thétique, mais, du fait qu'il décrit bien les données, tout se passe comme s'il avait une certaine réalité, et les conclusions qu'on en a tirées méritent réflexion.

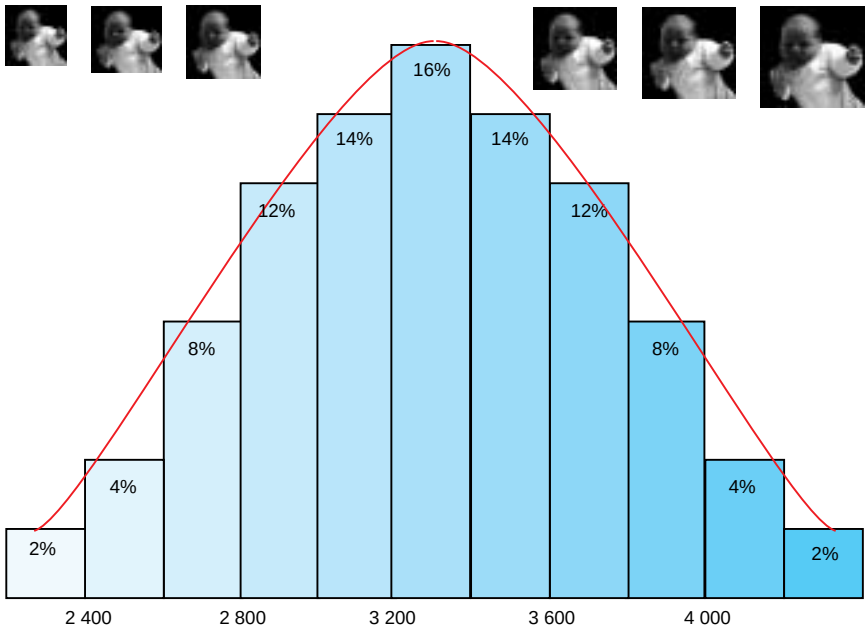
Le modèle probabiliste

L'évaluation de traitements (l'essai thérapeutique) consiste à comparer, pour la guérison d'une maladie par exemple, un groupe recevant un traitement *A* et un groupe recevant un traitement *B* (qui peut être l'absence de traitement).

La difficulté d'une telle comparaison provient de la variabilité des caractères biologiques selon les individus. Certaines personnes atteintes d'une maladie guérissent et d'autres non, et, parmi le premier groupe, le temps de maladie est variable selon les personnes. Le médecin qui veut évaluer l'efficacité d'un traitement ne peut étudier ce traitement chez une seule personne, et doit mesurer une proportion de guéris ou des durées moyennes de maladie. Comment déterminer ces valeurs? Ces moyennes vraies, ces taux vrais, ou probabilités, nous échappent nécessairement, car on ne peut tester le médicament chez tous les individus ; nous ne pouvons qu'estimer ces valeurs sur des échantillons de taille finie. Et ces estimations diffèrent plus ou moins des vraies valeurs, en raison des fluctuations d'échantillonnage, ou fluctuations de hasard : si on tire un échantillon de boules dans une urne à 20 pour cent de boules blanches, on n'observera pas nécessairement 20 pour cent de blanches. Voilà le problème : l'épidémiologiste veut comparer le taux de guéris chez des malades traités par le médicament *A* ou par le médicament *B*, ou la durée moyenne de la maladie, mais la connaissance de ces taux ou de ces moyennes est hors de portée.

Avant le XX^{e} siècle, la méthode était simple, mais absolument pas rigoureuse : on concluait selon la grandeur de la différence observée. Grande de combien? Le hasard est capable de tous les caprices : il peut conduire à des taux de guéris très différents, même si les probabilités de guérison sont les mêmes avec *A* et *B*, et à des taux voisins ou identiques, alors que les chances de guérison sont différentes. C'est ici que le calcul des probabilités donne une solution, avec le test statistique.

Supposons que, sur deux échantillons de 200 malades, on observe 100 guéris avec le traitement *A* et 60 guéris avec le traitement *B*. Le calcul



4. LE POIDS des nouveau-nés est distribué de façon «normale» : la proportion de nouveau-nés ayant un poids donné, en fonction du poids, est sous-tendue par une courbe en forme de cloche symétrique. Une telle distribution se retrouve pour plusieurs paramètres physiologiques.

des probabilités indique que, si les vrais taux de guérison étaient identiques, on aurait moins de une chance sur 10 000 d'obtenir une telle différence de taux. L'hypothèse d'égalité des vrais taux est donc très peu vraisemblable.

La différence observée est généralement considérée comme significative quand la probabilité d'atteindre la différence observée (le fameux «petit p» des statisticiens) est inférieure à cinq pour cent. Pour la probabilité précédente, de 1 pour 10 000, la différence est donc très significative.

Si la probabilité est supérieure à cinq pour cent, on ne dira pas que les vrais taux de guérison sont identiques, mais seulement qu'on n'a pas mis en évidence de différence : une telle attitude de «non conclusion» peut être illustrée par une analogie : si on trouve l'arme du meurtre chez un accusé, on pourra conclure, avec un risque d'erreur, qu'il est le criminel, mais si on ne trouve pas l'arme, on ne conclura pas qu'il est innocent.

On comprend le progrès apporté par le test statistique : on conclut non d'après l'importance de la différence observée, qui n'a pas de sens, mais d'après la probabilité d'obtenir une telle différence si les traitements sont équivalents.

La différence admise comme significative, un nouveau problème se pose : la différence peut-elle être attribuée aux traitements? Cette imputation causale n'est possible que si les deux groupes sont, à part les traitements, comparables à tous les points de vue. On peut montrer que cela n'est réalisé que si les deux groupes de malades ont été obtenus par tirage au sort dans l'ensemble des 200 malades considérés et que les groupes ainsi constitués sont restés comparables pendant l'essai.

Connaissance ou décision

Parmi de nombreuses classifications des modèles, l'une des plus importantes porte sur la distinction entre les modèles de connaissance ou de compréhension, d'une part, et les modèles de prévision ou de décision, d'autre part. La quête de connaissance et l'aide à la décision sont deux attitudes si différentes qu'elles devraient nécessiter des approches radicalement différentes. Toutefois cette distinction est souvent mal perçue, conduisant à une mauvaise formulation des problèmes.

Le protocole de tout essai visant à comparer les effets de traitements doit définir avec précision quatre mots clefs : comparer, effets, traitements, patients. Deux exemples montreront comment ces définitions peuvent revêtir des aspects très différents, y compris celle du mot «comparer» qui, contrairement aux apparences, est ambigu.

Les essais de connaissance

Soit à éprouver l'efficacité d'une nouvelle molécule, par exemple un hypocholestérolémiant ou un hypnotique. Pour attribuer une différence observée

aux traitements (problème d'imputation causale), il est essentiel que les deux groupes soient identiques, à part le traitement, et qu'ils le restent. Aussi les traitements doivent-ils être donnés avec le même mode d'administration (voie sanguine, ou bien voie orale, etc.) et sous une présentation identique (une pastille rouge n'a pas le même effet qu'une pastille verte). Plus généralement, la connaissance du traitement par le malade ou par son médecin peut provoquer, par autosuggestion ou par hétérosuggestion, des différences, non seulement dans l'appréciation des résultats, mais aussi dans l'évolution de la maladie. Toutes les fois qu'on le peut, on présentera donc les traitements sous des formes indiscernables, leur identité étant ignorée du malade (essai en «aveugle»), voire du malade et du médecin (essai en «double aveugle»). Si le groupe témoin peut ne pas être traité, on lui administrera un placebo, substance inerte imitant, dans sa présentation, le traitement à évaluer.

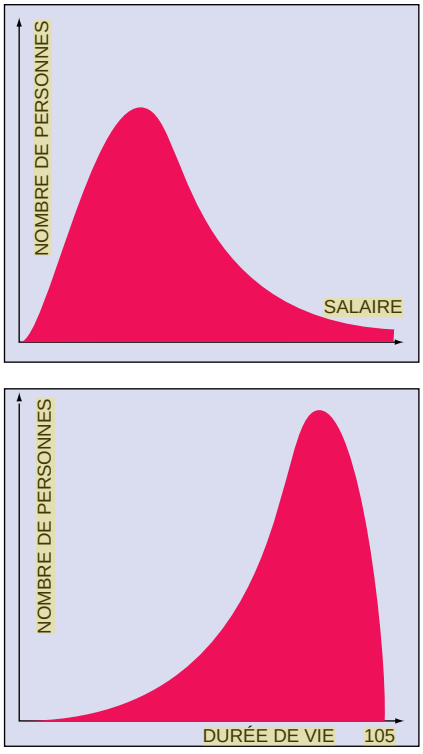
Les critères de jugement doivent être peu nombreux, un seul si possible : baisse de la cholestérolémie ou nombre d'heures de sommeil.

Les patients seront sélectionnés selon des critères d'optimisation : homogénéité, sensibilité..., le souci de représentativité étant secondaire : on teste une propriété pharmacologique vraisemblablement extrapolable d'une population à l'autre, l'essai prolongeant l'expérimentation sur l'animal. Enfin la comparaison sera effectuée par un test statistique. Cette démarche est celle de l'essai thérapeutique classique, qui est le plus couramment pratiqué.

Les essais de décision

Supposons que nous devions choisir entre deux traitements de la polyarthrite rhumatoïde : par exemple, entre un nouveau traitement de fond *N* et un traitement de fond classique *C*. Un essai de décision est entrepris pour déterminer quel traitement est le plus intéressant ou le plus utile dans ses conditions habituelles d'emploi, de manière à opter pour l'un ou l'autre, à la fin de l'essai. Les quatre mots clefs du protocole doivent alors être définis d'une nouvelle façon.

On administrera le traitement nouveau *N* pour un groupe, le traitement classique *C* pour l'autre, chacun avec ses conditions d'administration optimales : la meilleure dose (il n'est pas question de donner des doses égales, si les substances sont différentes), la meilleure voie d'administration, éventuellement les traitements adjuvants (si le régime sans sel est indiqué pour un des traitements, on ne privera pas de sel les patients de l'autre groupe). Ainsi on ne se préoccupe pas de réaliser des conditions égales, mais des conditions optimales.



5. LA DISTRIBUTION des durées de vie après 20 ans *(en bas)* peut se ramener à une distribution log-normale, telle que celle des salaires *(en haut)*, quand on considère la différence de la durée de vie à une valeur limite, égale à 105 ans dans la population féminine française, en 1969.

Pour juger des résultats, on retiendra non seulement les nombreux critères cliniques et de qualité de vie, permettant d'apprécier l'état d'un malade souffrant de polyarthrite rhumatoïde, mais aussi les inconvénients de ces traitements. Il faudra ensuite bâtir, à partir de ces critères, un score exprimant le bilan. Il est cette fois indispensable d'avoir un échantillon de patients aussi représentatif que possible de la population à laquelle on voudra appliquer les conclusions de l'essai. En effet, le résultat (*N* plus ou moins intéressant que *C*, dans un bilan global) n'est pas, comme la propriété pharmacologique testée dans le premier exemple, une conclusion de portée générale vraisemblablement extrapolable d'une population à l'autre.

Enfin la comparaison est l'élément le plus original de la démarche. Elle ne fait pas appel à un test statistique. En premier lieu, dans un bilan, l'hypothèse soumise à un test, qui serait l'équivalence des deux traitements, n'a aucune raison d'être. En second lieu, un test statistique peut conduire à une non conclusion, alors qu'il faut, en l'occurrence, aboutir à un choix : on doit dire lequel des deux traitements est le plus intéressant.

La solution est à emprunter à la théorie de la décision. À Villejuif, avec Joseph Lellouch, professeur à l'Université Paris XI, nous avons mis au point une méthode simple : elle consiste à opter, sans test statistique, pour celui des deux traitements dont le bilan dans l'essai a été le meilleur, sous la condition que le nombre de sujets de l'essai, calculé à l'avance, minimise le risque que le traitement ainsi adopté soit le moins bon.

Le premier essai, qui vise à éprouver une hypothèse biologique, a tous les caractères d'une expérience de laboratoire. Le second, qui vise à choisir le plus intéressant de deux traitements pour une population donnée, est une sorte de répétition générale, un banc d'essai des traitements en conditions réelles. La première démarche est une quête de connaissance : c'est l'attitude explicative. La seconde aboutit à une prise de décision : c'est une attitude pragmatique.

On conçoit facilement qu'un choix inapproprié à l'un quelconque des niveaux de définition peut rendre inutilisables, ou du moins limitées, les conclusions d'un essai. Citons, à titre d'exemple, la célèbre étude sur l'effet de l'aspirine dans la prévention de l'infarctus du myocarde. Sur près de 260 000 sujets sollicités, les nombreux critères exigés pour la participation à l'essai n'en laissaient subsister que 22 000, soit huit pour cent.

L'inconvénient de cette sélection n'était pas tant la réduction de l'effectif, qui restait notable, que la non-représentativité résultant des critères d'exclusion. L'essai, de type explicatif, prouva l'efficacité de l'aspirine, dans ces «conditions de laboratoire», ce qui était déjà beaucoup, mais l'extrapolation à l'ensemble de la population n'était pas possible.

En particulier, la prise d'aspirine au long cours présente aussi des inconvénients, notamment un risque d'hémorragie. Les clauses d'exclusion éliminaient, entre autres, tous les sujets pour lesquels cet événement risquait

de se produire : l'échantillonnage n'était pas satisfaisant. C'est sans doute une des raisons pour lesquelles la communauté médicale n'a pas emboîté le pas d'emblée. Au fond, la démarche logique que l'on devrait adopter est : d'abord des tests sur l'animal, puis des essais cliniques chez l'homme en échantillons sélectionné (essai explicatif), puis des essais chez l'homme en échantillon représentatif (essai pragmatique).

Néanmoins la recherche médicale s'est, presque toujours, limitée aux essais explicatifs. Le développement des essais pragmatiques a été lent, pour plusieurs raisons. D'abord parce que la méthode est nouvelle, donc mal connue. Ensuite parce qu'elle donne l'impression d'un manque de rigueur : absence de test statistique, recherche d'une situation aussi voisine que possible de la pratique courante, souplesse dans la définition des traitements. Bref on se rapproche de la médecine quotidienne, ce qui fait craindre un retour à l'empirisme, alors qu'il s'agit d'un essai de rationalisation de la conduite médicale.

En revanche, ici comme dans d'autres domaines, un «effet SIDA» s'est révélé bénéfique : l'essai pragmatique est apparu soudain comme une panacée. La plupart des molécules étant toxiques, on a d'abord demandé que leur évaluation tienne compte d'un bilan coût-avantage. Ensuite il a paru intolérable que des essais, souvent très longs et portant sur des milliers de malades, risquent d'aboutir à la non conclusion, à la suite d'un test statistique. Enfin c'est surtout la définition des traitements qui pose problème. En raison de leur toxicité, un grand pourcentage de malades ne les supportent pas et doivent les abandonner. En général, dans la comparaison de deux traitements *A* et *B*, ceux qui ne supportent pas *A* l'abandonnent au profit de *B* et réciproquement. La comparaison des deux groupes en ressort très affaiblie et, dans la plupart des cas, la différence est non significative.

C'est ici que l'attitude pragmatique est avantageuse : ce qui arrivera, en dehors de tout essai, dans les conditions de la pratique courante, c'est que les sujets intolérants à *A* recevront *B*, et réciproquement. On ne comparera donc pas les traitements *A* et *B*, mais les stratégies «traitement *A*, puis traitement *B* si intolérance au traitement *A*» et «traitement *B*, puis traitement *A* si intolérance au traitement *B*» ; autrement dit, on cherchera s'il faut d'abord administrer *A*, puis *B*, ou *B*, puis *A*. Remis en selle par le SIDA, les essais pragmatiques sont revenus à l'ordre du jour, et nul doute qu'ils ne se répandront.

Il serait utile que la distinction soit perçue dans de nombreux domaines de recherche. En agronomie, par exemple, un chercheur voulut un jour comparer deux races *A* et *B* de haricots. Il constitua des parcelles de haricots *A* et des parcelles de haricots *B*. Les premiers poussaient tout en hauteur, les seconds formaient des boules s'étendant en largeur et se gênant dans leur croissance. L'agronome avait utilisé le même nombre de graines par mètre carré. Pourquoi? Parce que, pour comparer l'effet d'un facteur, il devait, pensait-il, constituer deux groupes comparables pour tous les autres facteurs. Que vient faire ici ce souci d'égalisation? Le problème était de nature pragmatique. Il n'était pas de tester l'hypothèse, tout à fait invraisemblable, d'égalité des deux races, mais de choisir la meilleure. Il fallait donc donner à chacune ses chances optimales.

Les modèles

Qu'est-ce qu'un modèle? Cette question, éludée au début de l'article, avait été reportée avec l'idée que la réponse se concevrait mieux à la lumière de quelques exemples. Nous avions annoncé qu'il y avait, dans l'utilisation du mot, différences et points communs. Les différences sont évidentes : entre le modèle de conception, l'insémination artificielle et la théorie des signatures, elle saute aux yeux.

Cette différence existe d'ailleurs dans l'utilisation du mot en langage courant : on parle des petites filles modèles, du fusil modèle 1936, d'Harpagon comme modèle de l'avare, de la personne qui pose pour les artistes. Dans le domaine scientifique, on justifie cette diversité en instituant des catégories : il y a les modèles dialectiques, parmi lesquels figurent les modèles mathématiques, et les modèles physiques (comme le pendule, modèle du mouvement périodique) ou biologiques (comme la souris, modèle de l'homme pour l'étude d'un médicament).

D'autres parlent de modèles abstraits et concrets, mais la question reste le pourquoi de l'utilisation d'un même mot. Une définition courante du modèle est «instrument d'étude de la réalité», mais cette définition manque de spécificité : ne s'applique-t-elle pas à un microscope, voire à un crayon? Suzanne Bachelard dit plutôt «instrument d'intelligibilité». Il faudrait ajouter «ou d'aide à la décision». Même imparfaites, ces définitions mettent l'accent sur ce but commun des divers modèles, concrets comme abstraits. Ce but ressort clairement des exemples que nous avons présentés.

Nous avons dû faire un choix : ainsi nous avons ignoré les modèles à compartiments, qui décrivent la circulation d'éléments entre des départements qui communiquent, par exemple les échanges d'hormones entre divers organes. De tels modèles font l'objet de nombreuses études, mais ils n'auraient rien ajouté à notre propos. Nous avons préféré retenir quelques modèles très différents et que nous avons nous-mêmes utilisés.

Certains pourront paraître s'écarter de la notion stricte de modèle. Le modèle probabiliste, par exemple, est-il vraiment un modèle? Il en crée en tous cas plusieurs. Ainsi l'essai thérapeutique avec le tirage au sort des patients, son protocole précis, un médicament administré à l'aveugle, n'est-il pas typiquement un modèle construit pour prouver la causalité? Où s'arrête d'ailleurs la frontière des modèles ? Selon Boltzmann, l'hypothèse, la théorie, la science elle-même sont des modèles de la réalité.

Daniel SCHWARTZ est professeur émérite à la Faculté de médecine de Paris Sud.

Henri LERIDON, *Les enfants du désir*, éditions Julliard, 1995.

H. Le BRAS, *Lois de mortalité et âge limite*, in *Populations*, n° 3, 1976.

G. DAVID, *L'insémination artificielle et le système CECOS*, in *L'insémination artificielle*, éditions Masson, 1991.

Ph. LAZAR, *La naissance prématurée*, in *L'anthropologie*, tome 90, n° 3, 1986.

G. BOUVENOT, E. ESCHWÈGE et D. SCHWARTZ, *Le médicament*, éditions de l'INSERM et Nathan, 1993.

J. M. LEGAY, *La méthode des modèles, état actuel de la méthode expérimentale*, éditions Informatique et biosphère.

D. SCHWARTZ, *Le jeu de la science et du hasard*, éditions Flammarion, 1994.

D. SCHWARTZ et J. LELLOUCH, *Explanatory and Pragmatic Attitudes in Clinical Trials*, in *J. of Chronic Diseases*, vol. 20, 1967.