenir la culture du tabac en Europe ?
Point important : de toutes les études cliniques, les essais thérapeutiques sont sans doute celles dont les résultats –ou la conclusion- sont le plus appliqués. Pourtant, les éléments confondants en cours d'essai, contrairement aux éléments confondants initiaux, sont rarement mesurés et analysés.

IV- Applicabilité, force des conclusions, et force de l'évidence. Valeur des critères de jugement:
Il est rare que l'application des résultats d'une étude soit directe… L'information, pourtant accessible directement par la lecture des revues, passe en règle générale par toute une série de relais avant d'être livrée au prescripteur : les laboratoires pharmaceutiques font d'importants efforts de promotion en délivrant un message simple, les conférences de consensus, les sessions de formation médicale continue 'digèrent' l'information et la présentent sous forme de conduite à tenir… Pourtant la réalité n'est pas aussi simple, et l'on a sans doute tort de considérer qu'un médecin 'Bac + 10' ne devrait pas comme tout scientifique aller vers la source des données.

Les anti-aromatases ont obtenu très rapidement l'autorisation de prescription en première ligne dans le cancer du sein, au détriment du plus ancien tamoxifène (Nolvadex®). Cette autorisation est survenue après l'essai randomisé publié dans le New England Journal of Medicine, concluant à l'amélioration notable du pronostic des patientes traitées par anti-aromatase par rapport au bras 'tamoxifène' (4500 patientes incluses), amélioration saluée par un éditorial dithyrambique titrant sur 'les nouvelles étoiles dans le ciel du cancer du sein' [5]
Les publicités en dernière page dans les revues de médecine générale françaises suivaient de peu l'article original (15 jours…).
Pourtant, les courbes de survie publiées dans cet article ne montrent aucune différence de mortalité. Il existe simplement une différence statistiquement significative en terme de survie sans rechute, retardée de quelques semaines dans le bras 'anti-aromatase'. En parallèle, les patientes sous anti-aromatase présentent fréquemment des arthralgies, et leur risque de fracture ostéoporotique est significativement augmenté par rapport à celles sous tamoxifène, du fait du blocage de la sécrétion résiduelle d'oestrogènes. Le bénéfice global, quantifié sous forme d'amélioration de la qualité de vie, est ainsi difficile à transposer dans la pratique

clinique. Quel en est l'impact vrai dans la vie des patientes ? Cela justifie-t-il un coût 10 fois plus élevé que le tamoxifène, en sachant que les sommes investies dans une direction ne le seront pas dans une autre, peut-être plus efficace ?

La même question peut se poser face à la généralisation des traitements anti-Alzheimer : ils sont coûteux, ralentissent (peut-être ?) la dégradation du Mini Mental Score (MMS), mais changent-ils vraiment la vie des patients et de leur famille ? Faut-il dépenser des millions d'euros dans une thérapeutique pour l'instant très fragile, ou vaudrait-il mieux les investir dans la recherche fondamentale susceptible un jour d'apporter une amélioration réelle ?

Points importants :

- *La justice a compris depuis des siècles l'importance de la séparation des pouvoirs : le juge d'instruction, l'avocat, le procureur, le juge, sont des personnes différentes et supposées être indépendantes. Toutes ces fonctions sont à l'heure actuelle réunies en une seule dans l'évaluation des thérapeutiques : les laboratoires pharmaceutiques produisent la molécule, construisent l'essai randomisé, l'organisent, l'analysent, le publient -assez souvent à l'heure actuelle- et en présentent les résultats, notamment dans les congrès qu'ils financent très largement. Ils possèdent la puissance financière nécessaire à l'ensemble, mais les payeurs in fine sont les systèmes d'assurance maladie. Ne vaudrait-il pas mieux appliquer le vieux principe de la justice, avec des laboratoires –indispensables- effectuant la recherche pharmacologique, défendant leur molécule, les sociétés savantes -après tout, des docteurs en médecine et en sciences…- évaluant la molécule de façon indépendante sur le plan clinique, des instances nationales telles l'AFSSAPS ou la FDA instruisant le dossier, avant de faire passer une information plus objective ?*
- *Les résultats sur critères de jugement 'légers' (MMS pour Alzheimer, chiffres tensionnels pour hypertension, densité osseuse pour ostéoporose…) doivent être considérés avec plus de recul que les critères cliniques lourds (survie, incidence d'AVC ou d'infarctus du myocarde, incidence de fracture…) avant de généraliser une thérapeutique.*

V- Applicabilité des études médico-économiques:

Les médecins ont tendance à ne pas s'intéresser aux études médico-économiques… et de fait, laisser les décisions aux économistes non médecins. Il est important de connaître les types d'études économiques, leur philosophie de base (coût-utilité, coût-efficacité, coût-bénéfice, etc), qui elle-même peut déjà faire l'objet de discussions. Il est important ensuite, d'en connaître les modes de calcul et de se rappeler que comme toute étude, leurs résultats ne s'appliquent qu'aux populations sur lesquelles elles ont été faites : les coûts diffèrent considérablement d'un pays à l'autre au sein même des pays d'Europe de l'Ouest, le montant de la consultation d'un médecin généraliste est 5 fois plus élevé aux Etats-Unis ou au Royaume-Uni qu'en France, le coût total du travail varie en fonction des charges, très différentes d'un pays à l'autre, et le coût des médicaments lui aussi est décidé au niveau national et non pas international. Enfin, les 'ingrédients' (les coûts) inclus dans les calculs varient tout aussi considérablement en fonction de ce que l'on veut montrer : faut-il n'inclure que les coûts directs (coût total d'une hospitalisation par exemple), ou une partie de coûts indirects (allocations, estimation du coût du handicap, suite des soins à domicile, manque à

gagner….) ? Jusqu'où faut-il inclure les coûts indirects (eux-mêmes très variables en fonction du statut social ou familial, de la couverture sociale, et bien sûr des pays) ?
Dans un même contexte national et économique, les coûts peuvent varier considérablement en fonction même des caractéristiques cliniques du patient : le traitement de l'ostéoporose de l'homme de plus de 60 ans peut ainsi varier du simple au décuple si l'on prend en compte l'âge et les antécédents [6] Par exemple, il est coût-efficace, dans la plupart des systèmes de santé, de traiter les individus à fort risque de fracture à court terme, plutôt que de prescrire des traitements préventifs de la perte osseuse à une grande partie de la population préalablement sélectionnée sur la base d'une mesure de densité minérale osseuse. Ainsi ce sont surtout les sujets âgés, dont le risque de fracture est le plus grand, qu'il est le plus coût-efficace de prendre en charge. Certains pays font des choix radicaux, en ne remboursant qu'une seule ligne de traitement, sous forme de générique. Par exemple, au Royaume-Uni, seul l'alendronate générique est remboursé en première ligne du traitement de l'ostéoporose post-ménopausique. De même, seule la première ligne de chimiothérapie est remboursée dans le traitement du cancer du sein métastatique.

Points importants :

- *Les sciences économiques, tout en utilisant les outils mathématiques et le langage scientifique, sont plus proches des sciences sociales dans les bases de leur raisonnement : elles traduisent donc plus un mode de fonctionnement social qu'une 'réalité' au sens scientifique du terme, et les résultats de leurs études doivent être interprétés en gardant en mémoire cette relativité.*
- *On ne fait jamais la preuve que telle ou telle mesure est 'économiquement' efficace ou rentable : on apporte juste quelques indices pouvant aider à la décision, valables uniquement dans un contexte géographique et temporel étroit.*
- *Les médecins doivent comprendre la méthodologie de ces études pour pouvoir discuter, arguments solides à l'appui, de leur application face aux personnes qui prendront les décisions, non-médecins le plus souvent dans le contexte actuel.*
- *Il n'y a, dans la presse médicale nationale ou internationale, presque aucune étude médico-économique française. Quelle est la validité de décisions prises sur des études non publiées, dont la méthodologie n'est pas connue, et qui peut-être n'ont tout simplement pas été faites ?*
- *Il reste là un très large domaine d'étude pour les médecins cliniciens.*

VI- Application des résultats… et traditions ancrées :

Il arrive également que des résultats d'étude s'opposent à certaines règles de prescription solidement ancrées dans l'enseignement et les habitudes : la 'double antibiothérapie adaptée au germe', répétée par des générations d'étudiants en est certainement une illustration exemplaire. La 'double antibiothérapie' sous-entend habituellement l'adjonction d'un aminoside.
De multiples essais randomisés comparant une mono-antibiothérapie au même antibiotique avec adjonction d'un aminoside ont été effectués, dans des situations cliniques très différentes. Ils n'ont que très rarement soutenu l'emploi de l'aminoside, et la constance de leurs résultats est en soi remarquable. Deux méta-analyses de ces essais, l'une chez les patients immuno-compétents, l'autre chez les patients immuno-déprimés, confirment les résultats individuels de la grande majorité de ces essais : les aminosides n'améliorent en rien le pronostic *tant clinique que bactériologique* (guérison/mortalité, éradication du germe). Par contre, leur seul effet statistiquement et cliniquement significatif est représenté par… leur

néphrotoxicité [7] [8]. Les résultats de ces multiples études et de leur synthèse seraient facilement applicables. Quelles résistances les empêchent d'être au moins cités dans les conférences de consensus ?

Points importants :

- *Les études médicales apportent un lot important de connaissances… mais doivent être considérées comme le socle sur lequel bâtir ensuite le front mouvant des connaissances applicables.*
- *Les conclusions d'une conférence de consensus ne devraient jamais être enseignées sans qu'en soient décortiquées les fondations et la façon d'arriver au dénominateur minimal commun (entre qui ?) qui les établissent.*

B- Valorisation de la recherche clinique :

L'application aux patients des résultats lorsqu'ils sont applicables constitue sans doute l'élément primordial de la valorisation de la recherche clinique : le but unique d'une recherche clinique est l'amélioration, d'une façon ou d'une autre, de la condition des patients. D'autres éléments cependant ne sont pas à négliger:

I- Valorisation et enseignement du deuxième cycle :
L'enseignement des sciences fondamentales en premier cycle de faculté de médecine est souvent directement en prise avec la recherche correspondante, et il ne s'est pas écoulé des années entre la découverte des introns et leur introduction dans les cours de génétique ou de biologie moléculaire. La communication des résultats des études cliniques en cours passe un peu plus mal, car elle est trop souvent considérée comme étant réservée aux spécialistes. Les résistances à l'enseignement de la lecture critique, et ses reports successifs dans les programmes officiels en témoignent. Les résistances à l'enseignement des biostatistiques, outil indispensable à une lecture critique documentée, sont encore plus importantes tant de la part des étudiants que des enseignants. Tout le monde cependant s'accorde à dire, voire à penser, que les sciences cliniques sont profondément évolutives…. Alors ? Comment concilier le caractère évolutif de la connaissance et un enseignement d'allure fixée ?
Certaines facultés ont introduit depuis plus d'une décennie des enseignements de raisonnement médical : d'autres ont encore à le faire, et la lecture critique ne peut se concevoir sans enseignement sérieux des outils dont elle se nourrit. On ne peut pas interpréter un article sans en comprendre les méthodes tant épidémiologiques que statistiques, et l'on ne peut appliquer ses conclusions sans en comprendre les fondations. La recherche clinique doit donc être introduite dans l'enseignement à deux niveaux :

a- L'indispensable enseignement des méthodes et techniques, de leurs applications, mais aussi de leurs forces et faiblesses et de la variabilité des interprétations en découlant. Ceci doit faire l'objet d'un enseignement spécifique avant les modules de lecture critique ;
b- L'intégration de lecture d'articles spécifiques dans les modules de spécialités, vise à faire démonter par les étudiants, la naissance, l'évolution, et dans certains cas, l'aboutissement d'un concept médical : pourquoi favorise-t-on, ou dénigre-t-on, telle technique diagnostique ou telle thérapeutique à tel moment ?

Exemples :

- *Il s'est écoulé 10 années entre l'abandon de l'artériographie pulmonaire, examen 'gold standard' pour l'embolie pulmonaire, et la démonstration que l'angio-scanner qui l'a remplacée était faussement négatif dans… 50 % des cas. Pendant ces 10 années, la nature a remplacé l'action thérapeutique*

médicale pour des dizaines de milliers de patients… à l'époque des soins intensifs performants. La comparaison des deux techniques n'avait tout simplement pas été faite auparavant. Combien d'étudiants, et a fortiori combien de praticiens, le savaient ? La valeur diagnostique du scanner, y compris par technique hélicoïdale, reste débattue et mal évaluée cinq années plus tard [9]

- *Il s'est écoulé 30 ans, de multiples débats et congrès et des dizaines de milliers d'articles, entre l'application généralisée de l'hormonothérapie post-ménopausique dans les années 70 et la mesure de ses effets positifs et négatifs au début des années 2000 : par chance, les effets négatifs (augmentation de l'incidence des cancers du sein et des accidents cardio-vasculaires) ont été plus ou moins compensés par les effets positifs (diminution de l'incidence des cancers du colon et des fractures ostéoporotiques). Mais fallait-il laisser un tel rôle à la chance, et n'aurait-on pu prévenir les effets négatifs en les connaissant, c'est-à-dire en randomisant le traitement avant sa généralisation à des centaines de milliers de femmes NON malades ? Là encore, combien d'étudiants et de praticiens connaissaient les bases scientifiques sur lesquelles la prescription s'était généralisée ?*

II- Valorisation et enseignement du troisième cycle :
Le troisième cycle des études médicales comprend bien sûr la thèse et le mémoire de spécialité, mais aussi de multiples diplômes universitaires dont l'obtention, bien souvent, se fait sur la base d'un examen et d'un mémoire. Thèses et mémoires peuvent être bibliographiques, et reposer ensuite non lus dans les bibliothèques des services, car d'un intérêt modéré, … ils peuvent également -pour un investissement en temps et travail presque similaire (à condition d'être bien encadrés)-, être le sujet d'une recherche clinique originale conduite par le candidat, validant son diplôme et faisant l'objet d'une publication pour le bénéfice des patients, du candidat, du service et du directeur du projet…

III- Valorisation de la recherche et communication des résultats :
Tout résultat doit être accessible, et par conséquent publié. Un résultat négatif est parfois aussi intéressant qu'un résultat positif, même s'il est plus difficile à publier en pratique.

a- Les congrès :
Les résultats peuvent être communiqués dans les congrès nationaux et internationaux, et permettre ainsi l'établissement de liens et de coopérations avec d'autres équipes travaillant sur des thèmes voisins. Les communications et posters permettent aussi de motiver et promouvoir les jeunes collègues dans la discipline, et d'accroître l'attractivité d'un service. En pratique, il n'y a pas de recherche efficace sans l'apport des plus jeunes, et cette recherche en retour est utile pour leur carrière… et donc pour le maintien à un bon niveau de la qualité de soins offerte par un service.

b- Les publications écrites :
Il existe une stratégie de publication, et il ne faut pas craindre –ou se laisser décourager- par les refus des revues : une publication peut être refusée plusieurs fois pour des motifs plus ou moins justifiés. Les motifs justifiés servent à améliorer le papier pour une prochaine soumission. Les motifs moins justifiés ne font que traduire la subjectivité inhérente à toute activité humaine… et ne doivent pas empêcher la poursuite du travail.
Les revues médicales existant depuis quelques années sont pourvues d'un facteur d'impact (impact factor), calculé sur la base du nombre de lecteurs, du nombre de citations de leurs

articles dans des travaux ultérieurs et d'autres indices d'impact dans la communication des connaissances. Ce facteur d'impact n'est pas obligatoirement le reflet de la qualité d'un travail (une revue généraliste avec de multiples lecteurs, abordant des pathologies fréquentes comme les maladies cardiovasculaires aura un facteur d'impact supérieur à une revue très spécialisée de neurochirurgie ou de biostatistiques médicales), mais est souvent pris en compte dans l'évaluation du curriculum vitae d'un candidat. Il est pris en compte dans l'évaluation de l'activité de recherche d'un service ou d'une unité. On peut donc, pour un travail donné, tenter une revue susceptible d'accepter l'article dans la spécialité correspondante avec le facteur d'impact le plus élevé, quitte à essuyer un refus et descendre ensuite l'échelle des facteurs d'impact.
Les publications sur Internet se multiplient à l'heure actuelle : elles ne sont pas toutes pourvues d'un facteur d'impact pour l'instant, mais la donne sera sans doute considérablement modifiée dans la prochaine décennie et les stratégies de publication devront s'adapter.

IV- Valorisation de la recherche et amélioration de la formation des professionnels de santé :
Il n'y a pas de soins sans acteurs de santé (médicaux ou para-médicaux), et il n'y a pas de soins de haut niveau sans acteurs de haut niveau participant à leur amélioration de façon continue… une carrière bien construite doit donc l'être -entre autres- sur une recherche bien conduite, et cet aspect de la carrière médicale doit être abordé très tôt avec les internes en début de cursus. La plupart des masters de recherche sont validés après *soumission* d'un article dans une revue internationale. Même s'il n'existe pas de règle écrite, la plupart des universités demandent, pour qu'une thèse de science puisse être soutenue, qu'elle ait fait l'objet de deux articles originaux en premier auteur *acceptés* dans des revues internationales avec comité de lecture, d'impact facteur supérieur à 2 le plus souvent (à l'exception des spécialités très pointues à lectorat étroit). Le nombre de publications originales *acceptées en premier auteur* sur le même thème monte souvent à 6 pour la soutenance d'une habilitation à diriger les recherches. Ceci ne peut pas s'improviser et il vaut mieux prévoir le la mise en œuvre des travaux, et donc des publications, plusieurs années à l'avance en fonction de l'âge du candidat et du calendrier des opportunités : un peu de réalisme ne nuit pas à la conduite d'une recherche de bonne qualité et à l'essor d'un service.
La médecine générale ne fait plus exception à cette règle : elle est devenue depuis la transformation du concours d'internat ancien en examen classant national, une spécialité comme les autres. La création de départements de médecine générale dans les universités, et avec eux, de postes de professeurs, maîtres de conférence, et plus récemment, de chefs de clinique en médecine générale officialisent *de facto* des carrières en médecine générale similaires aux carrières effectuées dans les autres spécialités. La médecine générale reste un immense champ d'investigation clinique pouvant s'organiser au sein de réseaux de médecins libéraux, et les méthodes de l'épidémiologie clinique sont tout particulièrement adaptées à ce type de recherche appliquée.

V- Valorisation des découvertes :

Il faut envisager le dépôt de brevet ou la licence d'une technique le plus tôt possible dans le processus de recherche. Y penser lorsque l'on publie la découverte est déjà trop tard car tout le monde pourra s'approprier gratuitement le travail effectué. Le dépôt de brevet est très important dans une carrière de recherche, souvent plus que les publications elles-mêmes. Ils sont sources de revenus pour l'institution, pour un groupe de recherche, et contribuent au rayonnement d'une école.

Il s'agit d'un domaine d'abord très difficile, car maîtrisé seulement par des spécialistes du droit des brevets. Il convient donc de prendre conseil auprès des cellules de valorisation des universités ou des instituts de recherche (INSERM, CNRS).
Les retombées sont aussi importantes pour les patients, car le dépôt d'un brevet, puis son transfert industriel, sont la clé du développement rapide d'une nouvelle technique.

Chapitre XX

La synthèse de l'information scientifique

Yves Matillon

La finalité de la recherche clinique est d'améliorer la qualité des soins donnés aux malades… y compris pour les stratégies de prévention. Certains domaines (la thérapeutique et notamment les médicaments) sont privilégiés pour développer des études de recherche clinique, si on les compare à l'étude de la prévention, des soins de premier recours, sans parler des médecines dites « alternatives ».

Dans ces domaines privilégiés, les études sont parfois très nombreuses. Les données sont alors difficilement accessibles pour le praticien et le soignant.

En effet, le nombre des publications médicales est croissant mais leur qualité est inégale. Il est difficile pour un praticien de connaître, d'évaluer et d'assimiler toutes ces données nouvelles, et *a fortiori* de les intégrer dans sa pratique quotidienne. D'autant que les journaux électroniques se multiplient et que la qualité de l'information fournie est inégale !

D'où l'idée de favoriser les travaux permettant la synthèse de l'information en médecine. Des méthodes quantitatives existent -notamment la méta-analyse qui fait l'objet d'un chapitre de ce livre. Nous évoquerons plutôt ici les méthodes dites « qualitatives », en sachant que celles-ci, chaque fois que possible, peuvent avoir recours et/ou bénéficier de méthodes statistiques de type « méta-analyse ».

Par ailleurs les contraintes financières dans les pays développés sont telles que les démarches appelées « évaluatives » ont justifié dans les années quatre-vingt, notamment en France, le développement de ces synthèses d'information, avec un objectif soit individuel (décision malade-médecin), soit collectif (décision politique/de santé publique).

Enfin l'utilisation de l'EBM (*evidence based medicine)*, pour améliorer la pratique médicale, pour diffuser et rendre accessible une information scientifique est toujours une source de questions.

Est- elle bénéfique ?...ou s'agit-t-il d'une illusion perdue ?

L'évolution de l'EBM vers l'«evidence based management », voire l' «evidence based decision »…voire le « managing evidence based knowledge » (1) sont des questions souvent débattues dans les journaux anglo-saxons (2). Concrètement, les données scientifiques publiées de bonne qualité, de plus en plus nombreuses, doivent être synthétisées. Tout cela pour mieux former/informer les médecins et mieux informer patients et professionnels de santé pour une décision basée sur des données objectives.

La définition des recommandations pour la pratique clinique utilisée en France dérive de celle proposée par l'Institute of Medicine en 1990 aux USA.

Les recommandations sont " des propositions développées méthodiquement pour aider le praticien et le patient à rechercher les soins les plus appropriés dans des circonstances cliniques données ". Un soin est approprié lorsque " le bénéfice clinique qu'il procure est supérieur aux risques et aux coûts qui en découlent ". Cette définition sous-entend la capacité de quantifier préalablement les rapports bénéfice/risques et coût/efficacité d'une intervention diagnostique et/ou thérapeutique. Les recommandations visent à mettre à la disposition des professionnels de santé et des patients une synthèse objective des données disponibles afin de les aider dans leurs choix de soins. Elles constituent aussi l'étape préalable au développement de standards ou de référentiels de pratique destinés à l'évaluation, voire au contrôle de la pratique professionnelle.

Des recommandations ont été développées dans le domaine de la santé dans les années 1970, en Amérique du Nord, puis dans les pays européens. Un réseau international, le *Guidelines International Network*[1] (GIN) a été créé en 2002 afin de coordonner les efforts dans ce domaine.

Le développement des recommandations médicales et professionnelles cherche à répondre à l'amélioration de l'information des professionnels de santé et de celle des usagers du système de santé.

[1]Site internet : http://www.g-i-n.net

L'objectif principal des recommandations est d'exposer le plus clairement possible les interventions et les stratégies appropriées, celles qui ne le sont pas ou ne le sont plus, et celles pour lesquelles les connaissances sont insuffisantes. Les recommandations peuvent s'appliquer à la prévention, au diagnostic, au traitement ou au suivi d'une maladie. Le champ d'application des recommandations ne se limite pas à l'aide à la décision médicale.

Il peut aussi concerner la formation des professionnels de santé. L'information pour les patients et les familles doit être accessible y compris dans ses composantes économiques, organisationnelles, juridiques, sociales ou éthiques de la pratique médicale. Les recommandations peuvent acquérir une portée juridique lorsqu'elles sont intégrées dans des textes officiels, circulaires, décrets ou arrêtés. Les " données acquises de la science " sont intégrées aux objectifs réglementaires. Elles peuvent aider à construire des référentiels pour évaluer les pratiques professionnelles. Elles peuvent guider la recherche clinique en mettant en évidence les domaines de soins encore inexplorés ou controversés. C'est notamment le cas de thérapeutiques dites « alternatives ».

Les recommandations ont été voulues initialement comme une aide à la prise de décision par le patient.

Ainsi, beaucoup de recommandations publiées par des organisations nationales comprennent aujourd'hui une version destinée aux patients (version « papier » mais aussi parfois audio ou vidéo). C'est ce principe qui est à l'origine du programme MedlinePlus, mis en place par la National Library of Medicine aux USA.

Les patients pourraient aussi intervenir dans le processus d'élaboration d'une recommandation mais leur place reste à préciser : qui doit participer ? des représentants d'organisations de patients ? de consommateurs ? Comment apprécier leur représentativité ? A quel moment doit se faire cette participation ? Lors de leur élaboration ? de leur diffusion ?

1. Méthodes d'élaboration des recommandations

1.1 Méthodes standardisées

Les méthodes standardisées d'élaboration de recommandations diffèrent selon la place donnée aux trois sources possibles d'information : littérature scientifique médicale, avis d'experts et investigation (réalisation d'enquêtes, en particulier concernant la pratique) et sur les moyens utilisés pour collecter l'information et en faire la synthèse. Les recommandations doivent répondre à des critères de qualité**.** Une grille de qualité des recommandations a été validée au niveau international.
En France, deux méthodes ont été proposées aux promoteurs de recommandations : **la conférence de consensus et la méthode dite RPC (recommandations pour la pratique clinique)**. Leurs méthodologies détaillées sont disponibles sur le site de la Haute Autorité de Santé. Ces deux méthodes ont été initialement codifiées par deux organismes fédéraux américains, le NIH (*National Institutes of Health*) pour la conférence de consensus, l'AHCPR (*Agency for Health Care Policy and Research*, devenue AHRQ, *Agency for Health Research and Quality*) pour la RPC.

Dans la conférence de consensus, les recommandations sont élaborées par un groupe de professionnels de santé, en principe non experts du sujet, au décours d'une séance publique. Durant deux jours ils ont entendu des experts chargés de répondre à des questions précises prédéfinies. Cette méthode utilise trois modèles différents : le modèle judiciaire où des témoins (experts) sont écoutés par un jury « impartial » ; la réunion scientifique au cours de laquelle des experts exposent et discutent leurs travaux ; le débat démocratique où chaque personne peut exprimer son point de vue. Les recommandations sont écrites par le jury « à huis clos » dans les vingt-quatre ou quarante-huit heures suivant la séance publique puis sont présentées au public. Cette rédaction souvent nocturne et dans l'urgence (qui limite par ailleurs l'influence des lobbies) est un point négatif de cette méthode. La deuxième critique majeure est la façon dont est prise en compte la littérature scientifique. Quels moyens a-t-on de contrôler les dires des experts ? C'est pour cela que, dès 1990, l'ANDEM (Agence Nationale pour le Développement de l'Evaluation Médicale) avait demandé que, pour qu'une conférence de consensus soit valide, une revue systématique de la littérature soit effectuée par des personnes indépendantes des experts, afin que le jury soit mieux informé et

préparé. Dans ces conditions, la méthodologie de la conférence de consensus se rapproche de la RPC.

Dans la RPC, l'étape clé est la revue de la littérature effectuée selon des modalités prédéfinies et standardisées (définition précise du sujet, critères de recherche de l'information scientifique, de sélection et d'analyse des articles, enfin critères de synthèse d'information), indépendamment des experts afin d'en garantir la qualité. La méthode fait intervenir : un **promoteur** (qui prend l'initiative de l'élaboration des recommandations et en assure le financement), un **comité d'organisation** (qui précise le thème et les questions à résoudre, décide de l'organisation générale du travail, choisit les participants au groupe de travail et assure la logistique de l'ensemble du processus), un **groupe de travail** (qui réalise l'analyse et la synthèse des données disponibles, fait la synthèse des avis d'experts et rédige les recommandations), et enfin un **groupe de lecture** qui donne son avis sur le fond, la forme et l'applicabilité des recommandations et apporte des informations et des avis d'experts complémentaires au groupe de travail. Cette méthode est longue (au moins un an) ce qui en fait le principal inconvénient. Elle aboutit à des documents en général de grande qualité mais longs. Plus que des documents à large diffusion, le texte de ces recommandations doit être pris comme document de travail destiné à rédiger des référentiels de pratiques simples pouvant constituer des outils d'amélioration de la qualité directement applicables.

1.2 Aspects méthodologiques spécifiques

Deux aspects clés dans l'élaboration de recommandations, la prise en compte de la preuve scientifique et celle de l'avis d'experts, font l'objet de travaux de recherche.

- **La prise en compte de la preuve scientifique**

Il est important de connaître la force d'une recommandation et la qualité des éléments de preuve sur laquelle elle est fondée. Le concept de niveau de preuve a été proposé à la fin des années 1970

par la *Canadian Task Force on Periodic Health Examination* [2] *puis* par l'*US Preventive Task Force* pour élaborer des recommandations concernant les examens médicaux à réaliser régulièrement dans le domaine de la médecine préventive [3- 4].

Le niveau de preuve d'une étude peut être défini comme une gradation standardisée de la validité scientifique de l'étude, en fonction de la qualité de sa méthodologie et de sa réalisation, de l'analyse de ses résultats et de la pertinence de ses conclusions. En utilisant une échelle préétablie de niveau de preuve, il est possible de classer systématiquement la littérature médicale en fonction de la qualité de chaque étude. Dans ce contexte, les essais contrôlés randomisés sont supérieurs aux études de cohorte. Ils sont même supérieurs aux études cas-témoins et à tout autre type d'étude.

Malgré un gain d'objectivité et de reproductibilité apporté par la quantification standardisée de la preuve scientifique, il persiste une part subjective liée au jugement des experts. Par exemple, quelle est la valeur de plusieurs études de cohorte aux résultats convergents comparées aux résultats d'un essai contrôlé randomisé ? Quelle valeur accorder aux résultats d'une méta-analyse de petits essais comparés à ceux d'un grand essai multicentrique ? Comment prendre en compte l'avis d'experts lorsqu'il n'y a pas d'étude de niveau de preuve scientifique élevé disponible ?

La quantification de l'avis d'experts

La *RAND Corporation* a proposé une méthode du type " groupe nominal " dérivant de la méthode Delphi, qui vise à définir toutes les indications possibles d'une intervention dans une pathologie donnée : une revue de la littérature scientifique est d'abord réalisée sur le thème traité, à partir de laquelle une liste exhaustive des indications médicales possibles de l'intervention étudiée est établie. Cette revue et cette liste sont alors envoyées à neuf experts, généralistes et spécialistes hospitalo-universitaires et libéraux de la région géographique à laquelle les recommandations sont destinées. Chaque expert attribue un score de 1 à 9 à chaque indication de la liste selon qu'il considère l'intervention comme " appropriée " (score = 9) ou " non appropriée " (score = 1) dans

[2] Site internet : http://www.ctfphc.org

cette indication. Les scores attribués sont ensuite compilés, et le résultat de cette première cotation de la liste est présenté et discuté lors d'une réunion plénière des neuf experts. Au terme de cette discussion, chaque expert effectue une seconde cotation de la liste. La compilation de l'ensemble des scores attribués lors de cette seconde cotation permet d'établir la liste définitive des indications dans lesquelles l'intervention est considérée comme " appropriée ", " non appropriée " ou " douteuse ". Cette méthode aboutit à un répertoire d'indications, que l'on peut utiliser comme autant de recommandations, ou qui peut permettre de construire un arbre de décision [5].

Cette méthode, utilisée en Europe [6], a l'inconvénient d'aboutir à des conclusions qui risquent d'être influencées par le choix des experts [7], et dans lesquelles la place respective de la preuve scientifique et de l'avis d'experts est difficile à situer. Elle peut cependant être utilisée au cours de l'élaboration d'une RPC afin de formaliser l'avis d'experts.

2. Utilisation des recommandations

Les RPC largement diffusées en France sont acceptées par les professionnels de santé. Une base de données de recommandations de langue française est en développement depuis 2005[3], sur le modèle de ce qui existe aux USA[4]. En dehors de la HAS, de nombreuses sociétés savantes de spécialités médicales ainsi que la Fédération Nationale des Centres de Lutte Contre le Cancer ont établi leurs propres programmes de recommandations (programme SOR)[5]. L'élaboration de recommandations par de telles structures thématiques soulève la question de la difficulté à maîtriser des intérêts divergents, parfois, institutionnels et scientifiques…

Les recommandations ont par ailleurs jusqu'à présent concerné essentiellement la pratique des médecins, moins souvent celle des autres professionnels de santé, souvent faute d'information scientifique valide. Sans recherche clinique crédible et valide, comment peut-on disposer

[3] Site internet : http://bfes.anaes.fr

[4] Site internet : http://www.guideline.gov

[5] Programme des Standards Options et Recommandations

d'informations objectives pour décrire les processus de soins et leurs impacts économiques ou sociologiques ?

Les RPC ont été vues en France comme un moyen de régulation des pratiques médicales, comme cela avait été le cas aux Etats-Unis dans les années soixante-dix. Ceci a induit parfois une confusion entre le concept d'aide à la prescription et celui de contrôle des pratiques. Deux logiques différentes ont été prises en compte, l'une pour l'exercice hospitalier, l'autre pour la médecine libérale.

A l'hôpital, les ordonnances de 1996 ont instauré l'accréditation des établissements hospitaliers publics et privés, selon le modèle anglo-saxon, dans lequel l'évaluation des pratiques médicales avait, dans sa version initiale, une place limitée. Cette place doit se renforcer dans l'avenir.

En médecine libérale, l'orientation vers la maîtrise médicalisée des dépenses de santé a clairement été affirmée avec l'instauration en 1993 des références médicales opposables (RMO).

Ensuite, le concept d'évaluation des pratiques professionnelles (associant revue par les pairs, le *peer review* des anglo-saxons, et audit des pratiques) a été proposé. Il s'agissait d'un dispositif professionnel non sanctionnant d'amélioration continue de la qualité.

La loi « Hôpital, patients, santé et territoires » adoptée en 2009, consacre le concept de fusion entre la formation médicale continue et l'évolution des pratiques personnelles pour conduire au développement professionnel continu.

Globalement la production de RPC est assez considérable à ce jour, en France mais aussi dans tous les pays européens.

L'étude de sites internet (www.g-i-n.net et www.guidelines.gov) confirme ce mouvement de la production internationale : sur ces sites, il apparaît que certains sujets sont traités de nombreuses fois dans plusieurs pays (par exemple, l'asthme chez l'enfant, le traitement de la lombalgie ou de l'insuffisance cardiaque) et d'autres ne sont pratiquement pas traités (par exemple, les soins de premiers secours, tous les traitements mis en œuvre par les personnels paramédicaux et les

stratégies diagnostiques). Enfin sur les sites anglo-saxons, les recommandations en français ne sont qu'exception (2 sur 212 recommandations sur le site « guideline.gov » en 2007.

Conclusion

Les recommandations médicales et professionnelles ont pour objectif de répondre à une demande qui provient des médecins, des malades et des organismes de soins et de financement. Cette demande croît au fur et à mesure qu'augmentent la masse des publications scientifiques, les dépenses de santé, mais aussi le désir légitime de la collectivité d'obtenir une qualité optimale des soins au meilleur coût. Afin que les réponses apportées par l'élaboration de recommandations soient suivies d'un effet positif, alors que celui-ci est toujours limité aujourd'hui [8] [9], il est nécessaire que celles-ci soient développées selon une méthodologie rigoureuse même si elle doit être coûteuse, lourde et contraignante. Une des premières études publiées il y a plus de vingt ans aux USA, étudiant l'impact des conférences de consensus sur la connaissance des médecins américains de leurs résultats [8], montre un impact similaire à celui d'une étude européenne récente relative au traitement de la maladie coronarienne [9] : un médecin sur cinq en connait les résultats.

C'est pourquoi, produire des recommandations de bonne qualité, doit se faire en sélectionnant attentivement le thème clinique.

Des efforts importants doivent être consacrés à la diffusion des recommandations, à leur transformation en référentiels de pratique, enfin à leur mise en œuvre. L'élaboration de recommandations est un processus essentiellement national, tandis que leur mise en œuvre est un processus local. Il n'existe pas de solution unique pour améliorer les pratiques médicales. L'impact sera plus important en associant différentes méthodes d'intervention et en tenant compte du contexte de pratique (par exemple, la visite à domicile lorsque l'on s'intéresse à la prescription en ambulatoire, les leaders d'opinion en milieu hospitalier).

Il faut souligner que les modalités d'interventions décrites dans la littérature sont très variables et il est donc difficile d'en tirer des conclusions " universelles ". Les travaux publiés par des équipes françaises avec des méthodologies jugées valides ont eu des résultats identiques à ceux des travaux

anglo-saxons. Ceci montre bien que les recommandations ont un impact sur les pratiques quel que soit le système de santé au sein duquel elles sont élaborées. Les recommandations ont également une valeur pédagogique par le travail collectif de professionnels de santé, de « décideurs », de représentants du public. Leur efficacité dépend en grande partie des moyens que l'on se donne pour les élaborer, les actualiser [10] et les mettre en œuvre [11].

REFERENCES

1. Strauss S, Haynes RB. Managing evidence-based knowledge: the need for reliable, relevant and readable resources. CMAJ, 2009, 180(9)
2. Young C, Godlee F. The BMJ Evidence Centre: a novative approach to healthcare information. BMJ, 2008, 337:a2438 doi:10.1136/bmj.a2438
3. Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J, 1979, *121*: 1193-233.
4. Battista RN, Fletcher SW. Making recommendations on preventive practices: methodological issues. Am J Prev Med, 1988, *4*:53-67.
5. Brook RH, Chassin MR, Fink A, et al,. A method for the detailed assessment of the appropriateness of medical technologies. Int J Technol Assess Health Care, 1986, 2:53-63.
6. Nicollier-Fahrni A, Vader JP, Froehlich F, et al. Development of appropriateness criteria for colonoscopy: comparison between a standardized expert panel and an evidence-based medicine approach. Int J Qual Health Care, 2003,15:15-22.
7. Fraser GM, Pilpel D, Kosecoff J, et al. Effect of panel composition on appropriateness ratings. Int J Qual Health Care, 1994, 6 : 251-5.
8. Jacoby I, Rose M. Transfer of information and its impact on medical practice: the US experience. Int. J. Technol Assess Health Care, 1986, 2 : 107-15
9. Hobbs R. Erhardt L.
   - Acceptance of guideline recommendations and perceived implementation of coronary heart disease/ disease prevention among primary care physicians in five European countries : the Reassessing European Attitudes about Cardiovascular Treatment (REACT)
   - Family practice. 2002; 19 : 596-604
10. Shekelle P, Eccles M, Grimshaw JM, et al. When should clinical guidelines be updated? BMJ, 2001,323:155-7
11. Durieux P. Comment élaborer l'état des connaissances pour contribuer à améliorer les pratiques cliniques ? In Matillon Y, Maisonneuve H. L'évaluation en santé : de la pratique aux résultats. 3ème édition. Paris : Médecine Sciences Flammarion ; 2007

Encadré 1 : critères de qualité des recommandations pour la pratique clinique

Les recommandations doivent être :

1) développées par ou en collaboration avec des groupes de praticiens, selon un processus *multidisciplinaire* dont aucune des parties concernées par le thème ne doit être exclue, afin que tous les points de vue soient examinés ;

2) *valides*, car *fondées sur la totalité des informations disponibles* : preuves scientifiques publiées dans la littérature, opinions d'experts, éventuellement investigations complémentaires (enquêtes) ;

3) *documentées* selon une méthodologie explicite. Les recommandations doivent être *argumentées* et *vérifiables*. Les incertitudes (insuffisance de données scientifiques, impossibilité de parvenir à un accord professionnel) doivent être explicitées. Tous les moyens utilisés pour élaborer les recommandations doivent être décrits : stratégie de recherche documentaire, méthode de sélection et d'analyse de la littérature, gradation du niveau de preuve des études retenues, gradation des recommandations, noms et qualités des experts consultés et des personnes ayant réalisé le travail, financement ;

4) *détaillées* en ce qui concerne les situations cliniques et les contextes de soins dans lesquels elles s'appliquent (médecine ambulatoire, hôpital, bloc opératoire, services d'urgences, etc.), les types de patients concernés, les moyens nécessaires en personnel qualifié, en équipement et en structures ;

5) *spécifiques* d'une situation clinique précise. Les situations faisant exception, connues ou attendues, doivent être identifiées, permettant ainsi une certaine liberté d'action dans l'application des recommandations, ce qui définit leur *flexibilité* ;

6) *claires* dans leur rédaction et dans leur présentation : elles doivent être aisément utilisables en pratique quotidienne et être interprétées de la même façon par toutes les personnes cibles. La terminologie employée doit être adaptée à la cible visée ;

7) *applicables* en pratique : elles doivent être adaptées aux moyens disponibles et préciser les besoins humains, matériels et organisationnels (formation, planification) qu'elles nécessitent ;

8) *diffusées* largement auprès de tous les professionnels (et patients) concernés ;

9) *régulièrement révisées* afin de ne pas devenir obsolètes alors même qu'elles doivent constituer une référence durable.

Une grille de qualité reprenant ces différentes caractéristiques a été validée au niveau international et en langue française[6].

Seules des organisations spécifiques, nationales et indépendantes sont en mesure d'élaborer des recommandations qui répondent à l'ensemble de ces critères. Il a été suggéré que la plupart des recommandations soient révisées tous les 3 ans. Toute structure mettant en place un programme de recommandations doit déterminer, *a priori*, les critères de choix de thèmes de recommandations, les méthodes d'élaboration et de diffusion et enfin les méthodes et le calendrier de révision de ces recommandations.

[6] Site internet : http://www.agree.com

Tableau 1 : Classification des recommandations selon l'American College of Chest Physicians et reprise par l'ANAES puis par la HAS (www.has-sante.fr/**)**

Le degré (grade 1 ou 2) de la recommandation est une estimation du rapport entre les bénéfices issus de la mise en œuvre de la recommandation et ses risques et coûts.
Le niveau A, B ou C prend en compte la preuve scientifique apportée par l'analyse de la littérature.

Dans cette classification, le niveau 1C+ apparaît plus fort que le niveau 1B.
C+ par rapport à C signifie que les auteurs estiment qu'on peut en sécurité extrapoler les résultats d'un essai d'une population à une autre, ou que les données issues d'études observationnelles sont convaincantes. Le niveau C signifie cependant que la preuve n'est pas directement apportée par les résultats d'un essai contrôlé randomisé.

En cas d'essais à faibles échantillons, ou de résultats contradictoires, ou si les études sont de mauvaise qualité, le grade passe toujours du niveau A au niveau B.
Lorsque le taux d'événements est faible, ou que les résultats ne sont pas statistiquement significatifs, ou que l'addition d'un petit nombre d'effet adverses au bras traité rendrait le résultat non significatif, ou que l'importance de l'effet est faible, le grade est systématiquement dégradé du niveau 1 au niveau 2.

Tableau 2 : Gradation des recommandations utilisée par l'ANAES

Recommandation de grade A	Recommandation fondée sur une **preuve scientifique** établie par des études de fort niveau de preuve (par exemple : essais comparatifs randomisés de forte puissance et sans biais majeur, méta-analyse d'essai comparatifs randomisés, analyse de décision fondée sur des études biens menées)
Recommandation de grade B	Recommandation fondée sur une **présomption scientifique** établie par des études de niveau de preuve intermédiaire (par exemple : essais comparatifs randomisés de faible puissance, études comparatives non randomisées bien menées, études de cohorte)
Recommandation de grade C	Recommandation fondée sur des **études de niveau de preuve inférieur** (par exemple : études cas-témoins, séries de cas)

Glossaire

LEXIQUE

A

- AJUSTEMENT (adjustment):

Ensemble des procédures ayant pour objectif d'éliminer l'effet de certaines variables, considérées comme parasites, dans l'étude de la relation entre un facteur que l'on étudie et un critère de jugement. Le terme d'ajustement est réservé aux procédures utilisées après le recueil des données, dans l'analyse des résultats (par régression ou standardisation ou stratification).

- ALEATOIRE (random):

  - Répartition aléatoire: détermination au hasard (selon un plan prédéterminé, par exemple par tirage au sort) du groupe auquel est affecté le sujet qui est inclus dans l'étude (voir randomisation).

  - Variable aléatoire: variable dont les fluctuations obéissent à une loi donnée.

- ANALYSE COÛT-BENEFICE (cost-benefit analysis):

Type d'analyse visant à comparer le coût d'une stratégie et son bénéfice. Dans une telle étude, les coûts et les conséquences sont exprimés en unités monétaires.

- ANALYSE COÛT-EFFICACITE (cost-effectiveness analysis):

Type d'analyse permettant de comparer des stratégies qui diffèrent par leurs coûts et leurs effets. Elle s'exprime en unités monétaires par indicateur d'efficacité médicale (nombre d'années de vie sauvées, nombre de cas de maladie évités).

- ANALYSE COÛT-UTILITE (cost-utility analysis):

Type d'analyse comparable à l'analyse coût-efficacité, dont l'indicateur de résultat médical intègre plusieurs dimensions. L'indicateur généralement utilisé est le QALY (Quality Adjusted Life Years) dont le critère d'efficacité est le nombre d'années de vie sauvées, ou l'espérance de vie, ajusté sur un critère de qualité de vie.

- ANALYSE DE SURVIE (survival analysis)

Techniques statistiques d'étude de la survie et de recherche des effets de divers facteurs (traitement, expositions, facteurs pronostiques).

- ANALYSE EN INTENTION DE TRAITER (intention to treat analysis)

Méthode utilisée dans les essais comparatifs randomisés. Les patients de chaque groupe thérapeutique restent analysés dans ce groupe qu'ils aient ou non reçu / poursuivi le traitement.

- APPARIER (to match):

Consiste à rendre comparables deux ou plusieurs groupes de sujets en vue d'étudier les effets d'un facteur donné, et ces effets seuls, sur un critère de jugement. Cette technique vise à

neutraliser les facteurs de confusion éventuels, que sont en général l'âge, le sexe, la catégorie socio-professionnelle ou le groupe ethnique, en choisissant pour chaque individu inclus dans l'étude un ou plusieurs témoins semblables à lui pour ces caractéristiques. Dans l'appariement par paires (matching) les cas et les témoins forment des paires.

- ARBRE DE DECISION (decision tree):

Outil conçu pour déterminer la décision à rendement maximal quand différentes options sont possibles au cours d'un processus clinique à étapes (diagnostique ou thérapeutique). Il s'agit d'un graphique où à chaque étape les options possibles sont représentées par des branches, tout comme les conséquences de ces options.

Représentation graphique d'une analyse décisionnelle où chaque embranchement (ou nœud de décision) correspond à 2 (ou plusieurs) décisions possibles, chacune aboutissant à 2 (ou plusieurs) éventualités.

- ASSOCIATION (relation ou corrélation statistique) (association)

C'est une dépendance statistique entre deux ou plusieurs événements ou variables. L'association peut être positive ou négative.

- ASSORTIMENT (matching):

Voir appariement.

- AVEUGLE (blind):

Procédure destinée à éviter un biais de mesure ou un biais d'information. Synonyme: insu.

On distingue:

- les essais en simple aveugle, dans lesquels le patient ignore à quel groupe il appartient (groupe expérimental ou témoin), et par conséquent à quel type de traitement il a été soumis (par exemple, médicament ou placebo);

- les essais en double aveugle, dans lesquels le patient et l'expérimentateur ignorent tous deux si le patient appartient au groupe expérimental ou témoin. Cette procédure permet de réduire non seulement l'effet placebo dû à la subjectivité du patient, mais aussi les biais dus à la subjectivité de l'expérimentateur.

- Dans un même essai ouvert le patient et l'expérimentateur sont informés du groupe auquel appartient le patient. Dans les essais en triple aveugle, ni le patient, ni l'expérimentateur, ni le médecin chargé de surveiller le patient ne savent à quel groupe ce dernier appartient.

B

- BIAIS (bias):

Il s'agit de toute erreur systématique qui s'introduit dans une étude, au stade de la collecte ou de l'analyse des données, et de l'interprétation des résultats, et qui contribue à produire des estimations systématiquement plus élevées ou plus basses que la valeur réelle des paramètres que l'on étudie.

- BIAIS DE CONFUSION (confounding bias):

On regroupe sous ce terme toutes les erreurs qui peuvent survenir lorsque la mesure de la relation entre le facteur étudié et le critère de jugement est modifiée par un facteur, appelé facteur de confusion, lié à la fois au facteur étudié et au critère de jugement.

- BIAIS DE MESURE (measurement bias):

On regroupe sous ce terme toutes les erreurs systématiques qui peuvent s'introduire dans la mesure des phénomènes pris en compte chez les sujets qui entrent dans l'étude.

- BIAIS DE NON-REPONSE (response bias)

Erreur due aux différences de caractéristiques des personnes qui acceptent ou non d'entrer dans une étude. Les biais de non-réponse (patients inclus dans une étude ne répondant pas) peuvent représenter un biais de sélection important.

- BIAIS DE SELECTION (selection bias):

On regroupe sous ce terme toutes les erreurs qui peuvent faire que les sujets effectivement observés dans l'étude ne constituent pas un groupe représentatif des populations que l'on voulait étudier et ainsi ne permettent pas la généralisation des résultats.

C

- CAS RAPPORTES ET SERIES DE CAS (case report et case series):

Les cas rapportés, en décrivant une observation inhabituelle, constituent souvent la première étape de la reconnaissance d'une nouvelle maladie ou d'un nouveau facteur de risque. Les séries de cas représentent l'étape suivante en regroupant différentes observations similaires et en établissant ainsi l'existence probable d'une entité pathologique.

Les séries de cas et les cas rapportés traduisent avant tout l'expérience et l'observation d'un auteur et ne permettent pas de tirer de conclusions qu'on puisse généraliser à d'autres cas. Ils ne permettent pas d'établir la fréquence d'une maladie -ce qui nécessiterait une étude d'incidence ou de prévalence. Ils ne permettent pas non plus d'apprécier de manière statistique l'importance d'un facteur de risque qu'ils peuvent éventuellement suggérer.

- COHORTE (cohort):

Ensemble de sujets ayant vécu un événement semblable (par exemple, facteur étudié), et suivis dans le temps depuis la date de cette expérience. Le suivi de la cohorte est conçu pour

enregistrer un ou plusieurs événements donnés (ou critères de jugement), par exemple le décès avec sa cause.

- CO-INTERVENTION (cointervention):

Attitude particulière de l'investigateur vis-à-vis des patients inclus dans une étude, ou prise d'un traitement concomitant ayant des effets identiques, et pouvant résulter en une véritable intervention non planifiée qui peut s'avérer être la véritable raison d'une différence observée entre les groupes étudiés.

- COMPLIANCE (compliance):

Un patient est compliant lorsqu'il suit les consignes données par le corps médical sur le traitement instauré ou l'intervention proposée. On parle aussi d'adhérence au traitement. Voir aussi observance.

- CONFUSION (confounding):

Situation dans laquelle la mesure de la relation entre le facteur étudié et le critère de jugement est modifiée par un facteur, appelé facteur de confusion, lié à la fois au facteur étudié et au critère de jugement. Cette situation produit un biais, le biais de confusion.

- CORRELATION (correlation)

Il y a corrélation entre 2 variables quantitatives a et b lorsque les points représentant ces valeurs de l'une par rapport à l'autre se situent au voisinage d'une ligne droite. Ces points ne sont pas alignés, mais situés dans un nuage. La droite est construite selon la méthode des moindres carrés.

- COTE (odd):

Voir Rapport de cotes.

- COURBE ROC (Receiver Operating Characteristic curve):

Moyen graphique d'évaluation de la capacité d'un test diagnostique à différencier les sujets sains des sujets malades.

- CRITERE DE JUGEMENT (outcome factor):

Dans une étude analytique, le critère de jugement, ou facteur résultant, se définit comme la situation ou l'événement supposé être le résultat de l'influence du facteur étudié.

Les événements dignes d'intérêt, pour le patient comme pour le clinicien ou l'épidémiologiste, se répartissent en cinq catégories: la mort, la maladie, le handicap, l'inconfort, l'insatisfaction, auxquels on peut ajouter un sixième élément, la destitution, qui a une dimension plus sociale. Il peut aussi s'agir de leur inverse: survie, guérison, …

La définition de ces critères de jugement doit être aussi précise que possible. Un critère de jugement est une variable, dite dépendante, par rapport au facteur étudié qui est la variable indépendante.

- CROSS-OVER

Essai thérapeutique constitué de deux groupes où le sujet est son propre témoin. Le premier groupe reçoit dans un premier temps le traitement A puis le traitement B alors que les patients du second groupe reçoivent le traitement B puis le traitement A. Avec pour les deux groupes avec une période de « wash-out » (sans traitement) entre les deux afin de ne pas superposer leurs effets.

D

- DEGRE DE SIGNIFICATION (level of significance):

Voir valeur de p.

- DISTRIBUTION DE GAUSS (normal distribution):

La distribution est le résumé le plus complet d'une mesure quantitative faite sur un groupe d'individus. Elle révèle combien d'individus, ou quelle proportion du groupe, présentent chaque valeur parmi toutes les valeurs possibles que la mesure quantitative peut donner.

La distribution normale, ou gaussienne, est un cas particulier. La courbe de Gauss est symétrique, en forme de cloche. Elle possède la propriété mathématique d'avoir les deux tiers des mesures observées dans un écart-type, et environ 95 % des observations dans deux écarts-types. Il est souvent admis, pour des raisons pratiques, que les mesures en clinique ont une distribution gaussienne.

NB. Le terme "normal" n'est pas à prendre dans son sens clinique de "non malade" ou son sens de "commun" ou "ordinaire", mais au sens de "conforme à une règle ou à un modèle".

E

- ECART REDUIT

C’est le rapport de l’écart observé sur l’écart-type. Voir test statistique et écart-type.

- ECART-TYPE ou DEVIATION STANDARD (standard deviation):

C’est la racine carrée positive de la variance. Il s'agit de l'unité d'écart permettant d'exprimer un écart observé dans une unité non liée au contexte. C’est une mesure de la dispersion ou de la variation exprimant la distance moyenne des valeurs d’une variable par rapport à la moyenne de ces valeurs.

- ECHANTILLON

  - Echantillon représentatif (representative sample):

Un échantillon est représentatif d’une population s’il a les mêmes caractéristiques que cette population. Les conclusions d'une étude réalisée sur l'échantillon pourront alors être généralisées à l'ensemble de la population d'où est tiré l'échantillon.

- Echantillon aléatoire simple (random sample)

L'échantillon est dit aléatoire simple si tous les individus ont la même probabilité de figurer dans l'échantillon.

- EFFET CARRY OVER (carry-over effect)

Il s'agit de la poursuite de l'effet d'un médicament après son arrêt.

- EFFET PLACEBO (placebo effect)

Effet généralement bénéfique, attribué par les patients à un médicament ou à une intervention.

- EFFICACITE:

  - Efficacité potentielle ou théorique (efficacy): avantage théorique que présente l'action de santé considérée. Il s'agit d'une approche théorique, évaluée dans des conditions idéales, expérimentales. Efficacité théorique ne signifie pas nécessairement efficacité réelle.

  - Efficacité réelle (effectiveness): efficacité d'une action de santé mesurée par l'adéquation entre les résultats obtenus et les objectifs fixés lors de la mise en route de cette action de santé. Il s'agit d'une approche pragmatique, dans les conditions réelles.

  - Efficience (efficiency): efficacité d'une action de santé mesurée par la comparaison des résultats obtenus aux moyens mis en œuvre pour arriver à ces résultats optimaux. C'est la notion de capacité de rendement.

- EPIDEMIOLOGIE (epidemiology)

« Etude de la distribution et des déterminants d'états de santé dans les populations données, et application de cette étude à la lutte contre les problèmes de santé. »[1]

- ERREUR ALEATOIRE (random error):

Déviation d'une valeur estimée d'un paramètre à partir de l'observation d'un échantillon, par rapport à sa valeur réelle dans la population d'où provient l'échantillon. Il s'agit des fluctuations d'échantillonnage dues au hasard (par opposition aux biais).

- ERREUR DE TYPE I ou DE PREMIERE ESPECE (type I error):

C'est conclure, lorsque qu'on fait un test statistique, qu'il y a une différence entre les groupes étudiés alors qu'en réalité il n'y en a pas. La probabilité de cette erreur est le risque α.

- ERREUR DE TYPE II ou DE DEUXIEME ESPECE (type II error):

C'est conclure, lorsque l'on fait un test statistique, qu'il n'existe pas de différence entre les groupes étudiés alors qu'il en existe une en réalité. La probabilité de cette erreur est le risque β (voir puissance).

- ESSAI (trial):

[1] John M. Last. *Dictionnaire d'Epidémiologie*, Edisem inc. 2004

Expérimentation à visée interventionniste - intervention thérapeutique, de dépistage, de prévention ou d'éducation - dans laquelle l'évaluation des effets sur le groupe de sujets étudiés se fait par rapport à un groupe témoin de référence: on peut comparer une intervention à l'absence d'intervention, ou à une ou plusieurs autres interventions du même type, l'objectif étant de déterminer s'il existe une différence entre elles.

- ESTIMATION (estimate):

Estimer un paramètre (inconnu) caractéristique d'une population consiste à en proposer une valeur, appelée estimation, calculée à partir d'un échantillon. Une estimation est donnée avec son intervalle de confiance.

Exemple: la fréquence d'une maladie, observée sur un échantillon obtenu par tirage aléatoire simple, est une estimation de la fréquence de la maladie dans la population.

- ETUDE ANALYTIQUE (analytic study):

Type d'étude spécialement conçue pour vérifier une hypothèse qui a été soulevée par les résultats d'une étude descriptive, en cherchant à établir l'association possible entre un ou plusieurs facteurs étudiés et un ou plusieurs critères de jugement.

- ETUDE CAS-TEMOINS ou CAS-CONTROLES (case-control study):

Une étude cas-témoins est une étude d'observation, rétrospective, de type analytique, dans laquelle les caractéristiques de patients atteints d'une maladie (les cas), sont comparées avec celles de patients indemnes de la maladie (les témoins), ces caractéristiques étant en général l'exposition à des facteurs de risque que l'on recherche dans le passé des cas comme dans celui des témoins.

- ETUDE DE COHORTE (cohort study):

Une étude de cohorte est une étude d'observation, prospective, de type analytique, dans laquelle un groupe de sujets exposés à des facteurs de risque d'une maladie est suivi pendant une période de temps donnée. Le taux d'incidence de la maladie dans ce groupe exposé est comparé à celui d'un groupe témoin, suivi pendant le même temps, mais non exposé aux facteurs de risque.

- ETUDE DESCRIPTIVE (descriptive study):

Toute étude qui a pour but de rendre compte d'un phénomène de santé, de sa fréquence, de sa distribution et de son évolution, ainsi que des variables apparentées, au sein d'une population donnée. Ce type d'étude permet, entre autres, d'identifier certains groupes chez qui la fréquence de ce phénomène est anormalement élevée - ou anormalement faible - afin de déterminer leurs besoins en matière de santé.

- ETUDE D'OBSERVATION (observational study):

Etude caractérisée par l'absence d'intervention de la part du chercheur. Ce type d'étude vise, soit à décrire l'état de santé de la population ou ses déterminants, soit à étudier les relations entre l'état de santé d'une population et l'exposition à un ou plusieurs facteurs, celle-ci n'étant pas contrôlée. Elle ne permet pas d'établir de lien de causalité.

- ETUDE ECOLOGIQUE ou ETUDE DE CORRELATION (ecologic study ou correlational study):

Elles établissent la comparaison entre l'importance (ou la fréquence) d'un facteur de risque supposé au sein d'une population et la prévalence ou l'incidence de la maladie supposée secondaire, à partir de données déjà disponibles au niveau de cette population. Elles sont transversales dans la mesure où elles superposent deux types de données (fréquence du facteur de risque et fréquence de la maladie) recueillies dans une même période de temps.

- ETUDE EXPERIMENTALE (experimental study):

Etude dans laquelle l'expérimentateur choisit, dans un groupe donné, les sujets qui seront soumis à l'action du facteur étudié (intervention médicale, chirurgicale…), maîtrise dans la mesure du possible l'effet d'autres facteurs concomitants et mesure les variations de l'état des sujets soumis à l'expérience afin d'établir un lien de causalité entre le facteur étudié et l'effet mesuré. Quand les exigences de l'éthique médicale la permettent, l'étude expérimentale est préférable à l'étude d'observation, car elle permet d'établir la preuve directe d'une relation causale, l'exposition au facteur étudié étant contrôlée.

- ETUDE LONGITUDINALE (longitudinal study):

Etude qui consiste à suivre une ou plusieurs cohortes à l'aide d'examens périodiques et répétés pendant une assez longue période, contrairement à ce qui se passe dans l'étude transversale où la durée est suffisamment courte pour que l'élément temps devienne une donnée négligeable.

- ETUDE PROSPECTIVE (prospective study):

Etude qui comporte la récolte de données sur des événements à venir. Elle consiste généralement à suivre un groupe de sujets exposés à un facteur de risque particulier, afin d'étudier les phénomènes de santé qui affectent ce groupe au cours du temps. Le facteur étudié est enregistré avant que se produise le critère de jugement.

- ETUDE RETROSPECTIVE (retrospective study):

Etude dans laquelle on recherche un lien possible entre un phénomène de santé (maladie ou autre) présent au moment de l'étude et des événements (facteurs de risque) survenus dans le passé. Par exemple, on examine les antécédents d'un groupe de patients afin d'identifier les sujets qui, dans le passé, ont subi une exposition à un facteur de risque particulier. L'information concernant les événements passés est obtenue au moyen de documents d'archives ou par l'interrogatoire. Le facteur étudié est enregistré rétrospectivement, après que le critère de jugement s'est produit.

- ETUDE TRANSVERSALE ou ETUDE DE PREVALENCE (cross-sectional study):

Cette étude analyse la présence d'un facteur donné ou d'une maladie particulière dans une population P à un moment précis t, sans référence au passé ou sans suivi dans le futur. Elle représente l'équivalent d'un sondage rigoureusement et scientifiquement construit, ou de l'instantané photographique d'une situation précise dans la population étudiée.

Ce type d'études est surtout effectué dans un but descriptif. Il pose des problèmes d'interprétation liés au manque d'information sur la chronologie des événements et à des difficultés de sélection.

- EXPOSITION (exposure):

Proximité ou contact avec un facteur qui peut avoir un effet nuisible ou protecteur sur le sujet exposé.

F

- FACTEUR DE CONFUSION ou CONFONDANT (confounding factor):

Toute variable liée à la fois au facteur de risque étudié et à la maladie, et susceptible de modifier la liaison statistique entre ces derniers. Il est essentiel d'identifier les facteurs de confusion potentiels et d'en tenir compte au moment de la planification de l'étude ou de son analyse.

- FACTEUR DE RISQUE (risk factor):

Caractéristique individuelle ou collective, endogène (propre à l'individu) ou exogène (liée à l'environnement), qui augmente la probabilité de survenue d'une maladie ou de tout autre phénomène de santé. Le lien est purement statistique et ne préjuge en rien de la causalité.

- FACTEUR ETUDIE (study factor):

Dans une étude analytique, le facteur étudié est l'exposition ou l'intervention supposée avoir des conséquences sur un problème de santé, une maladie ou un état clinique.

- FACTEUR PRONOSTIQUE (prognostic factor):

Un facteur pronostique est l'état, la situation ou l'événement qui, quand il est observé chez un sujet qui présente déjà un état pathologique, est associé avec une conséquence de cet état pathologique.

- FAUX NEGATIF (false negative):

Test diagnostique négatif (normal) chez un sujet qui a la maladie.

- FAUX POSITIF (false positive):

Test diagnostique positif (anormal) chez un sujet qui n'a pas la maladie.

G

- GOLD STANDARD (gold standard)

Test considéré comme indiscutable pour déterminer l’état de maladie d’un patient (indépendamment des autres tests diagnostiques) ou mesurer l’efficacité supérieure d’un médicament ou d’une thérapie.

- GROUPE CONTROLE ou GROUPE TEMOIN (control group)

Dans le cas d’un essai contrôlé, il s’agit du groupe qui reçoit le médicament de référence ou le placebo.

H

- HYPOTHESE (hypothesis):

Peut être définie comme une supposition, née d'une observation ou d'une réflexion, qui conduit à des prédictions que l'on peut réfuter. En d'autres termes, l'hypothèse est avant tout l'expression d'une opinion. Une telle opinion peut être plus ou moins forte, plus ou moins hardie. Mais quoi qu'il en soit, l'objectif d'une étude analytique est de démontrer, à partir des données, et avec des méthodes statistiques appropriées, si cette supposition est une description exacte de la relation qui existe entre les facteurs que l'on étudie, ou si elle ne l'est pas.

- HYPOTHESE ALTERNATIVE (alternative hypothesis) :

C'est l'hypothèse qui sera acceptée si l'hypothèse nulle est rejetée (voir hypothèse nulle).

- HYPOTHESE NULLE (null hypothesis):

Correspond à l'hypothèse selon laquelle il n'y a pas de différence entre les groupes étudiés pour une variable donnée. Dans un test statistique, il s'agit de l'hypothèse qui est "soumise à épreuve". Si l'hypothèse nulle est rejetée, alors on retient l'hypothèse alternative avec un risque d'erreur α.

I

- INCIDENCE (incidence):

Nombre de nouveaux cas d'une maladie donnée ou de personnes qui sont atteintes de cette maladie, pendant une période donnée, dans une population déterminée.

- INCLUSION - NON INCLUSION (CRITERES D') (inclusion - exclusion criteria):

  - Les critères d'inclusion sont les caractéristiques que les sujets doivent obligatoirement présenter pour être inclus dans une étude donnée.

  - Les critères de non inclusion (parfois appelés critères d'exclusion), sont les critères qui interdisent l'inclusion d'un sujet, bien qu'il satisfasse aux conditions de la première liste.

Tout sujet entrant dans l'étude devra satisfaire à tous les critères d'inclusion, et ne présenter aucun des critères de non-inclusion. Les critères d'exclusion sont valables durant toute la durée de l'étude.

- INDICATEURS DE SANTE (health indicator):

Descripteur quantitatif d'un phénomène donné. Dans la pratique ce sont plutôt des critères de « non-santé ». Exemple : le nombre de jours d'arrêt de travail pour maladie, le taux de mortalité infantile.

- INSU (blind):

Voir aveugle.

- INTERACTION (interaction)

Action réciproque de 2 ou plusieurs causes pour produire ou prévenir un effet. L'interaction entre deux causes peut être synergique ou antagoniste.

K

- KAPPA (COEFFICIENT DE CONCORDANCE):

Un des coefficients les plus utilisés pour mesurer le degré de concordance entre des juges classant les mêmes individus (par exemple, des patients) dans une ou plusieurs catégories préalablement définies (par exemple, des diagnostics). On peut également calculer un kappa sur plusieurs mesures du même phénomène évaluées par un même juge (kappa intra-observateur).

M

- MEDIANE (median):

Valeur centrale qui sépare la population en deux parties d'effectifs égaux. Dans le cas d'effectifs impairs, la médiane est la moyenne des deux valeurs.

- MESURE (measurement):

Quantification d'un phénomène en utilisant un instrument ou en appliquant une échelle de mesure standard sur une variable ou un ensemble de valeurs.

- MODE (mode):

Valeur la plus fréquente pour une variable discrète.

Classe correspondant au maximum de l'histogramme pour une variable continue.

- MOYENNE (mean):

Caractéristique de valeur centrale, pour une distribution observée ou théorique. Si n est le nombre de valeurs observées, la moyenne arithmétique est la somme des valeurs observées divisée par le nombre de valeurs observées.

$x_1, x_2, \ldots\ldots\ldots, x_n: \dfrac{x_1 + x_2 + \ldots\ldots + x_n}{n}$

O

- OBSERVANCE (compliance):

Le fait de se conformer, pour le sujet inclus dans l'étude, aux prescriptions du corps médical, et pour le chercheur au protocole de recherche. Synonyme: compliance.

P

- PERCENTILE (centile):

Voir quantile.

- PERDU DE VUE (lost to follow-up):

Sujet dont on n'a plus de nouvelles et dont on a perdu la trace au cours d'une étude prospective et pour lequel les informations le concernant ne seront pas disponibles au moment de l'analyse.

- PHASE I (phase I):

Les essais de phase I visent à préciser la sûreté et la tolérance; ils sont faits chez un petit nombre de sujets.

- PHASE II (phase II):

Les essais de phase II précisent l'efficacité optimale du traitement.

- PHASE III (phase III):

Les essais de phase III établissent l'efficacité du traitement, le plus souvent grâce à des essais thérapeutiques comparatifs, idéalement randomisés.

- PHASE IV (phase IV):

Les essais de phase IV, après la commercialisation, visent à préciser les effets du traitement à long terme.

- PLACEBO (placebo):

Procédure ou substance inactive ou inerte que l'on substitue à la vraie procédure ou à la vraie substance médicamenteuse pour contrôler ou induire les effets psychologiques accompagnant la médication.

- POPULATION (population):

En statistique, une population est un ensemble d'unités susceptibles d'être observées (qui peuvent ne pas être des personnes physiques).

En épidémiologie, il convient de distinguer la population examinée, que l'on peut également qualifier d'échantillon, de la population d'où provient cet échantillon (à condition qu'on puisse la définir).

Dans l'un et l'autre cas, la population doit être, dans la mesure du possible, définie et décrite de manière précise.

- PRECISION (precision):

Propriété d'une mesure rendant compte de la dispersion des valeurs qu'elle fournit autour de sa moyenne pour une grandeur donnée. La mesure sera d'autant plus précise que la dispersion des valeurs autour de la moyenne sera faible.

- PREVALENCE (prevalence):

Nombre de cas d'une maladie donnée ou de personnes atteintes de la maladie ou de tout autre événement morbide, dans une population déterminée à un moment donné ou au cours d'une période donnée, sans distinction entre les cas nouveaux et les cas anciens.

- PREVENTION PRIMAIRE (primary prevention):

Ensemble de mesures destinées à diminuer l'incidence d'une maladie dans une population (en évitant cette maladie à des sujets qui n'en sont pas atteints). Il peut s'agir de mesures qui visent à augmenter la résistance des individus ou à éliminer les facteurs de risque de la maladie.

- PREVENTION SECONDAIRE (secondary prevention):

Ensemble de mesures destinées à diminuer la prévalence d'une maladie dans une population (en réduisant la durée de la maladie). Assurée par les organismes de santé publique et par le corps médical, elle se fonde sur le dépistage des maladies à un stade précoce, pendant que celles-ci peuvent être facilement curables.

- PREVENTION TERTIAIRE (tertiary prevention):

Ensemble des mesures destinées à diminuer les complications dégénératives d'une maladie ou les incapacités chroniques. Elle vise à diminuer la prévalence des incapacités chroniques.

- PROBABILITE *A POSTERIORI* OU PROBABILITE POST-TEST (posterior probability):

C'est la probabilité de la maladie étant donné le résultat du test (voir valeurs prédictives).

- PROBABILITE *A PRIORI* ou PROBABILITE PRE-TEST (prior probability):

C'est la fréquence de la maladie dans la population dont est issu le patient soumis à l'examen (c'est la prévalence lorsqu'il s'agit d'une population importante).

- PROTOCOLE (protocol):

Description de l'ensemble des étapes nécessaires pour la réalisation d'une étude. Dans une étude, la mise au point du protocole est au moins aussi importante que le recueil des données, car la validité des résultats dépend de cette mise au point.

- PUISSANCE (power):

La puissance d'un test (ou d'une étude) est la probabilité de conclure, à partir de l'échantillon, à une différence ou à un effet statistiquement significatif lorsque cette différence ou cet effet existe réellement dans la population (puissance = 1 - β). Une étude manque de puissance lorsque la conception de son protocole ou les conditions de sa réalisation ne permettent pas de mettre en évidence un effet ou une différence, alors que ceux-ci existent réellement dans la population.

Q

- QUANTILE (quantile):

Les quantiles d'une variable quantitative permettent de diviser la population en groupes d'effectifs égaux. Les quartiles divisent la population en quatre groupes d'effectifs égaux, les quintiles en cinq, les déciles en dix et les percentiles en cent.

- QUESTIONNAIRE (questionnaire):

Ensemble de questions prédéterminées pour récolter des données d'ordre clinique, social, comportemental ou de santé publique.

R

- RANDOMISATION ou REPARTITION ALEATOIRE (randomization, random allocation):

Répartition d'un échantillon ou d'une population en deux ou plusieurs groupes comparables, à l'aide d'une méthode fondée sur le hasard. La répartition est effectuée par tirage au sort, notamment à l'aide des tables de nombres au hasard.

C'est la meilleure solution pour contrôler l'influence des facteurs de confusion, qu'ils soient connus ou inconnus, en répartissant également ces variables dans les différents groupes, de sorte que leur effet sur le critère de jugement s'annule et qu'ainsi, si une différence est observée entre les deux groupes, elle puisse être attribuée à l'effet du facteur étudié.

- RAPPORT DE COTES (odds ratio):

C'est une estimation indirecte du risque relatif, ce dernier n'étant pas disponible dans les études cas-témoins. Il correspond au rapport de deux cotes qui sont respectivement le rapport des exposés sur les non exposés chez les malades et le rapport des exposés sur les non exposés chez les non malades.

Cote (odd) d'exposition parmi les cas (les malades):

$$\frac{\text{Proportion de cas exposés}}{\text{Proportion de cas non exposés}} = \frac{a / (a+b)}{b / (a+b)} = \frac{a}{b}$$

Cote (odd) d'exposition parmi les témoins (les non-malades):

$$\frac{\text{Proportion de témoins exposés}}{\text{Proportion de témoins non exposés}} = \frac{c / (c+d)}{d / (c+d)} = \frac{c}{d}$$

$$\text{Rapport de cotes} = \frac{\text{Cote d'exposition parmi les cas}}{\text{Cote d'exposition parmi les témoins}} = \frac{a / b}{c / d} = \frac{ad}{bc}$$

Le rapport de cotes exprime le risque relatif lorsque la maladie est rare, et se calcule très simplement par le rapport des produits en croix de la table 2×2.

- RATIO (ratio):

Rapport, parfois exprimé sous forme de pourcentage, entre deux grandeurs, deux quantités, ayant chacune une nature distincte. Exemple: sex-ratio. Le ratio est différent du taux dont le dénominateur contient le numérateur.

- REGRESSION (regression):

Modèle mathématique permettant de décrire une variable dépendante comme une fonction d'une ou plusieurs variables indépendantes. Le modèle le plus commun est le modèle linéaire. Lorsque la variable dépendante a deux modalités, on utilise un modèle de régression logistique.

- REPRODUCTIBILITE (reproductibility):

Un test est reproductible s'il donne les mêmes résultats dans des situations identiques. La quantification de la reproductibilité d'un test est fondée sur le coefficient KAPPA pour les variables quantitatives.

- RISQUE (risk):

Probabilité de survenue d'un événement pendant une période donnée. En épidémiologie, cette probabilité varie en fonction de certaines caractéristiques endogènes (âge, sexe, hérédité, …) ou exogènes (milieu, profession, …) qui peuvent constituer des facteurs de risque. La mesure du risque ne permet ni un diagnostic, ni une évaluation certaine du pronostic individuel.

- RISQUE ABSOLU (absolute risk):

Probabilité de survenue d'un événement exprimée sans référence à une autre probabilité. S'il s'agit de la probabilité d'être atteint d'une maladie, le risque absolu sera souvent estimé d'après le taux d'incidence de la maladie au sein de la population.

- RISQUE ATTRIBUABLE A UN FACTEUR DONNE (attributable risk):

Différence entre le taux d'incidence d'un phénomène de santé au sein d'une population exposée à un facteur donné, et le taux d'incidence de ce phénomène au sein de la population non exposée.

- RISQUE RELATIF (relative risk):

Taux d'incidence d'une maladie donnée au sein d'un groupe exposé à un facteur de risque donné, rapporté au taux d'incidence de cette maladie au sein d'un groupe non exposé. Ce rapport est généralement considéré comme une mesure de la force de l'association existant entre le facteur et la maladie. Il répond à la question: combien de fois les sujets exposés ont-ils plus de chance que les sujets non exposés de contracter la maladie ?

S

- SELECTION (selection):

Procédure qui consiste à choisir les sujets à inclure dans l'étude.

- SENSIBILITE (sensitivity):

C'est la probabilité d'avoir un test diagnostique positif (anormal) quand on a la maladie. Cela correspond au taux de vrais positifs, nombre de vrais positifs rapportés à l'ensemble des malades.

- SERIE DE CAS (case series):

Voir cas rapporté et série de cas.

- SPECIFICITE (specificity):

C'est la probabilité d'avoir un test diagnostique négatif (normal) quand on n'a pas la maladie. Cela correspond au taux de vrais négatifs, nombre de vrais négatifs rapportés à l'ensemble des non-malades.

- STANDARDISATION (standardization, adjustment):

Méthode qui rend comparables des taux bruts dans des groupes qui diffèrent par la distribution d'une autre variable (âge, sexe, niveau socio-économique, …). On distingue:

- la standardisation directe: pour l'âge, on calcule le taux que l'on observerait dans la population étudiée si elle avait la même structure d'âge qu'une population de référence;

- la standardisation indirecte: on calcule le nombre d'événements attendus (maladies, décès, …) dans la population étudiée, en appliquant à chaque classe d'âge les taux spécifiques d'une population de référence.

- STATISTIQUE (statistics):

C'est l'art de faire la collecte, la synthèse et l'analyse des données qui sont sujettes à des variations aléatoires.

- STRATIFICATION (stratification):

Répartition des sujets en sous-groupes ou strates, en fonction d'une ou plusieurs caractéristiques, de sorte que chaque strate soit homogène pour cette ou ces caractéristiques. On peut ainsi stratifier les individus en fonction de l'âge, du sexe, des deux, du niveau socio-économique, du niveau de risque, …

Cette méthode permet de rendre comparables, en procédant strate par strate, deux groupes différents.

- SUIVI (follow-up):

C'est l'observation pendant une période de temps donnée d'un individu, d'un groupe ou d'une population dont les caractéristiques ont été définies, afin de pouvoir observer et enregistrer tout changement ou tout événement concernant la santé de ces individus.

T

- TABLE 2 × 2 (2 × 2 table):

Mode de représentation, sous forme de tableau, de la relation entre deux variables qualitatives: chaque ligne représente une modalité d'une variable, chaque colonne une modalité de l'autre variable. La relation entre le résultat d'un test diagnostique et la survenue de la maladie, comme la relation entre les résultats d'un test statistique et la vraie différence entre les deux groupes, peuvent être représentés sous la forme d'une table 2 × 2.

- TAILLE DE L'ECHANTILLON (sample size):

C'est le nombre de sujets à inclure nécessairement dans l'étude afin que celle-ci ait une puissance suffisante (possibilité de conclure à une différence lorsque celle-ci existe réellement).

- TAUX (rate):

Dans un contexte épidémiologique et démographique, il s'agit:

- de l'effectif d'une sous population *n*, dénombré à une date déterminée et rapporté à l'effectif de la population *N* dont cette sous-population fait partie, soit *n/N* ;

- du nombre des événements (maladies, décès, …) *n* survenus dans une population donnée durant une période déterminée, rapporté à l'effectif de cette population *p* pour la même période, soit *n/p*.

- TAUX BRUT (crude rate):

- Taux qui rend compte uniquement d'une fréquence d'événements au sein d'une population donnée, au cours d'une période déterminée. Il permet seulement d'évaluer l'intensité d'un phénomène de santé dans une population, mais ne permet pas de comparer l'impact de ce phénomène dans des populations dont les structures (d'âge, de sexe, …) sont différentes.

- Taux calculé pour l'ensemble d'une population ou d'une fraction de cette population, exprimant la fréquence de la mortalité ou de la morbidité attribuables à une cause déterminée (par exemple, taux de mortalité par cause, taux d'incidence ou de prévalence d'une maladie donnée).

- TAUX D'ATTAQUE (attack rate):

Dans le contexte d'une épidémie, nombre de cas d'une maladie donnée survenus pendant une période d'observation déterminée, rapporté à la population exposée à cette maladie au début de la période. Celle-ci représente en général la totalité du temps pendant lequel la maladie peut se produire.

- TAUX SPECIFIQUE (specific rate):

Taux calculé pour une fraction d'une population déterminée par une ou plusieurs caractéristiques (par exemple, taux spécifique par tranche d'âge et par sexe, par catégorie socio-professionnelle, par ethnie, …).

- TAUX STANDARDISE (standardized rate; adjusted rate):

Taux calculé par des méthodes particulières (standardisation) permettant d'éliminer l'effet de certains facteurs (par exemple, l'âge, les facteurs socio-culturels, …) distribués différemment entre deux ou plusieurs populations. On peut ainsi comparer l'intensité d'un phénomène donné (la mortalité par exemple) dans des populations de structures différentes.

- TEMOIN (control):

Sujet ou groupe de sujets dont les caractéristiques servent de référence pour l'estimation de l'association entre un ou plusieurs facteurs étudiés et un ou plusieurs critères de jugement dans la population que l'on étudie.

- TEST STATISTIQUE (statistic test):

Lorsque l'on observe une différence, soit entre un paramètre observé sur un échantillon et une valeur théorique, soit entre les valeurs prises par ce paramètre sur plusieurs échantillons, se pose la question de savoir si l'écart observé est seulement dû au hasard. On ne peut jamais rejeter, de façon absolue, l'hypothèse nulle selon laquelle l'écart observé est dû au hasard, mais on sait estimer le risque d'observer un tel écart simplement par le fait du hasard, par le calcul de la statistique de test. Si ce risque est faible (en pratique inférieur à 5 %), on rejette l'hypothèse nulle.

La statistique de test est l'expression qui transforme un écart observé, tout à fait lié aux circonstances (en particulier à l'effectif de l'échantillon), en un écart exprimé dans une unité non liée au contexte, permettant la comparaison à des valeurs tabulées. Il s'agit de l'écart réduit.

- THEOREME DE BAYES (Bayes' theorem):

Moyen mathématique d'estimer un jugement correct (valeurs prédictives) sur une maladie donnée, en fonction de la prévalence de la maladie, de la sensibilité et de la spécificité du test.

V

- VALEUR DE p (p value):

La valeur de p est l'expression, en termes quantitatifs, de la probabilité que les différences observées dans une étude puissent être observées par la chance seule.

Une autre façon d'exprimer la valeur de p est de répondre à la question: "s'il n'y a en réalité pas de différence en termes d'effet de l'intervention étudiée et que l'étude était répétée de nombreuses fois, combien de fois conclurait-on cependant qu'un effet existe ?"

Il est habituel de donner une importance spéciale aux valeurs de p situées au-dessous de 0,05, parce qu'il est généralement admis qu'une chance sur vingt est un risque acceptable.

Les effets associés à une valeur de p inférieure à 0,05 sont dits statistiquement significatifs. Il faut cependant souligner que cette limite de 0,05 est totalement arbitraire. On pourrait exiger une valeur différente, plus élevée ou plus basse, selon les conséquences que l'on accepte de conclusions faussement positives. Pour $p = 0,05$, on considère que la probabilité de 1 sur 20

est trop faible pour que le résultat observé soit survenu par chance seule. L'intervention ou l'exposition étudiée explique le résultat observé.

Pour briser cette alternative, < 0,05 ou > 0,05, on peut se référer à la valeur exacte de p (par exemple, 0,00843 ou 0,07). L'interprétation de ce qui est statistiquement significatif est alors laissée au lecteur. Cette valeur numérique de p est appelée degré de signification.

- VALEUR PREDICTIVE NEGATIVE (negative predictive value):

Probabilité de ne pas avoir la maladie quand le test diagnostique est négatif (normal).

- VALEUR PREDICTIVE POSITIVE (positive predictive value):

Probabilité d'avoir la maladie quand le test diagnostique est positif (anormal).

- VALIDITE (validity):

Possibilité d'une méthode (enquête, examen, test diagnostique, dépistage) de fournir une valeur exacte de ce qu'elle est censée mesurer.

Dans le cas d'un test diagnostique, c'est la proportion des mesures qui représentent la valeur réelle de l'objet mesuré. Une bonne validité suppose un degré élevé de sensibilité et spécificité.

Dans le cas d'un instrument de mesure, on étudie plusieurs aspects de la validité, les deux premiers s'appliquant aux questionnaires:

- pertinence apparente du questionnaire (face validity): caractère sensé et rationnel des questions
- exhaustivité du questionnaire (content validity): capacité du questionnaire à mesurer tous les domaines qu'il est sensé mesurer et uniquement ceux-là
- comparaison d'un instrument de mesure à une "référence" ou "étalon" (criterion validity)
- lorsqu'aucun examen de référence n'existe, on peut évaluer l'exactitude d'un instrument en le comparant à un faisceau d'arguments ou à une mesure différente en se fondant sur l'hypothèse qu'ils mesurent la même chose ou au moins des caractéristiques très semblables (construct validity)

- VALIDITE EXTERNE (external validity):

Possibilité de généralisation des observations issues de l'échantillon à la population générale. La validité externe est menacée par le biais d'échantillonnage, lorsque les conclusions fondées sur un échantillon sont généralisées à d'autres groupes qui ne sont pas semblables à l'échantillon.

- VALIDITE INTERNE (internal validity):

Capacité intrinsèque de la méthode à fournir des observations exactes pour les sujets ayant participé à l'étude, dans les conditions particulières de cette étude. La validité interne est menacée par les erreurs systématiques (biais) et aléatoires (chance).

- VARIABLE (variable):

En général, toute grandeur susceptible de prendre n'importe quelle valeur dans un ensemble préalablement défini.

Dans un sens plus restreint, propriété ou caractère servant à distinguer les individus d'une population et pouvant être soit qualitatif (par exemple, le sexe) soit quantitatif (par exemple, l'âge).

Parmi les variables quantitatives, on distingue:

- les variables discrètes (ou discontinues) : le nombre d'enfants;
- les variables continues: le poids, la taille.

Parmi les variables qualitatives, on distingue:

- les variables nominales: le sexe, le groupe sanguin;
- les variables ordinales, à modalité ordonnée: classification TNM pour les cancers.

- VARIANCE (variance):

Caractéristique de dispersion, pour une distribution théorique ou observée.

Les auteurs remercient Peggy Falgon, Jérôme Mollard et Adeline Roux qui ont relu ce glossaire.