

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

ROBERT GIBRAT

**L'analyse des données. 3e partie. Causalité et analyse
des données en médecine**

Journal de la société statistique de Paris, tome 120, n° 4 (1979), p. 224-243

http://www.numdam.org/item?id=JSFS_1979__120_4_224_0

© Société de statistique de Paris, 1979, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

II

L'ANALYSE DES DONNÉES

3^e partie

CAUSALITÉ ET ANALYSE DES DONNÉES EN MÉDECINE (1)

Robert GIBRAT

Ancien président des Sociétés de statistique de Paris et de France

Nous avons, dans une première partie (Journal de la Société n° 3, 1978), étudié l'application aux techniques industrielles, puis dans une deuxième partie (Journal de la Société n° 4, 1978) commencé l'étude de l'application aux sciences humaines en esquisant une théorie des « oracles » et de la méthode Delphi.

Ici nous prenons acte à la fois de l'échec presque total de la recherche médicale actuelle dans la détermination des causes des maladies et de son succès prodigieux en matière d'accumulation de faits et surtout de corrélations.

Nous pensons que la recherche médicale doit s'orienter de plus en plus vers l'analyse des données, orientant le chercheur vers des évaluations de plus en plus scientifiques de l'effet des thérapeutiques et transformant peu à peu totalement à la fois nosologie et diagnostics.

Une analyse assez poussée de quelques études typiques, surtout par analyse factorielle des correspondances, déjà réalisées par association entre médecins et statisticiens est présentée ici.

In a first part (Journal de la Société n° 3, 1978), we studied the application to industrial techniques, then in a second part (Journal de la Société n° 4, 1978) we began to study the application to human sciences drafting a theory of "oracles" and of the Delphi method.

We state here both the almost complete failure of the present medical research to determine illness causes and its tremendous success in gathering facts and above all, correlations.

We think that the medical research should more and more deal with the data analysis, orientate the searchers towards more and more scientific evaluations of the effect of therapeutics and gradually transform nosology and diagnosis.

We give here a rather deep analysis of some typical studies, above all through the factorial analysis of correspondences, already realized by associated doctors and statisticians.

1. Conférence prononcée aux Secondes journées internationales : « Analyse des données et Informatique » organisées par l'I. R. I. A. du 17 au 19 octobre 1979.

Wir haben in einem ersten Artikel (Journal de la Société No. 3 1978) die Anwendung an den industriellen Techniken studiert, dann in einem zweiten (Journal de la Société No. 4 1978) eine Studie der Anwendung an die Sozialwissenschaften begonnen, indem wir eine Theorie der „Orakel“ und der Methode „Delphi“ gezeichnet haben. Hier stellen wir endgültig die Tatsache fest von dem fast vollkommenen Misserfolg der medizinischen Versuche die Ursachen der Krankheiten zu bestimmen und im Gegensatz dazu von dem fast unglaublichen Erfolg hinsichtlich der Sammlung von Tatsachen und besonders von Korrelationen.

Nach unserer Meinung muss sich die medizinische Forschung immer mehr auf eine Analyse der Tatsachen richten. Der medizinische Forscher muss sich immer mehr auf eine wissenschaftliche Abschätzung der Wirkung der Medikamente beschränken und nach und nach zu gleicher Zeit die Nosologie und die Diagnosenstellung umwandeln.

Eine sehr eingehende Analyse von einigen „typischen“ Arbeiten ist gegeben mit der faktoriellen Analyse der Beziehungen, die das Resultat einer Zusammenarbeit von Medizinern und Statistikern ist.

... Si certaines disciplines, comme les Sciences humaines ou la Biologie, ont été si lentes à se mathématiser, ce n'est pas tant, comme on le croit souvent, à cause de la complexité de leur objet (tout est complexe dans la nature), mais plutôt parce que la déduction qualitative et empirique y offrait, d'ores et déjà, un support suffisant aux possibilités de prévision et d'expérience.

René THOM, *Stabilité structurelle et morphogénèse*,
page 2, (Édition 1977).

I — INTRODUCTION

La recherche médicale est contestée aujourd'hui en France et à l'échelle mondiale : les pouvoirs publics s'interrogent discrètement devant l'importance des dépenses et le ralentissement des découvertes. Il n'est pas question de nier les fantastiques progrès obtenus de 1930 à 1960, mais de réfléchir sur ce qui se passe depuis, particulièrement pour les trois grandes maladies dominantes dans nos pays, cardiovasculaires, cancer et diabète. Cette question a été étudiée rarement, « sine ira et studio ». Nous allons, cependant, essayer d'apporter notre contribution en montrant *le rôle essentiel que devrait prendre l'analyse des données en informatique médicale.*

Trois directions doivent être examinées :

1. Fin 1960, les *maladies infectieuses*, tout au moins en Occident, étaient en voie de disparition rapide. Il y avait relation claire de cause à effet à l'échelle individuelle dans la méningite induite par le bacille de la tuberculose. Le succès était obtenu à 100 % au point que l'on prévoyait à l'époque où toutes les bactéries dangereuses auraient disparu de la surface du monde. Mêmes relations rigoureuses apparentes entre les insectes et le D. D. T.

Une résistance microbienne inattendue a balayé ces espoirs, celles des insectes a été stupéfiante. Comme l'a écrit le professeur J. Hamburger dans « l'homme et les hommes » (1976), « une espèce vivante, mise en présence de conditions extérieures apparemment mortelles, est capable de survivre et de « s'adapter » à des conditions adverses ». On connaît, depuis les travaux des deux prix Nobel Luria et Lenerberg, les causes exactes de ces réactions inattendues. « Les individus microbiens ont tous l'air identique, continue Jean Hamburger, mais ils ne le sont pas. Sur un millier de bacilles pyocéniques par exemple... quelques microbes

sur ces réactions
Jean Hamburger
ECN 01/11/1977

se distinguent des autres, précisément à l'un des points vulnérables... Ce point vulnérable peut différer d'un antibiotique à l'autre, mais il y aura toujours dans la masse des germes sensibles, quelques individus-microbes n'ayant pas le point de faiblesse et par conséquent non vulnérables ». Pour les insecticides, en principe mortels, on estime à 5 pour 1 000 la proportion des insectes spontanément résistants.

2. Autre domaine où les déceptions se sont succédées, la *nosologie* ou classification des maladies : Au Colloque U. N. E. S. C. O., sur « la nouvelle frontière technologique » (24-26 avril 1979), le professeur agrégé de médecine, J. F. Bach, traitant des progrès récents de l'immunologie, déclarait avec force (nous le citons de mémoire) : aujourd'hui où les maladies contagieuses ont proprement disparu, la maladie (diabète, bronchite chronique, maladies cardiovasculaires, etc.) est définie par un ensemble de signes cliniques correspondant à une thérapeutique bien définie, mais permettant des confusions d'une maladie à l'autre, ce qui est peu satisfaisant et détruit toute notion stricte de causalité au point qu'on a pu parler d'une appréciation « *botanique* » de la maladie. Botanique signifie *taxinomie*, ajoutons nous, et chacun sait que le tome I du célèbre ouvrage de Benzecri sur « l'analyse des données » porte ce titre. Le lien causalité en médecine et analyse des données étant ainsi établi, ajoutons que J.-F. Bach pense que l'immunologie devrait permettre d'utiliser à nouveau des relations de cause à effet, quand elle expliquera pourquoi et comment une cellule humaine est capable de différencier et donc de reconnaître 10^7 structures différentes. Le professeur Jean Bernard écrit de son côté : « les différentes classes du système de globules blancs connus sous le nom de H. L. A. représentent autant de prédispositions à diverses maladies du système nerveux, du tube digestif, de la peau qui peu à peu se définissent ».

En attendant il faut essayer de guérir le plus possible. Le professeur Tubiana, dans son livre « Le refus du réel » (1977) page 118, explique comment : « la méthode expérimentale permet des progrès empiriques, même si les mécanismes ne sont pas connus. Si l'on avait attendu pour soigner les maladies de connaître leur origine et leurs mécanismes, si l'on attendait pour utiliser un traitement d'avoir étudié son mode d'action, la médecine serait encore balbutiante. Or, dès le début du xx^e siècle, on guérissait des cancers avec le radium et les rayons X au moment où l'on n'avait que peu d'idées sur ce qui était un cancer et encore moins sur les effets biologiques des rayonnements. Plus récemment, si l'on avait attendu pour faire des greffes de rein de connaître les lois de l'immunologie des centaines d'hommes seraient morts qui sont vivants ».

Mais évidemment, la recherche sans idée directrice coûte très cher et amasse des faits sans relations logiques, les thérapeutiques sur des maladies mal définies auront des effets statistiques, donc conduiront à des dépenses individuelles aujourd'hui presque inacceptables. On ne saurait continuer sur cette lancée.

3. Les *facteurs de risques*. Le lecteur s'attend peut être à nous voir traiter maintenant sur le plan philosophique du principe de causalité : « tout phénomène a une cause ». Aristote, Descartes, Kant, Auguste Comte, Schopenhauer et mille autres philosophes en ont traité savamment. Mais qui a écrit : « la notion de cause dans les sciences est d'un usage d'autant plus rare qu'elles sont plus développées ». Des allusions aux théories quantiques paraîtraient aussi nécessaires. (Nous recommandons alors le livre de Max Jammer, « The philosophy of the quantum mechanics »), etc.

Volontairement nous y renonçons, laissant la parole seulement aux statisticiens et aux médecins. Daniel Schwartz, directeur de l'unité de recherches statistiques de l'I. N. S. E. R. M., ancien président de la Société de Statistique de Paris, dans sa trop brève

« Contribution de la statistique à des conceptions nouvelles de la médecine » (*Journal de la Société de Statistique de Paris*, n° 3, 3^e trimestre 1978, pages 192-194) estime que c'est « en très grande partie à l'apport de la statistique que la médecine doit, d'une part, son accès au rang de science, d'autre part son orientation actuelle ». Il regrette qu'en matière de comparaison des thérapeutiques le corps de doctrine statistique très satisfaisant aujourd'hui n'a pratiquement pas trouvé d'écho chez les médecins, « trop coûteux » en efforts plus qu'en argent », mais insiste sur ses succès marqués en *épidémiologie*, c'est-à-dire dans la « recherche des facteurs influençant la fréquence d'une maladie ».

L'étude épidémiologique la plus célèbre, explique-t-il, est celle de la corrélation entre tabac et cancer du poumon, bien attestée par des enquêtes considérables, faites dans des conditions rigoureuses (l'une d'elles, aux U. S. A., a suivi un million de personnes pendant 5 ans). Mais comment passer de la corrélation à la cause? Comme l'écrit Daniel Schwartz, « tout permet de supposer que fumeurs et non fumeurs diffèrent, non seulement par l'usage du tabac, mais par toute une série de caractéristiques dont l'une pourrait bien être — pourquoi pas? — la cause du cancer ». L'utilisation d'analyses des statistiques les plus raffinées dans des travaux immenses, ne pouvait évidemment aboutir à la preuve d'une relation causale, ce que pourtant l'immense majorité des médecins et du public semble admettre. Il existe à la fois de nombreux fumeurs qui mourront avant d'avoir un cancer du poumon bien que celui-ci sera noir de goudron à l'autopsie et de nombreux cancéreux du poumon qui n'ont jamais fumé. Cependant, chacun sent très bien qu'il est mieux de s'abstenir de fumer, mais c'est de la nature d'une assurance accidents à laquelle personne n'attribuerait de relation causale. « L'épidémiologie *explicative*, conclut-il, est devenue épidémiologie *pragmatique* ».

La situation n'est pas différente en matière d'infarctus. On a décelé, dès les premières enquêtes, divers « facteurs de risques » : le cholestérol, la tension artérielle, certains électrocardiogrammes, etc.; des expressions mathématiques très sophistiquées ont parfaitement défini le « candidat » sérieux à l'infarctus par combinaisons des facteurs de risques, mais sans preuve d'une relation causale. Comme l'a écrit Daniel Schwartz « cette performance très belle, quasi magique, ... n'apportait rien d'autre que la connaissance d'un danger ». Dans ce cas, comme dans presque tous les autres, on a cherché la diminution du risque non dans une action sur les facteurs de risques, mais par des procédés indépendants. Ainsi, on dirige les femmes dont la grossesse présente un haut facteur de risques sur lequel on ne peut rien, l'âge par exemple, vers des maternités spécialement équipées.

Dans un livre très récent (1978), « L'espérance ou le nouvel état de la médecine », le professeur J. Bernard écrit, page 41, à propos des facteurs de risques dans les maladies cardiovasculaires : « et d'abord quel est le sens exact du terme « facteur de risques ». A quel moment, à quel niveau interviennent ces facteurs? Avec l'ensemble des phénomènes qui vont préparer l'apparition de lésions artérielles? Dans la genèse de la toute première lésion artérielle? En aggravant les premières lésions artérielles et favorisant leurs extensions? En créant les conditions propices à l'éclosion des accidents eux-mêmes, de l'hémorragie cérébrale, de l'infarctus du myocarde, de la gangrène des membres? Peut-être certains facteurs interviennent-ils à plusieurs niveaux? Peut-être d'autres facteurs exercent-ils une action spécifique à un moment précis? ... Notre ignorance des causes profondes de l'athérosclérose est grave. Il est grave de ne pas connaître la cause de la maladie la plus fréquente de l'homme de la fin du xx^e siècle ».

On voit le chemin parcouru en lisant ce qu'il écrivait en 1973, dans « Grandeur et tentative de la médecine » : « La médecine si longtemps empirique et balbutiante devint une science fondée sur des règles précises ».

Certains professeurs donnent à l'ignorance une valeur particulière : ainsi, le professeur Schwarzenberg écrit dans « Changer le monde » (1977), page 84 : « Il existe dans le monde des cas de régression spontanée d'un cancer apparemment sans espoir. Actuellement, 176 cas bien étudiés avec documents à l'appui peuvent être considérés comme des guérisons spontanées. Les tumeurs ont cessé de grossir, ont régressé et ont disparu. Qu'est il arrivé? Personne ne le sait réellement, on est réduit aux hypothèses ».

Citons encore une fois le professeur J. Bernard pour préciser cette introduction sur la nature de la causalité en médecine (« L'homme changé par l'homme ») : « la recherche des causes est une des plus nobles fonctions de la médecine. Elle est rarement exhaustive. Elle est avant tout pragmatique : elle a pour objet d'améliorer les conditions de la prévention et du traitement. Cette recherche se développe habituellement sur trois plans :

- les causes externes sont parfois évidentes, l'accident d'automobile, l'arsenic absorbé, la contagion en période épidémique,
- les causes internes, tenant à l'individu, à l'altération de sa constitution, sont aisément reconnues dans les grandes maladies génétiques, maladie de l'hémoglobine, idiotie mongolienne, etc.,
- l'analyse du mécanisme même du trouble, de sa physiopathologie, pour employer le langage des médecins, rend de grands services dans les cas où les causes externes ou internes sont ignorées. Ainsi de nombreuses opérations biochimiques concourent à la fabrication du globule rouge normal. Pour l'anémie pernicieuse, nous ne connaissons pas les causes de la maladie, mais nous savons à quel point la chaîne de fabrication est altérée. Nous savons que la vitamine B 12 corrige cette altération. L'anémie n'est plus pernicieuse ».

Inégale « sensibilité » des êtres vivants mise en évidence dans la réaction aux antibiotiques ou aux insecticides, donc a priori à toute thérapeutique, classification « botanique » des maladies peu satisfaisante ou peu fiable, facteurs de risques sans relations causales traduisant des « prédispositions » à diverses maladies, mots-clefs qui rappelleraient trop la vertu « dormitive » de l'opium, si le médecin n'acceptait pas, un peu à son corps défendant, de devenir probabiliste et si les théories statistiques n'avaient fait de très importants progrès précisément ces dernières années.

On devine donc le rôle essentiel que peut avoir l'analyse des données pour relancer la recherche médicale, car elle s'est heurtée aux mêmes problèmes de causalité. La deuxième partie de cet exposé va analyser quelques études typiques de premier plan, déjà réalisées en collaboration entre médecins et statisticiens. Une troisième partie fera une évaluation des moyens d'action après un retour vers la causalité à travers la théorie des catastrophes de Thom. Des conclusions brèves termineront cet exposé.

II — LES DIX PREMIÈRES ANNÉES DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES (A. F. C.) EN MÉDECINE

D'abord un bref historique : l'analyse *factorielle* a été développée dès le début du siècle par les psychologues (Spearman, 1904) pour « expliquer » des résultats par des facteurs cachés (mémoire, intelligence, etc.). Elle supposait un modèle a priori et ne portait, faute d'ordinateurs, que sur des petits ensembles de données. Puis l'analyse dite en *composantes principales*

a cherché sans hypothèses particulières a priori à représenter le mieux possible des ensembles de données de plus en plus grands dans des espaces de petites dimensions. Toute une série de méthodes en ont dérivé : analyse de covariances partielles, analyse de rangs, analyse canonique, analyse discriminante ainsi qu'une floraison extraordinaire de méthodes de classification automatique. Mais la méthode la plus puissante, la mieux structurée, la plus indépendante des modèles a priori, celle de l'analyse *factorielle des correspondances* (A. F. C.) date seulement de 1964 et est animée par une équipe française. Elle a évidemment été favorisée par l'essor concomitant des grands ordinateurs. Les premières leçons ont été données à cette époque à l'Université de Rennes par son animateur le professeur J.-P. Benzecri, une thèse de mai 1965 de son élève, M^{me} Escoffier-Cordier apportant presque immédiatement des compléments théoriques importants. Une nouvelle approche des problèmes statistiques en a résulté sans conteste...

Son succès près des jeunes chercheurs, les critiques aiguës et les résistances dont elle fait toujours l'objet et surtout son adaptation remarquable aux problèmes où la causalité individuelle disparaît expliquent l'attention presque exclusive que nous lui portons ici. (On pourra consulter, sur le plan général, nos deux études parues dans le *Journal de Société de Statistique de Paris*, 3^e trimestre 1978, pages 201 et 228 et 4^e trimestre 1978, pages 312-331.)

L'A. F. C. prend peu à peu une certaine place dans la recherche médicale et nous avons eu la joie de présider, les 8 et 9 mars 1979, à la faculté de médecine de Dijon, des journées sur ce sujet où assistaient, en particulier, onze des quatorze professeurs français de médecine agrégés d'informatique, et où ont été passées en revue les principales applications. (Malheureusement rien n'a été publié.)

Nous n'avons été, ni en médecine, ni en statistique, élève ou professeur. Nos réflexions sont celles d'un ingénieur de bonne volonté; elles ne visent ni à être complètes, ni à être définitives. Nous avons pu recueillir des documents dans trois domaines de la médecine (affections hépatiques, infarctus graves, leucémies); ils donnent croyons-nous, une vue assez représentative de l'effort médical encore assez modeste dans le domaine de l'A. F. C.

1. *Recherches hépatologiques* (Rennes)

Il s'agit des travaux de l'unité 49 de l'I. N. S. E. R. M., à l'hôpital du Pont-Chaillou, à Rennes (professeurs M. Bourel et P. Lenoir); unité particulièrement active en matière d'A. F. C. M. Kerbaol, leur principal statisticien, nous a adressé un dossier d'une dizaine d'études. La première, de 1967, signée du professeur Benzecri ⁽¹⁾, est par sa date et son contenu un document historique car elle est une des toutes premières études A. F. C. Elle part de 613 dossiers de patients, rassemblant 12 catégories d'affections hépatiques, chacun comprennent six mesures : fer sérique, cuivre sérique, etc. Les analyses avaient trois buts :

- vérifier, sur les données médicales, la validité de la méthode utilisée;
- essayer de distinguer, à l'aide des mesures, deux des douze catégories, syndromes inflammatoires et cancers viscéraux;
- faire des remarques d'ordre médical.

L'A. F. C. a utilisé uniquement ici la projection sur le plan des deux premiers facteurs.

L'apparition d'un groupement « franc » de maladies dans le coin inférieur droit (les hémachromatoses primitives) avait paru une preuve de la validité de la méthode, même alors qu'il était le seul ainsi mis en évidence. On a aussi vu que trois des 16 mesures auraient pu être éliminées ayant peu d'effet discriminant. Enfin, un fait médical peu connu (une seule

1. P. Benzecri, P. Lenoir, M. Kerbaol et M. Bourel. Premier congrès d'électronique médicale et de génie biologique. Tours, 1^{er} juillet 1967.

référence médicale à l'époque) est apparu avec la forte liaison des valeurs d'alpha 2 globuline et du cuivre.

Dans cette étude et celles qui l'ont complétée (deux analyses de 180 et 135 patients) on voit apparaître les principales difficultés de la technique A. F. C. que le professeur Benzecri et ses élèves ont pu résoudre ensuite peu à peu, en particulier dans les deux célèbres volumes, parus chez Dunod en 1973, (1. Taxinomie, 2. Correspondances) qui constituent aujourd'hui la référence fondamentale. Citons en exemple la valeur explicative, plus ou moins assurée, des proportions d'inertie totale représentée dans le plan choisi, la prise en compte de l'hétérogénéité des données tantôt quantitatives, tantôt qualitatives (« la subjectivité de la limite de signification contraste scandaleusement avec la rigueur de la puissance de l'analyse des correspondances »).

La première analyse a permis de retrouver, dans le plan factoriel, des relations déjà classiques en médecine. « Le groupement extrêmement serré de la plupart des hémochromatoses primitives comparé à celui beaucoup plus lâche des cirrhoses et des hémochromatoses secondaires valide, à nos yeux, l'autonomie pathogénique des hémochromatoses, fait toujours discuté... ».

La deuxième a permis de répondre à la question de savoir si on pouvait discriminer entre syndromes inflammatoires et cancers viscéraux.

La conclusion de cette série « historique » mérite d'être reproduite :

« Cette méthode présente, pour nous cliniciens, un intérêt majeur dans, au moins trois domaines privilégiés :

1° Celui de la nosologie pathologique ⁽¹⁾. Notre découpage de la pathologie est-il licite? Sinon est-il réalisable et comment? ...

2° Celui de l'optimisation de la sémiologie médicale ⁽²⁾.

3° Celui du diagnostic chimiobiologique des situations pathologiques complexes ⁽³⁾.

On retrouve les mêmes auteurs ⁽⁴⁾ dans une autre étude : celle semi-quantitative des associations naturelles de cinq types de cellules dans 439 primo cultures de foie humain adulte. L'A. F. C. a montré que « leurs formules sont liées à la pathologie du foie explanté, mais ne permettent pas une discrimination non équivoque des situations pathologiques ».

Nous signalerons encore une étude récente ⁽⁵⁾ A. F. C. sur l'exploration fonctionnelle de la thyroïde, en commun avec les services de l'hôpital Saint André à Bordeaux comportant 201 sujets et 16 caractères. Le but était surtout de confronter deux types d'associations d'examen (tests *in vivo*, tests *in vitro*). Les résultats ne paraissent pas définitifs, l'intérêt principal résidant « dans la réflexion, la critique, la mise en œuvre de nouvelles investigations que suggère la projection en un lieu inattendu d'un sujet ou d'un groupe de sujets liés par des critères particuliers ». Une analyse partielle a permis, enfin, de représenter sur le plan factoriel par des segments de droite l'évolution de la maladie et ainsi de mieux ajuster le traitement itératif à l'iode 131 en représentant les diverses étapes biologiques franchies.

1. Nosologie : classement général des maladies d'après un certain nombre de critères qui permettent de définir et s'isoler chacune d'elles.

2. Sémiologie : partie de la médecine qui s'occupe des signes ou des symptômes des maladies

3. Il s'agit ici de l'introduction d'un dossier nouveau en variable supplémentaire, point fondamental utilisé assez tardivement en A. F. C. Une première conclusion était plus prudente : « ... les problèmes restent encore nombreux et seule une étroite collaboration entre mathématiciens et médecins doit nous permettre de les dépister, puis de les résoudre ».

4. Relations entre les types cellulaires observés dans les primo cultures de foie humain adulte et la pathologie des malades dont elles proviennent. M. Kerbaol et alia dans « Biologie et gastroentérologie », 1971, tome II, pages 113 à 122.

5. Application de l'analyse factorielle des correspondances à l'exploration fonctionnelle de la thyroïde. M. Kerbaol et alia. *Pathologie biologique*. 1971, 23, n° 1, pages 21 à 28.

Nous citerons, enfin, pour clore cette rapide revue des études de Rennes un compte rendu à l'Académie des Sciences, en décembre 1970 (1). Il est reproduit *in extenso* par le professeur Benzecri (tome I A, n° 2, § 4, pages 37 à 42). 503 sujets avec 15 catégories de situation pathologique y compris la normale, 23 données biologiques bien étudiées et homogènes. Quarante analyses ont été faites (il s'agit donc d'un travail considérable). Deux seulement ont été retenues. La première donne dans le plan des deux premiers axes, avec 80 % d'inertie, cinq zones nosologiques bien discriminantes. La seconde utilise les trois premiers axes (87 % d'inertie) fournissant respectivement quatre zones (premier et deuxième axe) et trois zones (premier et troisième axe) donnant ainsi une séparation encore plus satisfaisante.

2. Pronostic de survie après infarctus grave (Lariboisière)

Il s'agit d'une étude parue dans les Cahiers de l'Analyse des Données (2). Le résumé offert indique que « l'étude vaut tant par la généralité des méthodes que par l'importance des résultats concrets », ce qui est tout à fait exact, la signature du professeur Benzecri explique notre assez longue analyse.

Il s'agit d'une technique d'assistance dans le cas d'atteinte grave au muscle cardiaque par introduction, dans l'aorte, d'un ballon (24 cm de long, diamètre inférieur à celui de l'aorte), une pompe assurant gonflage et dégonflage synchronisée sur le rythme cardiaque.

Cette technique, appliquée à l'Unité de soins intensifs coronaires de l'hôpital Lariboisière, est très lourde en personnel et en matériel. On ne veut donc l'appliquer qu'aux patients n'ayant pas de chance de survie sans une aide mécanique et on cherche à déterminer celle-ci par l'analyse statistique utilisée ici comme aide au pronostic.

Toutes les ressources de l'A. F. C. (analyse factorielle des correspondances) ont été utilisées pour assurer un choix impeccable des données. Elles ont été tirées des examens initiaux effectués à l'hospitalisation de 101 patients avec 17 variables mesurées (une fréquence, cinq pressions hémodynamiques, deux résistances, deux mesures de travail et deux de puissance, cinq index), deux variables quantitatives (âge, date de la première observation par rapport au début de l'infarctus), et quatre questions à deux modalités (sexe, infarctus dans les antécédents, troubles de la conduction, localisation antérieure ou postérieure. De plus, on a introduit en variable supplémentaire la variable à expliquer : survie au delà de quatre semaines (S) ou décès (D).

Les variables quantitatives ont été divisées en classes, trois, quatre ou cinq d'importance numérique à peu près égale, d'où 81 modalités au total et un tableau de base 101×81 .

Les contributions relatives ont été ici calculées, ce qui est à recommander d'une manière générale, leur pouvoir explicatif étant souvent considérable. Les plus importantes

1. Une étude en ordinateur des corrélations entre les modifications des protéines sériques en pathologie humaine. Note de G. Sandor, P. Lenoir et M. Kerbaol, présentée par J. Tréfouel. (Bibliographie de dix articles échelonnés entre 1963 et 1970). Des études préliminaires qui nous ont été communiquées portaient sur 201 cas et 16 données biologiques.

Le professeur Sandor a, depuis, en collaboration avec l'I. R. I. A. (E. Diday et Y. Lechevallier) approfondi ces corrélations en utilisant d'autres méthodes d'analyse des données, classification automatique par nuées dynamiques, utilisation de l'algorithme de Fisher sur un tableau de contingence, etc. Il obtient ainsi une grille de diagnostics (voir en particulier colloque I. R. I. A., Analyse des données, page 165. Versailles 7-9 septembre 1977). Nous tenions beaucoup à citer ces travaux afin que le lecteur ne croit pas que le choix fait par nous de l'A. F. C. est un choix exclusif.

2. Aspects pronostiques et thérapeutiques de l'infarctus myocardique aigu compliqué d'une défaillance sévère de la pompe cardiaque. Application des méthodes de discrimination (aorte). Volume II, 1977, n° 4, pages 415 à 434, par J. P. Nakache, P. Lorente, J. P. Benzecri, J. F. Chastang. Il faut signaler sur le même sujet une communication de P. Lorente et J. P. Nakache aux Premières journées internationales, « Analyse des données et informatique », I. R. I. A., Versailles 7-9 septembre 1977, pages 243 et 254, avec le titre : « Évolution de l'aide au pronostic et à la décision thérapeutique des méthodes statistiques multidimensionnelles au cours de l'infarctus myocardique aigu ».

pour le premier axe sont celles des paramètres de la fonction ventriculaire gauche et pour le deuxième axe celles de la fonction ventriculaire droite, discrimination que les cardiologues ne sauraient démentir (1).

Finalement, l'interprétation habituelle sur le plan factoriel des deux premiers axes montre que survivants et décédés occupent des zones en recouvrement partiel, d'où un

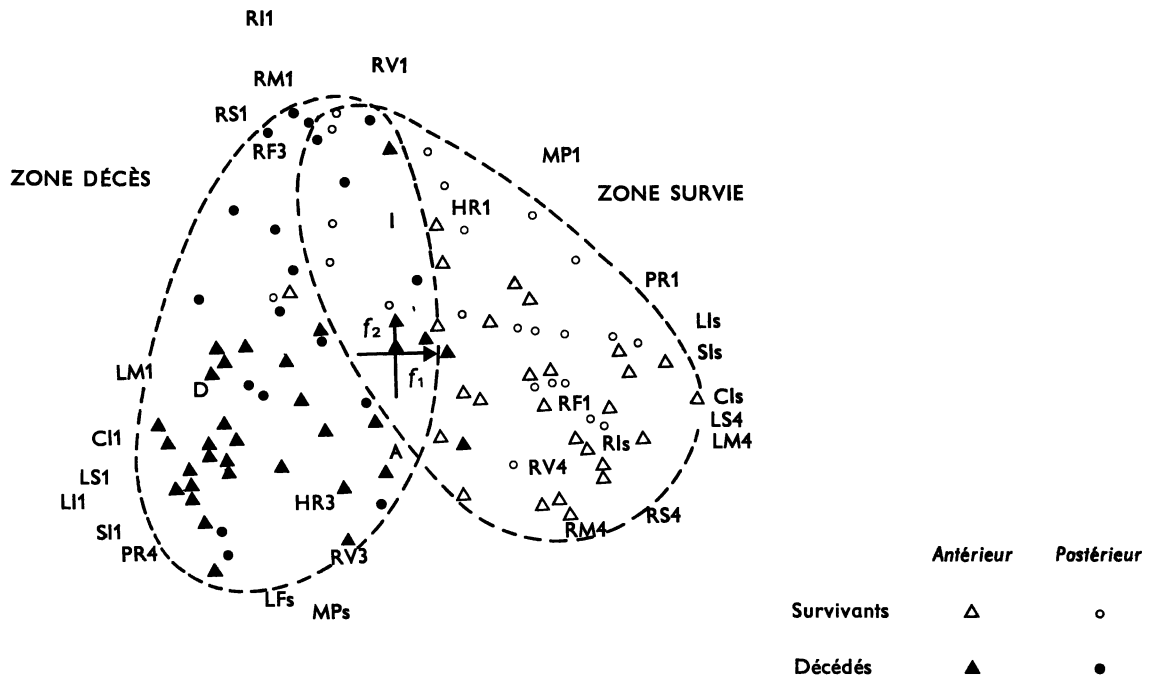


FIG. 1. — Projection sur (f_1 , f_2) des patients et des paramètres (parmi les 83 bien représentés sur ce plan). Cahiers de l'Analyse des données. Vol. II, n° 4.

classement en trois zones de 101 patients (zone survie, 43 survivants et 2 décédés; zone décès, 43 décédés et 2 survivants; zone mixte 5 survivants et 7 décédés). Les survivants sont donc bien classés à 86 %, les décédés à 82 %. Il y a 12 % de mal classés : trois pourcentages qui peuvent apparaître satisfaisants quand on se réfère à notre introduction et à l'état actuel de la recherche médicale (fig. 1).

1. En reprenant les notations de la page 417 des Cahiers et l'utilisation des relations du haut de la page 419, on montre que l'on peut déduire rigoureusement des huit variables CI, HR, MA, MP, LF, LS, RS, RF par les quatre règles de l'arithmétique les huit autres variables R, RI, RM, RV, LM, LI, PR, SI; par exemple, on a $1\ 000\ CI = SI \times HR$ ou $LS = LI \times LF$ ou $RS = RI \times RF$. Une seule variable est réellement indépendante des seize autres mesures hémodynamiques.

Les huit variables en dépendance représentent 36 modalités sur les 83 retenues dans l'impeccable codage. Il paraît, a priori, étonnant qu'elles aient été conservées dans les calculs et on peut se demander dans quelle mesure cette utilisation de variables dépendantes à côté d'un système de variables indépendantes n'a pas faussé certains résultats de l'étude. Nous avons posé la question aux auteurs.

Par exemple, les contributions relatives les plus importantes sur le premier axe factoriel sont les modalités extrêmes de CI (index cardiaque) et SI (index systolique), CI et SI correspondant par ailleurs à des contributions relatives faibles sur le deuxième axe factoriel. Doit-on alors s'étonner que HR (fréquence cardiaque) qui est le quotient des deux variables précédentes ait de faibles contributions au premier axe et de fortes au second, ceci ne faisant que traduire la relation mathématique sans apporter la moindre information supplémentaire au problème? Nous pourrions évidemment multiplier les exemples.

La signification de l'axe 2 (infarctus antérieur et postérieur) et le plan de la zone de recouvrement suggèrent que la séparation entre survivants et décédés est plus franche pour les antérieurs que les postérieurs, ce qui correspond à la gravité différente des obstructions des deux coronaires.

On avait utilisé en parallèle cinq méthodes différentes de classification. Les résultats ont été décevants malgré une étude critique très fouillée (1). Aussi, le processus utilisé pour la décision d'assistance utilise les trois zones précédentes du plan factoriel fourni par les 101 premiers patients en y plaçant successivement comme élément supplémentaire les données des 55 cas nouveaux. 21 se sont trouvés dans la zone de survie, ils ont reçu le traitement

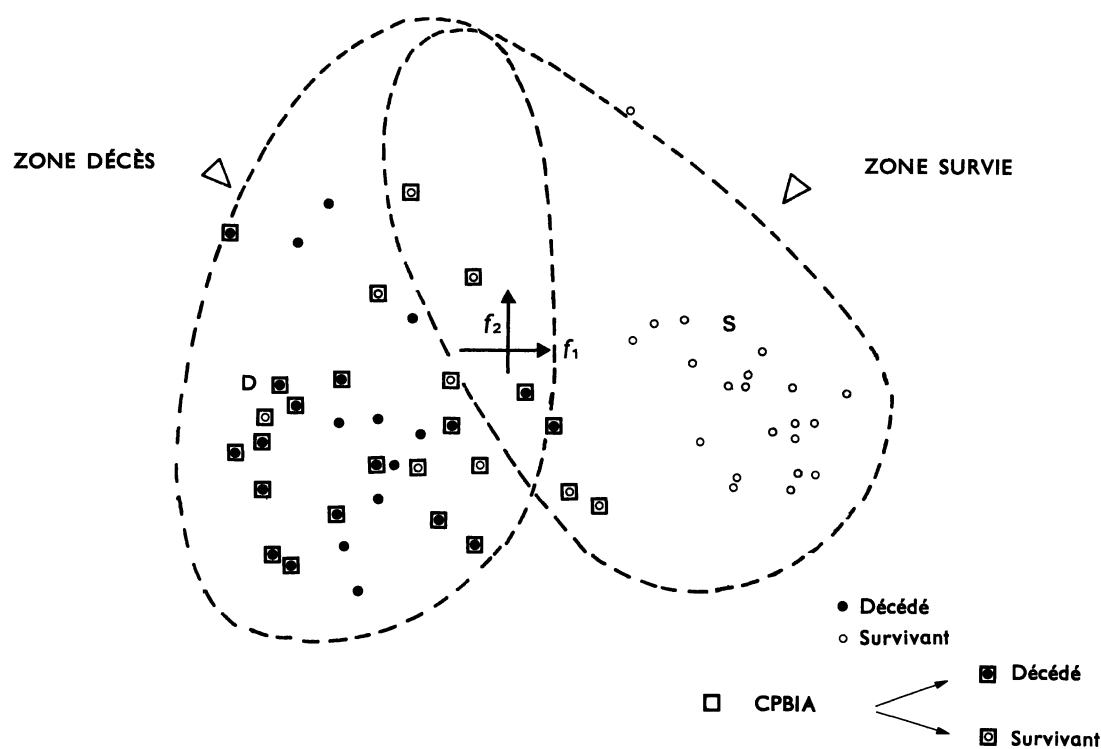


FIG. 2. — Projection des huit variables supplémentaires sur (f_1, f_2) . Cahiers de l'Analyse des données. Vol. II, n° 4.

normal sans assistance, *tous ont survécu*. Les autres se sont placés dans la zone de décès ou au voisinage, 25 ont reçu l'assistance de la contre-pulsion intra-aortique, 9 ont survécu au delà de quatre semaines, les 16 autres sont décédés ainsi que ceux qui n'avaient pas reçu l'assistance (*cf.* fig. 2, C. P. B. I. A.).

On notera que, malheureusement, sur les 9 survivants, 5 sont morts dans un délai de un à plusieurs mois. Cette étude ne permet donc pas, à notre avis, de porter un jugement sur la contre-pulsion elle-même.

1. Le véritable problème est de ne pas savoir *a priori* pourquoi ces méthodes sont décevantes. Il est rare que l'A. F. C. le soit aussi pleinement. Est ce dû au fait que les chercheurs en ont beaucoup plus l'expérience? L'épistémologie de l'analyse des données est tout entière à construire.

Les figures 1 et 2 sont très belles et expliquent la conclusion de l'étude : « Quant aux méthodes multidimensionnelles, elles ont aidé le praticien dans le pronostic et la décision thérapeutique en donnant des multiples informations recueillies ordinairement, une présentation géométrique simple se prêtant bien à la comparaison d'un cas nouveau aux cas antérieurs... » (1).

3. Les leucémies (hôpital Saint-Louis)

Des recherches sont en cours dans de nombreuses directions : nosologie encore très incertaine, mécanismes d'action de la chimiothérapie (dosage, calendrier des traitements), épidémiologie avec la détermination des « facteurs de risques » (virus peut être, hérédité peut être, géographie). Nous examinerons, ici, les analyses statistiques faites sur les données recueillies à l'U. E. R. d'hématologie (service du professeur Jean Bernard) à l'hôpital Saint-Louis, à Paris : leucémies oligo-blastiques (LOB) (statisticiens du service même sous la conduite de V. Izrael), leucémies lymphoïdes chroniques (LLC) (statisticiens de l'Université P. et M. Curie, Paris), leucémies myéloïdes chroniques (LMC) (2). Malgré leur commune origine médicale, ces études sont diverses et méritent une analyse approfondie.

a) Les leucémies oligo blastiques (LOB)

Suivant le professeur Jean Bernard, les LOB se situent entre les leucémies où la moelle est envahie par une prolifération uniforme de cellules souches et les aplasies médullaires où la moelle est complètement désertée : elles comportent un certain degré d'insuffisance médullaire. A ces deux frontières, il en ajoute une troisième, les syndromes myélo-prolifératifs chroniques pour des raisons de « maturation granuleuse ». Quoiqu'il en soit, les frontières étant mobiles, les états correspondants ambigus, les LOB elles-mêmes évoluant à travers des morphologies différentes et étant l'aboutissement d'anémies diverses, on est en présence d'une nosologie très mystérieuse. Ce concept LOB peut correspondre à des états pathologiques différents abusivement regroupés ou, au contraire, à un ensemble homogène. Les conséquences d'une mauvaise réponse peuvent être très graves pour le patient. « Porter un diagnostic erroné d'anémie pernicieuse ou carencielle entraîne la prescription de thérapeutiques aussi variées qu'inutiles ou contraires... Conséquence d'un diagnostic trop hâtif de leucémie aiguë, instituer une chimiothérapie plus agressive pour le malade que pour la maladie. »

1. Les pourcentages d'inertie expliquée par les deux premiers axes factoriels, respectivement 12,9 % et 8,2 %, ont apparu faibles aux auteurs dès le début de leur étude et il leur apparaissait choquant d'obtenir des pronostics aussi bons avec seulement 21,1 %. Ceci est à comparer avec les pourcentages d'inertie des premières études de Rennes (pour les trois premiers axes, 88,3 % dans l'étude sur les foies, 87 % dans celle des protéines sériques). Aussi, ont ils proposé (idée souvent reprise dans les Cahiers de l'Analyse des données) d'utiliser comme « conformes aux faits » les carrés des valeurs propres, ce qui conduirait à 35,6 % d'inertie expliquée. Cela ne nous paraît pas convaincant, surtout en nous référant aux travaux de L. Lebart, « Validité des résultats en analyse des données » (recherche financée par la D. G. R. S. T., novembre 1975, 158 pages).

2. J. BERNARD, V. IZRAEL, C. JACQUILLAT. « Les leucémies oligo blastiques (LOB) ». *La Nouvelle Presse Médicale*. 29 mars 1975, 4, n° 13. Éditorial.

Idem et alias. Les myéloblastielles partielles (ou LOB). L'apport de l'analyse des données à la détermination du pronostic exprimé sous une forme tronquée. *Bulletin du cancer*, 1974, 61, n° 3, pages 341 à 366.

Ch. BASTIN, D. RAIN (Chef de clinique) et alias. « Les leucémies lymphoïdes chroniques ». 7 9 septembre 1977, page 339. La diversité des cas et leur évolution (LLC). *Les Cahiers de l'analyse des données*. Volume I, 1976, n° 4, pages 419 à 440.

Cl. JACQUILLAT, J. BERNARD et alias. « Facteurs de pronostics de la leucémie myéloïde chronique. A propos de 498 observations. *Nouvelle revue française d'hématologie*. 1975, tome 15, n° 2, pages 29 à 240.

Pour le vocabulaire médical, nous conseillons aux non spécialistes le petit abrégé d'hématologie (238 p.) chez Masson (1978) par J. BERNARD et J. P. LÉVY. Ils pourront y vérifier les idées que nous avons présentées dans l'introduction de ce texte.

Parmi les méthodes de recherche, le professeur Jean Bernard a voulu utiliser la statistique, mais avant mars 1975 on examinait seulement que quelques dizaines de cas à la fois. Aussi, son équipe a sélectionné 120 cas parmi les malades des dix dernières années répondant à des limites bien définies situées à distance convenable des frontières dont nous avons parlé plus haut.

La première partie de l'étude examine, comme on le faisait autrefois en statistique classique, une dizaine de distributions à un paramètre en vue « d'esquisser un portrait type » des LOB, âge, sexe, tableau chimique, examen du sang (globules rouges, plaquettes, moelle, vitesse de sédimentation), etc., sans introduire aucune corrélation. Les résultats ne permettent pas de répondre à la question fondamentale : toutes ces LOB étaient-elles des leucémies aiguës dès le début ? Si oui, qu'est-ce qui ralentissait l'évolution ?

Dans une deuxième partie, l'A. F. C. a été appliquée d'une part pour obtenir des pronostics de survie comme pour les infarctus, d'autre part pour essayer par des zones discriminantes de préciser la nosologie. Une première A. F. C. semble montrer trois groupes (cette classification n'étant pas auparavant connue) ; les taux de survie exprimés par six paramètres binaires, portés en éléments supplémentaires, ont montré que les trois zones correspondaient à des taux différents de survie, respectivement inférieurs à 6 mois, proches de un an et au-delà. Une deuxième analyse, sur un tableau de contingence croisant les 27 paramètres de pronostic et les taux de survie confirmerait, d'après les auteurs, les résultats. Mais au total ne doit-on pas considérer comme optimiste la conclusion de l'étude Izrael et alia : « ... Quoiqu'il en soit, l'ensemble des données déjà recueillies plaide en faveur d'un concept de « maladie médullaire unique » (1).

b) *Les leucémies lymphoïdes chroniques (LLC)*

Il s'agit, ici, de la prolifération des lymphocytes (variété de globules blancs) dans le sang et dans la moelle, maladie très hétérogène, à délais de survie très différents, des critères précis proposent une nosologie analogue à celle du diabète. « Un sujet est atteint s'il présente au moins deux de trois critères :

- plus de 4 000 lymphocytes par mm³ de sang périphérique et dans la moelle osseuse lors d'un seul examen plus de 25 % de cellules observées constituées par des lymphocytes ;

1. Nous lirons dans l'étude des Professeurs Cl. Jacquillat, J. Bernard, à propos de l'A. F. C. : « On peut la concevoir intuitivement de la manière suivante : chaque individu est représenté par un point hérissé de tous ses paramètres et tous ces individus se regroupent dans l'espace par analyse des paramètres. Le nuage de points et de paramètres ainsi obtenu est, ensuite, projeté sur des places, selon les axes d'élongation maximale du nuage que détermine la machine ». C'est exact certes, mais la suite est dangereuse parce qu'équivoque. « Sur les projections ainsi obtenues, deux individus seront d'autant plus proches que leurs profils se ressemblent et plus la liaison entre deux paramètres est plus grande, plus ils seront voisins. De plus, un individu se projette près des paramètres qu'il présente et inversement autour d'un paramètre gravitent les individus qui présentent ce caractère. »

Les erreurs qui peuvent résulter d'une telle affirmation sont si fréquentes qu'on les a appelées « classiques » ; il est faux de considérer que le voisinage d'un point individu et d'un point observation a un sens. Les deux nuages de points sont en relation *dual* : au facteur $1/\sqrt{\lambda}$ près (λ valeur propre) la représentation d'un point du nuage dual des paramètres est au *centre de gravité* des points du nuage direct des individus convenablement pondérés. C'est ainsi qu'il faut interpréter l'expression « autour d'un paramètre gravitent les individus » que l'on a pu lire plus haut.

On retrouve les mêmes affirmations dans l'étude, en date de janvier 1975, sur l'exploration fonctionnelle de la thyroïde que nous avons citée plus haut : « Un sujet proche d'un point variable aura toute chance de voir cette variable prendre une part prépondérante dans son profil ». On peut aussi y lire ce que l'on sait aujourd'hui être presque toujours inexact : « En règle générale, le premier ou les deux premiers facteurs sont les seuls significatifs, c'est à dire que la quasi totalité de l'information peut se projeter dans un espace à une ou deux dimensions ». Tout ceci montre les pièges que peut présenter une A. F. C. trop rapide.

- plus de 4 000 lymphocytes par mm^3 de sang périphérique à trois reprises en moins de trois mois;
- infiltration de lymphocytes dans la moelle osseuse reconnue par un examen histologique. »

L'étude statistique de Bastin, Rain et alii (1977) est une des premières A. F. C. homogènes et complètes. Elle porte sur 102 malades, mais la maladie n'a pu être suivie que sur 90 (1967-1975). Les diverses données biologiques et cliniques, codées qualitativement, ont fourni un tableau de chiffres 0,1 de 102 individus et 91 modalités dont 21 en éléments supplémentaires. (Pour l'étude de l'évolution, 90 individus et 42 modalités dont 4 en éléments supplémentaires, le tableau 392×42 est obtenu par superposition des divers stades d'évolution, chaque ligne décrivant un état d'un individu à un certain stade).

● *Première analyse* : Les pourcentages d'inertie sont, ici, très petits, respectivement 8, 7, 6 et 5 %, ce qui n'est pas étonnant étant donné le nombre élevé des axes factoriels mais paraît nouveau en recherche médicale. Trois directions se dessinent sur le plan des premiers axes montrant d'une part que pour toute une série, les examens sont fortement liés, ce qui pourrait permettre de supprimer certains particulièrement douloureux, d'autre part qu'un certain classement des individus en cinq régions serait « possible » celles-ci n'étant toutefois pas très nettes et ne permettant pas de classer tous les malades. L'axe 3 serait lié à une forme spéciale d'atteinte (cytopénie ⁽¹⁾) dont il y aurait deux formes, centrale et périphérique, les deux étant nettement différenciées sur la plan factoriel (1, 3). L'étude des conditions de survie met en évidence la gravité du diagnostic de cytopénie et, écrivent les auteurs, « ceci est un résultat suggéré déjà par certains médecins mais pour la première fois mis en évidence d'une façon aussi nette par cette analyse et confirmé par l'étude de l'évolution de la maladie. Il peut, bien sûr, avoir une influence sur la thérapeutique à envisager dans le cas d'un tel diagnostic ».

● *Deuxième analyse* : Avec des valeurs propres et des taux d'inertie plus élevés (il y a moins d'axes factoriels), elle confirme dans l'ensemble les résultats précédents, extrême importance du diagnostic mais pas de classification nette des malades permettant une discrimination des diverses LLC.

c) *La leucémie myéloïde chronique (LMC)*

Il s'agit ici d'une des modalités de la troisième frontière des LOB, le syndrome myéloprolifératif (ici, en général, plus de 50 000 leucocytes par mm^3). C'est une étude considérable, 798 malades entre 1957 et 1973, avec 156 données cliniques, hématologiques, anatomiques et biologiques ⁽²⁾.

Deux séries distinctes d'analyses statistiques :

- analyse des survies, avec calcul d'écart type pour les diverses caractéristiques du malade ou de la maladie, âge, sexe, activité clinique, taille de la rate, hémoglobine initiale, globules blancs, plaquettes, chromosome Philadelphie, etc.; certains éléments sont des pronostics favorables, d'autres non;

1. Ce mot n'existe pas dans l'index alphabétique de l'Abrégé d'hématologie (1978) du professeur Jean BERNARD qui comporte pourtant près de cinq cents termes.

2. Dans son tome II de l'Analyse des données, page 22, édition 1973, le professeur J. P. BENZECRI prend comme exemple pour discuter de l'exhaustivité d'un ensemble de données les analyses de sang. « On ne sait au juste ce qu'est un inventaire complet de ce que recèle le sang, on se borne dans les examens les plus détaillés à quelques éléments, quelques substances dont l'importance a déjà été connue. » Le « Blood Cells Club », réunissant les principaux hématologues du monde entier, a tenu précisément du 15 au 17 juin 1979, à Kremlin Bicêtre, sous la conduite du professeur Jean Bernard, une session de trois jours intitulée : « What to measure and why? » Nous avons eu l'honneur d'y être invité et attendons avec une impatience statistique l'édition des textes.

— analyse multidimensionnelle très riche : A. F. C., discrimination, régressions multiples, analyse en composantes principales et classifications automatiques.

Nous n'avons d'informations précises que pour l'A. F. C. : elle a porté sur 460 LMC et 13 facteurs pronostiques « sans valeur marquante » (sexe, signes cliniques et biologiques) conduisant à 36 paramètres. Le premier facteur oppose des types de patients bien définis par des signes biologiques et le volume de la rate, le deuxième oppose des signes cliniques sauf la fièvre et l'aspect en parabole du nuage dans le plan factoriel (1, 2) suggère une rotation du second degré. (Ici, nous ferions attention à l'effet Guttman qui intervient quand il y a une variable unidimensionnelle cachée.)

Le troisième facteur sépare les sujets dont les signes biologiques s'accompagnent ou non de signes cliniques. On obtient ainsi un certain ensemble de facteurs pronostiques. Mais aucun résultat saillant n'a pu être énoncé : unique conclusion déjà importante par elle-même, même si elle n'est évidemment que provisoire, « les protocoles futurs de LMC devront tenir compte de ces facteurs pronostiques dans la vigueur des traitements proposés et l'interprétation des résultats obtenus ».

III — ÉVALUATION DES MOYENS D'ACTION

1. La théorie des catastrophes

Avant de reprendre nos réflexions sur l'analyse des données, nous voudrions faire un bref retour vers la causalité en montrant comment cette notion peut être considérablement enrichie par une branche des mathématiques en plein essor, la topologie. Celle-ci peut, en effet, éclairer d'un jour nouveau certains problèmes de recherche médicale comme le cancer ou la réanimation médicale et suggérer utilement des directions de recherche.

Tout a commencé avec un livre de René Thom ⁽¹⁾, mathématicien français, tirant de la topologie une série de modèles possédant, grâce à leur structure interne, des propriétés globales *qualitatives* très importantes et dont on n'avait jamais pu obtenir jusqu'ici une représentation mathématique adéquate.

Nous avons, en 1976, essayé nous-mêmes de montrer comment cette théorie pouvait permettre de faire de grands progrès dans les sciences de l'environnement ⁽²⁾.

Un de ses disciples, E. C. Zeeman ⁽³⁾, l'a utilisée pour les sciences humaines après l'avoir mise au point pour les battements du cœur, en l'appliquant aux agressions, aux émeutes de prisons, à la croissance économique, au krach des bourses de valeurs, à la course aux armements, aux stratégies militaires, etc.

Mais si le livre de R. Thom était surtout inspiré par l'embryologie, peu de biologistes, sans doute à cause de son niveau mathématique exceptionnellement élevé, ont utilisé ses idées. Aussi, nous voulons en dire quelques mots ici, renvoyant à notre article déjà cité pour des compléments nécessaires.

1. R. THOM, Stabilité structurelle et morphogénèse. Benjamin, New York 1972. Édition française 1977. L'utilisation pour sa théorie de l'appellation « théorie des catastrophes » a apporté un succès journalistique ~~consi-~~ dérable et tout le monde en a parlé souvent de manière très inexacte.

2. R. GIBRAT, Énergie et environnement après l'an 2000. *Sciences et Techniques*, n° 3, novembre 1976, pages 5 à 14.

3. *Scientific American*. Avril 1976.

Les systèmes que nous rencontrons en environnement et dans de nombreux domaines de la recherche médicale sont caractérisés par un très grand nombre de variables et même en acceptant a priori que la variable que l'on cherche à déterminer (par exemple, le déclenchement de la métastase dans un cancer) soit une vraie fonction des autres variables définissant une vraie relation cause effet, l'introduction de *toutes* les interactions sous forme d'équations différentielles, aux dérivées partielles, intégrales différentielles, etc., paraît au delà des forces humaines bien que théoriquement le résultat cherché soit la conséquence du jeu de ces équations.

Car recherche médicale et écologie (pensons aux problèmes soulevés par les climats) sont souvent caractérisées par un manque presque général de données valables et une compréhension plus que médiocre des mécanismes fondamentaux. Suivant l'expression des deux chercheurs de l'I. I. A. S. A. (1), les solutions sont « localement rigides et globalement vagues » alors que l'on observe les résultats globaux, la situation étant « globalement rigide et localement vague ». Les outils de l'analyse ont été créés pour l'analyse locale et non pour le survol global (2). La causalité est entièrement masquée par la complexité. La topologie, au contraire, paraît être adaptée aux nouveaux problèmes et permet de respecter le rapport cause effet, auquel, au fond de nous mêmes, nous tenons tant.

Car Thom a démontré que chaque fois que l'on peut séparer les variables définissant un système en un nombre k petit de variables dites lentes, les autres en nombre aussi grand que l'on veut, étant qualifiées de rapides, *le comportement en période exceptionnelle* (ce que Thom appelle « catastrophe ») ne dépend *que du nombre des variables lentes* et toute une série de propriétés très importantes ne dépendent que de leur nombre, ce qui est un résultat considérable et inattendu.

Théorème :

Si $k \leq 5$, sous des conditions très générales :

1. Le lieu des points d'équilibre est une multiplicité de dimensions k .
2. Une projection sur l'espace des variables lentes est normalement stable sous de petites perturbations de la fonction f .
3. Toute singularité de cette projection équivaut à une d'un nombre fini K de catastrophes élémentaires donné par le tableau suivant :

k	1	2	3	4	5
K	1	2	5	7	11

Par exemple, pour $k = 2$, il a défini le *pli* et le *point de rebroussement*; pour $k = 3$ il y a en plus la *queue d'hirondelle* et les *ombilics* hyperboliques et elliptiques, etc.

1. S. BEER et J. CASTI, Investment against disaster in large organisation I. I. A. S. A. Research Memorandum RM 75.16 (avril 1971).

2. On notera, cependant, que cette idée de Thom que la topologie est la science du passage du local au global n'est pas acceptée de façon générale. (C.P. BRUET, Topologie, Encyclopédie Einaudi, 1978). Celui-ci, par exemple, n'accepte pas le parallèle trop rigoureux entre « les changements qualitatifs observés en fonction du nombre et de l'organisation en mathématiques d'une part et dans les domaines de la vie d'autre part ». On lira avec fruit aussi F. PERROUX : « Unités actives et mathématiques nouvelles », pages 44 à 74 (1975).

Au-dessus de 5, le nombre est infini et la classification des catastrophes relève du continu. Mais 5 paramètres donnent une richesse de situation probablement suffisante pour la plupart des cas qui se présentent en pratique.

L'observateur d'un système sait maintenant, grâce à cette théorie, que des changements continus ou des perturbations d'un système peuvent donner de grandes discontinuités et que le retour du système initial ne se fera pas toujours en le ramenant aux conditions initiales. Il s'en doutait peut être déjà; mais de plus il saura qu'il y a de grandes chances que quelle que soit la complexité de son système (le nombre de paramètres) et quelle que soit son ignorance sur le comportement détaillé de chaque réaction, il n'aura devant lui qu'un nombre très limité de comportements en époque exceptionnelle.

Si le chercheur a observé dans la nature un comportement de sauts en « catastrophes » il pourra essayer, d'après leur aspect, de deviner combien il y a de paramètres lents à mettre en évidence.

On devine facilement l'application, par exemple, aux problèmes du cancer. Cependant, R. Thom, dans l'édition de 1977, n'y fait allusion qu'une seule fois, page 219, dans le paragraphe 10.3.B intitulé « la reproduction dans le domaine hydraulique »; il présente un modèle topologique qu'il considère comme une sorte de volcan, dominant tout le paysage, la lave s'écoule lorsqu'elle atteint le bord du volcan et il écrit : « il n'est pas interdit de se représenter ainsi les faits de *cancérisation* » (en italique dans le texte).

En vue d'éclairer notre propos, nous citerons le cas de l'insecte dévorant *parfois* les pins résineux du Canada ⁽¹⁾.

Un équilibre très satisfaisant existe normalement entre le nombre des insectes dévoreurs et le feuillage des arbres et peut durer soixante dix ans; mais, parfois, brusquement en 3 ou 4 années la densité des insectes est multipliée par plus de cent mille. Tout le nouveau feuillage et une partie de l'ancien est consommé et la mortalité des arbres est presque générale. Heureusement, les très jeunes arbres sont indemnes ou à peu près et délivrés de leurs parents dont l'ombre ralentissait leur croissance, ils croissent pendant que les insectes reviennent faute de nourriture à leur état endémique; la crise complète peut durer de 7 à 14 ans.

La variable rapide est évidemment la densité des larves de l'insecte. On a commencé l'étude avec une seule variable lente, la surface du feuillage. Mais la *catastrophe pli* seule possible avec $k = 1$ ne présente qu'un retour avec saut à la position d'origine et cela paraissait insuffisant pour expliquer les faits. Il fallait introduire la possibilité d'un retour continu. La variable lente fut donc divisée en deux. La première fut la surface utilisable par l'insecte pour y vivre (surface des branches) fonction croissante avec l'âge de l'arbre, l'autre paramètre fut la surface du feuillage par unité de surface de branches, fonction admise en première approximation comme indépendante de l'âge de l'arbre. On a alors $k = 2$ (habitat et nourriture), d'où possibilités d'une *catastrophe rebroussement* nécessaire pour les retours continus. Ceci a permis, par un modèle adéquat reliant les paramètres entre eux, d'étudier très à fond ce problème fondamental pour les forêts du Canada. On a en particulier pu déterminer les conditions assez inattendus dans lesquelles il fallait faire intervenir les insecticides, l'attention étant attirée non pas comme auparavant par la dynamique de croissance des larves, mais sur les deux paramètres lents.

Tout spécialiste du cancer saura traduire dans son langage les paragraphes précédents. Il s'agit de rechercher pour chaque cas, d'après l'aspect global, le nombre de variables lentes qui le représente le mieux.

1. DIXON P. JONES, The application of catastrophe theory to ecological systems. I. I. A. S. A. Austria Research Report 75 — 15 juin 1975.

Tout cela n'est, cependant, que pour demain et l'analyse des données reste la ressource pour aujourd'hui comme nous allons le voir.

2. *L'analyse des données*

Un gros effort a donc été fait dans les dix dernières années par certaines unités de recherche médicale pour utiliser l'analyse des données et ceci dès l'apparition de l'A. F. C. Les buts ont été ambitieux et les résultats peut-être un peu décevants pour des raisons qu'il convient d'analyser mais qui ont, pour la plupart, disparu aujourd'hui.

a) D'une part, les premières études ont souffert de ce que l'A. F. C. n'a pas eu tout de suite son *aspect définitif*. Le professeur Benzecri, lui-même, écrit, dans les *Cahiers de l'Analyse des données* (avril 1977) qu'elle vient seulement d'établir ses structures essentielles : « L'analyse des données, telle qu'on la pratique en 1977, ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle donne, pour la préparation des données, des règles telles que le codage, sous forme disjonctive complète, aide à critiquer la validité des résultats, principalement par des calculs de contribution, fournit des procédés efficaces de discrimination et de régression, se conjugue harmonieusement avec la classification automatique. Ainsi, une méthode unique dont le formulaire reste simple est parvenue à s'incorporer des idées et des problèmes nombreux apparus d'abord séparément, certains depuis plusieurs décennies ».

b) D'autre part, les précautions nécessaires n'ont peut être pas été toujours observées dans l'*interprétation des résultats*. Nous l'avons montré à diverses occasions dans notre examen des études des faits. A de nombreuses reprises, dans les deux tomes du professeur Benzecri, celui-ci a mis en évidence les difficultés que peut entraîner l'interprétation des résultats de l'analyse des correspondances et multiplié les mises en garde. Mais sa marche continue en avant ne lui a pas encore permis, à notre connaissance, d'écrire la synthèse correspondante.

On doit, par contre, beaucoup aux recherches de L. Lebart ⁽¹⁾. Ainsi, à propos des valeurs propres et des taux d'inertie, contrairement à l'opinion générale, il écrit : « L'utilisation de ces paramètres n'est pas toujours justifiée en pratique. Leur interprétation est de plus délicate ... car ils sont étroitement liés au codage des données ». Il a cherché à donner des « garde fous » en calculant des seuils de signification. Il a construit des tableaux de données avec indépendance totale entre lignes et colonnes. Les valeurs propres et les taux d'inertie correspondants résulteront donc seulement des fluctuations d'échantillonnage et pourront être considérés comme des seuils. Toutes valeurs des paramètres en dessous, ou seulement voisines, seront donc sans signification.

Il veut ainsi dissuader les chercheurs d'utiliser sans réfléchir les taux d'inertie pour « noter la qualité d'une représentation » et de les interpréter en « pourcentage d'information ». Ainsi, il a montré comment des « options » de calcul peuvent, pour un même ensemble d'observations et pour une même description finale, modifier totalement des paramètres auxquels on aurait voulu attribuer un « pouvoir explicatif ».

Il a essayé aussi d'approfondir la notion même d'information et les différents concepts réunis sous ce nom. Sa conclusion essentielle est qu'en analyse des données, il vaudrait mieux parler de forme que d'information. La validité d'une représentation serait mieux établie en étudiant la stabilité de la forme plutôt qu'en mesurant une quantité d'information.

1. Validité des résultats en analyse des données. (Recherche financée par le D. G. R. S. T.). Novembre 1975, 158 pages.

Tout ce travail de Lebart est fondamental et doit être connu de tous ⁽¹⁾.

Aujourd'hui, on dispose donc d'un outil statistique au point et d'un certain nombre de statisticiens expérimentés. On a, cependant, continué à diriger toute une série de critiques contre l'analyse de données. L. Lebart, dans son chapitre III de l'étude citée, en examine successivement quatre types :

- l'utilisation inconsiderée. Il parle des « monomaniaques » de la technique dont il faut réfréner l'ardeur dévastatrice;
- le fait que l'évidence *a posteriori* n'est pas toujours celle *a priori* (« beaucoup d'assertions sont possibles, aussi les résultats apparaissent évidents *a posteriori* »);
- l'usage de l'ordinateur rend le travail trop facile et introduit des spécialisations regrettables. On perd le contact avec le matériel brut, point capital dans le domaine médical ⁽²⁾;
- on ne travaille pas directement sur des structures mais sur des tableaux. Il s'agit ici des hypothèses sous-jacentes, qu'elles soient ou non exprimées, car il est toujours difficile de savoir si elles vont ou non introduire des erreurs. Leur importance dépendra de leur « robustesse » et ce serait une force de l'A. F. C. d'engendrer « presque toujours » les hypothèses plus faibles.

Le professeur Benzecri a, certes, écrit ⁽³⁾ : l'A. F. C. est « une méthode qui, bien mieux que toute autre, nous a permis de découvrir les faits de structure que recèle un tableau de données quel qu'il soit ».

Mais le doute a percé en quelques rares aveux qu'il nous faut garder en mémoire pour éviter les abus : « Dans quelles limites sommes nous capables d'accomplir ce magnifique programme? ... Comment démontrer que, de la manipulation purement mathématique des tableaux de données, sortent des résultats assez significatifs pour que le spécialiste les lise et les accepte comme une évidence? Ne nous vantons pas trop... » Sa méthodologie, cependant, est impitoyable. « Le modèle doit suivre les données, non l'inverse. Découvrir sans participer, sans *a priori*, quels courants de lois traversent l'océan des faits », et encore cette belle formule, « asservir la chair des données à l'âme des formules », et ce cri du cœur « la statistique n'est pas seulement un mode de connaissance, elle est un mode d'être ».

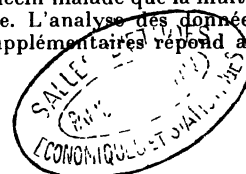
Pour notre part, nous nous rangerons volontiers à la conclusion de G. Morlat dans son excellente introduction à l'ouvrage de P. Caillez et Pages ⁽⁴⁾ : « L'analyse des données doit rendre service partout où l'on se soucie d'accumuler des observations... Les services rendus montrent bien que l'analyse des données constitue aujourd'hui, et de loin, la partie

1. Citons aussi deux importants ouvrages : L. LEBART et J.-P. FENELON, « Statistique et informatique appliquée ». Dunod, 1973 et L. LEBART, A. MORINEAU, N. TABARD, « Techniques de la description statistique ». Dunod, 1977.

2. On nous permettra, pour montrer combien nous y sommes sensibles, un très bref historique. Au début du XIX^e siècle, malgré les travaux d'Auenbrugger (1761) sur la percussion, le diagnostic était fait sur la base des comptes rendus de sa maladie par le patient lui même et certaines observations du médecin *faites à distance*, couleur de la peau, état de la langue, respiration, etc. T.-H. Laennec (1816) inventait le stéthoscope et l'auscultation dans un monde où la discrimination des maladies des poumons et du cœur était bâtie sur la conviction que des défauts anatomiques caractérisaient les désordres. En 1895, la découverte des rayons cathodiques par Roentgen donnait une extraordinaire impulsion à la vision directe. Cependant en 1979, percussion et auscultation occupent toujours en France une grande place, aussi bien dans l'enseignement et dans les livres que dans la pratique médicale même à haut niveau. Ainsi s'est conservé le contact personnel médecin malade que la multiplication presque automatique des analyses de laboratoire auraient pu faire perdre. L'analyse des données en personnalisant le patient dans les diagrammes par la technique des éléments supplémentaires répond au même souci.

3. La place de l'*a priori*. Texte ronéotypé.

4. Introduction à l'analyse des données. Smash. 1976.



la plus immédiatement rentable de la statistique... Cela permet, selon les cas, de découvrir dans les phénomènes étudiés des structures directement visibles sur les résultats de l'analyse, alors qu'elles ne l'étaient pas sur les données originelles, ou de retrouver en les précisant des des structures que l'on soupçonnait déjà pour telle ou telle raison »

Les buts des pionniers de l'A. F. C. en recherche médicale n'étaient donc pas trop ambitieux, bien au contraire.

IV — CONCLUSIONS

Nous avons conclu dans la première partie : inégale « sensibilité » des êtres vivants mise en évidence dans la réaction aux antibiotiques ou aux insecticides, donc a priori à toute thérapeutique, classification « botanique » des maladies peu satisfaisante ou peu fiable, facteurs de risques sans relations causales traduisant des « prédispositions » à diverses maladies, mots clefs qui rappelleraient trop la vertu « dormitive » de l'opium, si le médecin n'acceptait pas, un peu à son corps défendant, de devenir probabiliste et si les théories statistiques n'avaient fait de très importants progrès précisément ces dernières années.

Depuis vingt années au moins, l'essentiel de la recherche médicale est consacré à la recherche des causes des maladies. Sur ce point précis l'échec est complet ⁽¹⁾ dans tous les domaines ou presque et il en sera ainsi très probablement tant que la recherche fondamentale biologique n'aura pas élucidé les raisons profondes des différences entre individus (immunologie et H. L. A., sensibilité ou prédisposition). Nos connaissances de base sont insuffisantes aujourd'hui pour permettre d'élaborer des directives de recherche.

Aussi l'impression que donne la recherche médicale à un observateur extérieur est d'être une recherche de luxe. On ne veut négliger aucune piste si faible que puisse apparaître la probabilité de réussite et, si on n'en a pas les moyens, on les réclame au nom du droit à la santé de chaque homme et que le financier ait mauvaise conscience s'il refuse ! Tant que l'on a travaillé sur des domaines où régnait la causalité et où pouvaient s'appliquer les règles classiques de la recherche scientifique, par exemple construire une expérience telle que les faits observés permettent de répondre par oui ou par non à une question précise, le succès a été prodigieux et comparable à ceux des physiciens. Certes, on a peu à peu mis en évidence autour de maladies aux frontières souvent mal définies des ensembles de facteurs dont on est à peu près assuré qu'ils définissent un risque. Certes, aussi le « génie » (le mot n'est pas trop fort) de certains médecins a su parfois tourner l'obstacle et obtenir des résultats splendides en s'attaquant, sous des formes variées, aux *conséquences* faute de pouvoir attaquer les *causes* (anémie pernicieuse, réanimation médicale, etc.) ⁽²⁾.

Certes, la biologie et particulièrement la biologie génétique n'ont pas cessé de progresser, « la déduction qualitative et empirique leur offrait (Cf. notre épigraphe) un support suffisant aux possibilités de prévision et d'expérience ». Elles sauront se mathématiser à temps. Aussi, il faut désormais, diminuant progressivement l'effort vers les causes, d'une part mettre l'accent sur le fondamental pour permettre la recherche médicale de demain, d'autre part prendre en compte fermement le fait que la recherche médicale d'aujourd'hui est avant tout analyse des données.

1. Au moment où nous écrivons ces lignes une association déclenche dans les mass media une campagne pour obtenir des fonds privés pour la recherche dans le cas de la sclérose en plaques. Ici encore l'aveu est franc : on ne sait rien sur les causes.

2. Nous n'oublierons pas les résultats extraordinaires obtenus à ce jour par les médecins, ainsi le risque de mort en cas de grossesse aux U. S. A. est passé de 582 pour 100 000 en 1935 à 33 pour 100 000 en 1964. Les médecins ont aussi constitué un ensemble de données absolument prodigieux que l'informatique va leur permettre d'utiliser.

Cette évolution vers l'analyse des données est aussi celle de toutes les techniques industrielles. Partout l'aspect « analyse des données » domine les recherches avancées. Le premier effet en recherche médicale d'une utilisation générale de l'A. F. C. sera de réduire considérablement le nombre d'analyses nécessaires à un diagnostic ou à une réanimation médicale et permettre des pronostics de survie indispensables pour orienter les thérapeutiques (1). A plus long terme, le développement généralisé de l'analyse des données orientera le chercheur, jusqu'ici réticent, vers une évaluation de plus en plus scientifique des thérapeutiques, évitant l'évident gaspillage actuel et diminuant le danger pour le malade des prescriptions mal adaptées. La nosologie sera complètement transformée. Des zones convenables dans les plans principaux d'inertie affermiront les diagnostics cliniques, etc. Les conséquences sur le budget de la Sécurité Sociale en seront très grandes, n'en doutons pas. Les ouvrages statistiques destinés aux médecins devront être rédigés d'une toute autre manière en insistant sur l'analyse des données et la nécessité de ne faire que de la « bonne » statistique, ce qui exigera un haut niveau. Il faudra y multiplier les exemples de ce qu'il ne faut pas faire (2).

Cet aspect statistique n'est pas définitif, bien au contraire. G. Morlat, dans sa préface de 1976 déjà citée, s'exprimant pour tous les domaines d'utilisation écrivait : « *Pour quelques années encore, priorité à l'analyse des données* ». Pourquoi? Parce qu'un jour le chercheur « voudra prévoir les phénomènes (et il lui faudra adopter un modèle probabiliste, ...) ou éclairer des décisions et il se posera des problèmes du type de ceux qu'abordait la statistique mathématique classique ». Retour éternel des disciplines...

Dans quelques années ou quelques décades, la biologie mathématisée rendra possible des modèles, rétablissant les lois scientifiques et la causalité en recherche médicale. Les professeurs J. F. Bach et J. Bernard ont entrevu des directions possibles.

Sans oublier, comme l'écrit le professeur Benzecri, précisément à propos de l'A. F. C. « le trait dont nous dessinons nos conclusions doit être aussi large que la place laissée au doute » *priorité donc pour quelques décades dans la recherche médicale à l'analyse des données.*

Alors, après avoir ressenti longtemps la joie de connaître, le médecin jouira de celle de comprendre.

1. Dans son livre sur « La puissance et la fragilité », le professeur J. HAMBURGER parlant des « choix subtils qu'une enquête exige parmi dix explorations possibles et la moisson de documents qu'il faudra dépouiller pour répondre à la question posée... » parle du « vertige numérique (souligné dans le texte) des données qui peuvent intervenir aujourd'hui pour obtenir un diagnostic précis ». Malheureusement il ajoute : « Mais ce vertige de la quantité n'est rien.. » Or, pour nous, sa suppression seule permettra de désembourber le char.

2. Il faudra, aussi, étendre les connaissances probabilistes de chaque médecin. Un médecin célèbre incontestable a écrit récemment : « Un malade qui s'entend dire : si l'on applique pas le traitement que je vous soumetts, vous avez huit chances sur dix de mourir la semaine prochaine; le traitement a trois chances sur dix de vous tuer, mais il réduit à 30 % les craintes que nous avons de vous voir succomber. Ce malade là est un homme perdu ... par l'extrême anxiété qu'on a suscité en lui ». Nous craindrions surtout qu'il n'ait absolument pas compris qu'avec le traitement il aurait presque une chance sur deux de guérir (49 %). Mais est ce cela qu'a voulu dire le médecin en appliquant la théorie des probabilités conditionnelles?