How many general practitioners will fit on a ROC curve?

The impossibility of medical decision analysis in daily practice

J.O.M. ZAAT

Zaat JOM. How many general practitioners will fit on a ROC curve? The impossibility of medical decision analysis in daily practice. Huisarts Wet 1993; 36(Suppl): 54-7.

Abstract Medical decision making is becoming more and more important. Ardent admirers plead not only for its use in making standards but also in daily practice. With more knowledge about test characteristics GPs would become more rational in requesting diagnostic tests. But there are some major methodological and cognitive psychological problems. First: the diagnosis is often not a well defined reality, but is often only a statement about prognosis. A golden standard is therefore almost always lacking, so calculations with sensitivity and specificity are rather hazardous. Logistic regression analysis is one of the more sophisticated techniques to make prognostic scales, but the results can often not be reproduced in slightly different populations. Calculations in individual cases are therefor quite risky. Second: human beings do not use prior and posterior probabilities in reasoning. This is not only an understandable aversion but mainly a basic psychological fact. Nobody can calculate Bayes theorem or use complex scoringsystems by head. Clinical epidemiology is very useful as basic science in general practice but you can not use its principles in detail when Mr Jones is sitting in front of you. Scientists should realize that they only look through a narrow hole and that the reality is often more complex than their most sophisticated statistical technique.

J.O.M. Zaat, MD, PhD, Vrije Universiteit, Department of General Practice and Nursing Home Medicine/Institute of Extramural Medical Research, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. Telephone: +31 (0)20.548.4553.

Introduction

Diagnostic tests are more and more frequently the topic of publications, including articles published in *Huisarts en Wetenschap*. This concerns not only tests that are considered 'classical', but also tests that use patient characteristics for making predictions about the presence or absence of a disease. The techniques and measures that are being used become more and more sophisticated: only a few years ago it was considered highly proper if sensitivity, specificity and predictive value of the test studied were reported;[1] nowadays likelihood ratios, odds ratios and ROC curves appear to be the minimum that should be available.[2-4] Publication of the Dutch College of General Practitioners standards for sore throat and urinary infections elicited letters in the correspondence section of *Huisarts en Wetenschap* that described various ways of calculating a predictive value.[5,6] This journal must naturally provide a forum for methodological disputes between scientists, but as far as the average general practitioner is concerned, these discussions are beginning to resemble the sixteenth century polemic about the number of angels that can dance on the head of a pin.

There is a steady flow of publications advocating training in clinical epidemiology for medical (general) practitioners, so that they may become more rational decision makers.[7,8] The training of general practitioners almost universally includes separate 'medical decision analysis' modules and general practitioners are taught in several postgraduate training courses that thinking along decision analysis lines also has a useful place in the daily care for individual patients. It even appears from discussions with researchers that a lot of confidence is placed in the possibility that diagnostic and therapeutic decisions could be rationalised, should physicians use medical decision analysis in the office. However, I feel more and more that the application of medical decision analysis in the daily practice of medicine - the direct doctor-patient contact - is hardly reasonable or even possible. In this contribution to the discussion I shall indicate where difficulties arise in daily practice if medical decision analysis were to be used for solving *diagnostic* problems, first using a methodological, then a cognitive-psychological approach.

What is wrong?

A diagnosis is not always a given, well defined reality. This is the most important problem with many diagnostic investigations. Diseases are not evil spirits that attack us, but man-made concepts, as *Wulff* so aptly formulated.[9] Physicians do not discover facts through digging with perseverance, but instead they arrive gradually at a concept of reality.[10] Many 'diagnoses' are not much more than conventions among physicians. For example, to satisfy the diagnosis of rheumatoid arthritis, the patient must meet the criteria of the American Rheumatism Association. Preference for certain diagnoses is culturally determined: in Germany one can die of cardiac diseases that do not exist in other EC countries.[11] At least in general practice a diagnosis is no more than a starting point for thoughts about therapy or a basis for statements about prognosis.[12]

The lack of objective reality, the true gold standard, makes it rather hazardous to carry out calculations based on test characteristics such as sensitivity and specificity. Even using the course of an illness episode as gold-plated standard, for lack of a solid gold standard, does not solve this problem: in final analysis one is faced with the same problem of definition. Obviously this does not hold for all disorders: Chlamydia can be cultured, so that a culture appears to be a reasonable gold standard - until it appears that a Chlamydiazym estimation reveals more Chlamydias than a culture does.[13] Even the result of coronary angiography, considered a gold standard in clinical investigation, proves to show considerable variation between different radiologists. In short, there is a vast difference between our diagnoses and the diseases that are actually present.[14]

Application of medical decision analysis to daily practice is limited by the

frequent lack of necessary data, such as the sensitivity and specificity of many of the diagnostic tests.[15] Even if these should be available, the physicians' life would not become any easier. *Dekker et al.* have, for instance, considered the diagnostic values of several of the characteristics of a diagnosis of Chlamydia.[13] Applying a multiple logistic regression analysis, they produced a 'Chlamydia scale': intercourse without condoms, 19 points; age under 28 years, 11 points; no itch, 11 points; intercourse with more than two men, 10 points; either Surinam or the Antilles as country of origin 10 points; vaginal irrigation 10 points. A score of 37 points or more indicates a chance of over 10 percent that there is a Chlamydia infection, therefore a test should be performed. These characteristics were compiled from a large group of women in Amsterdam. The scale, however, has not (yet) been validated for other groups, so that it would be difficult for a general practitioner in The Hague whose practice includes, for example, many patients from other parts of the world, to assess the value of this instrument. Even without considering the practical difficulties, would anyone be prepared to quickly add all these items, during a consultation, and then, should their sum exceed 37, remember to ask for a Chlamydiazym test? Also, the reproducibility of these 'predictive rules' is somewhat unclear. A reproducibility study has shown that the results of logistic regression analysis are only seldom identical.[16]

Small errors in the assessment of a priori chances are of great consequence for the results of the decision tree and thus for reaching the 'correct' decision.[17] Moreover, these small errors at the beginning of the diagnostic process cannot be eliminated, even by using better techniques. Meteorologists have long thought that, given more refined software and better computers, they would produce longer range forecasts. Meanwhile, they have had to accept the idea that they cannot forecast the weather to any extent for longer than a few days. Small deviations that appear at the beginning of all calculations about the course of a depression eventually, after many calculations, appear to lead to considerable uncertainty.[18] It is much the same with the diagnostic process in general practice. The level of the probability that a patient suffers from a particular disease is the threshold value which the physician will apply when deciding to test, to treat, or to do neither. There are advanced methods to calculate these thresholds,[19] but it appears that there are few investigations regarding the threshold values actually used by physicians.[20] If physicians assess threshold values differently, due to their own experience or willingness to take risks, the result will immediately be quite different. Training as to how to deal with exact numbers will not help many of us. As the individual patient does not offer exact a priori values, physicians use assorted terms such as 'possible', 'perhaps', 'almost certain', etc. Asking physicians to quantify these terms is not a solution, because translations into exact numbers will differ for different people.[21,22] A study of this 'translation' showed that the concept 'never' had a dispersion of 0 to 72 percent.[23] These findings were confirmed by recent Dutch research.[24] It is not only the descriptions or translations of probabilities that differ between physicians, but the probability estimates themselves vary for any given well-known problem. In the case of one problem, the estimate of the chance that a particular disorder would occur varied from 2 percent to 90 percent.[25] Decision analysis and research produce statements concerning groups (the 'objective probability'), but the general practitioner faces a patient with a 'subjective probability'; the patient either is or is not ill. This difference between group and individual can hardly ever be reconciled.

Why does it not work?

One of the assumptions of the 'Expected Utility Theory' common in medical decision analysis is that people are prepared to make the most rational choice. In reality, however, people frequently choose strategies that do not yield maximal gain (or ensure minimal loss), as supposed by the theory. This aspect has been reviewed excellently by *Yates*.[26] Even if the theory did apply, it still remains difficult, if not impossible, for ordinary physicians to work with this theory. There is an abundant, often controversial, literature about diagnostic thinking by physicians. This will not be reviewed in detail, but some relevant points will be selected.

The human brain is not structured particularly rationally.[27] We need a great deal of information to be able to solve complicated problems, but we can only store 5 to 7 bits of information in our short-term memory. Only calculation fiends or 'idiots savants' can do a mental calculation of the result of the Bayes formula. If our working memory is too limited for dealing with somewhat more complicated decisions we actually fall back on all sorts of rules of thumb. Such rules are often prone to bias of various types.[28,29] Several of these rules have been tested in medical situations. For instance, if physicians have extra and irrelevant information, they do not take a priori chances into account. Because of the 'noise' in the information they wrongly assess the similarity of the problem presented with that of a classical anamnesis.[30] When they estimate the probability of a disease physicians are influenced to some extent by the possible seriousness of the illness and they allow information obtained later in the diagnostic process to be coloured by opinions formed earlier.[31] Also, they emphasize positive findings more strongly than negative ones and are inclined to ask for more information than they are able to assimilate.[32] *Eddy & Clanton* analysed the diagnostic process involved in solving a complicated case from the New England Journal of Medicine.[33] As is well known, physicians appeared to gather up the data into a few working hypotheses and then select a key symptom. They produced a listing of diagnostic possibilities around this key symptom. They subsequently used this key symptom to pair and weigh the possibilities (the favorite diagnosis against a possible other diagnosis). Finally they tried to place the entire clinical pic-

ture within the framework of the one remaining diagnosis. According to them, not a single chance estimate is involved in the entire process. By quickly selecting a key symptom and thereafter reasoning by paired analysis the physician is able to oversee the whole process. Comparable processes have been described by others.[34-36]

Conclusion

To use decision trees and a priori and a posteriori probabilities in ordinary patient care is not characteristic of general human methods for problem solving. We shall never be able to do this in the office or at the bedside because of our cognitive limitations. Those who are optimistic about the benefit of decision analysis for the rationalisation of medical functioning should above all take note of the relevant literature not directly related to decision analysis. A great many factors influence diagnostic performance. Knowledge about tests is of only secondary importance.[37] With scientific approaches on all sides, part of the medical act remains magic and is based on 'pre-scientific thought'.[38,39] The eventual construction of a reality in which problem solving remains central includes various patient characteristics, organisation of care, possibilities for therapy, test characteristics, time, physician characteristics, etc. To act as though an understanding of test characteristics would on its own, produce better, more rational physicians is a gross oversimplification of reality.[40] Investigators need to simplify a given situation if they are to be able to investigate it, but when relating their results to the real situation they must not forget that they have had only a small look through a narrow keyhole.

Some time ago *Vandenbroucke* wrote about a persistent misunderstanding.[41] I did not and do not agree with his reasoning that in individual cases it is the obligation of the physician to strive towards maximal certainty, but I do agree with the conclusion: training in decision analysis is a necessary mental discipline for general practitioners but is unusable as an instrument in their daily contact with patients. When authors discuss their results concerning the predictive values of tests or patient characteristics they should devote a special discussion to the practical relevance of the results of their studies and should take into account the cognitive psychological (im)possibilities of normal human beings.

Acknowledgments

I would like to thank Prof. Dr. J. Th. M. van Eijk and Prof. Dr. L. M. Bouter for their comments on earlier versions of this paper. Had there not been stimulating discussions about this topic in the Research Committee of the Dutch College of General Practitioners I would not have written this paper. Although the points of view differ I am most grateful to the committee members for their (unconscious) contributions to my way of thinking.

References

1 Arrol B, Sheps SB, Schechter MT. The assessment of diagnostic tests. A comparison of medical literature in 1982 and 1985. J Gen Intern Med 1988; 3: 443-7.
2 Van Duijn NP. De likelihood ratio en de unlikelihood ratio; twee praktische maten voor diagnostische tests. Huisarts Wet 1989; 32: 478-82.
3 Dinant GJ, Knottnerus JA, Van Wersch JWJ. Het onderscheidend vermogen van de BSE-bepaling in de dagelijkse praktijk. Huisarts Wet 1992; 35: 197-203.
4 Van der Schouw YT, Verbeek ALM, Ruijs JHJ. De ROC-curve voor de eerste evaluatie van nieuwe diagnostische tests. Huisarts Wet 1992; 35: 204-8.
5 Van Geldorp WJ. Streptest. Huisarts Wet 1990; 33: 205-5.
6 Roodenburg P. Diagnostiek urineweginfecties. Huisarts Wet 1991; 34: 430.
7 Ziekenfondsraad. Grenzen aan de groei. Amstelveen: Ziekenfondsraad, 1991.
8 Gezondheidsraad. Medisch handelen op een tweesprong. Den Haag: Gezondheidsraad, 1991.
9 Wulff HR, Pedersen S, Rosenberg R. Filosofie van de geneeskunde; een verkenning. Amsterdam: Meulenhof, Utrecht: Bunge, 1988.
10 Berg M. The construction of medical disposals. Medical sociology and medical problem solving in clinical practice. Sociology of Health & Illness 1992; 14: 151-80.
11 Payer L. Medicine and culture. New York: Henry Holt and Company, 1988.
12 Van der Velden, HGM. Diagnose of prognose. De betekenis van de epidemiologie voor het handelen van de huisarts. Huisarts Wet 1983; 26: 125-8.
13 Dekker JH, Boeke AJP. Vaginale klachten in de huisartspraktijk. [Dissertatie]. Amsterdam: Vrije Universiteit, 1992.
14 Kraemer HC. Evaluating medical tests. Newbury Park: Sage, 1992.
15 Harris JM. The hazards of bedside Bayes. JAMA 1981; 246: 2602-5.
16 Heckerling PS, Conant RC, Tape TG, Wigton RS. Reproducibility of predictor variables from a validated clinical rule. Med Decis Making 1992; 12: 280-5.
17 Young MR, Poses RM. Can physicians be rational about diagnostic tests? Clin Lab Med 1984; 4: 25-9.
18 Tennekes H. Dan leef ik liever in onzekerheid. Bloemendaal: Aramith, 1990.
19 Ament A. Optimaal gebruik van diagnostische tests [Dissertatie]. Maastricht: Rijksuniversiteit Limburg, 1992.
20 Allman RA, Steinberg EP, Keruly JC, Dans E. Physician tolerance for uncertainty. JAMA 1985; 254 : 246-8.
21 Swets JA, Feehrer CE, Greenes RA, Bynum TE. Use of probability estimates in medical communications and decisions. Meth Inf Med 1986; 26: 35-42.
22 Kong A, Barnett GO, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. N Engl J Med 1986; 315: 740-4.
23 Bryant GD, Norman GR. Expressions of probability words and numbers. N Engl J Med 1980; 302: 411.
24 Eekhof JAH, Mol SSL, Pielage JG. Is doorgaans vaker dan dikwijls; of hoe vaak is soms? Ned Tijdschr Geneeskd 1992; 136: 41-2.
25 Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. Med Decis Making 1986; 6: 216-23.
26 Yates JA. Judgment and decision making. Englewood Cliffs: Prentice Hall, 1990.
27 Ornstein R. Multimind. A new way of looking at human behavior New York: Doubleday, 1986.
28 Evans JStBT. Bias in human reasoning. Hove: Erlbaum Publishers, 1990.
29 Riegelman RK. Minimizing medical mis-

takes. The art of medical decision making. Boston : Little Brown Company, 1991.
30 Balla JI, Elstein A, Gates P. Effects of prevalence and test diagnosticity upon clinical jugdments of probability. Meth Inf Med 1983; 22: 25-8.
31 Wallsten TS. Physician and medical student bias in evaluating diagnostic information. Med Decis Making 1981; 1: 145-64.
32 Hershey JC, Baron J. Clinical reasoning and cognitive processes. Med Decis Making 1987; 7: 203-11.
33 Eddy DM, Clanton CH. The art of diagnosis. Solving the clinicopathological exercise. N Engl J Med 1982; 306: 1263-8.
34 Snoek JW. Het denken van de neuroloog. [Dissertatie] Groningen: Rijksuniversiteit Groningen, 1989.
35 Kassirer JP, Kopelman RI. Cognitive errors in diagnosis: instantiation, classification and consequences. Am J Med 1989; 86: 433-41.
36 Kassirer JP. Diagnostic reasoning. Ann Intern Med 1989; 110: 893-900.
37 Greenland P, Mushlin AI, Griner PF. Discrepancies between knowledge and use of diagnostic studies in asymptomatic patients. J Med Educ 1979; 54: 863-9.
38 Wulff HR. Rational diagnosis and treatment. In Van Eijk J, Bakker S, Gubbels J, Wulff HR. Research in general practice. Rijswijk: Uitgeverij Gezondheidsbevordering, 1988.
39 Gerrity MS, Earp JAL, DeVellis RF. Uncertainty and professional work: perceptions of physicians in clinical practice. Am J Sociology 1992; 97: 1022-51.
40 Balla JI, Elstein AS, Christensen C. Obstacles to acceptance of clinical decision analysis. Br Med J 1989: 298: 579-82.
41 Vandenbroucke JP. De halsstarrigheid van een professie. Ned Tijdschr Geneeskd 1989; 133: 2540-2. ■

Samenvatting

In toenemende mate wordt er over diagnostische waarde van tests gepubliceerd. Zowel over laboratoriumbepalingen als over kenmerken van patiënten of anamnestische vragen die als voorspeller voor een bepaalde aandoening worden gebruikt. Zowel bij beleidmakers als wetenschappers lijkt de neiging te bestaan om een op klinisch epidemiologische principes gebaseerde diagnostische strategie ook in het individuele patiënten contact te propageren. Er zijn grote methodologische en cognitief psychologische bezwaren tegen het daadwerkelijk toepassen van het theorema van Bayes in de gewone huisartspraktijk. De klinische epidemiologie doet immers uitspraken over groepen en niet over individuele patiënten. Bovendien is er in de huisartsgeneeskunde veelal geen gouden standaard, waardoor uitspraken over testkenmerken altijd een beetje twijfelachtig blijven. Naast deze methodologische problemen zijn er cognitief psychologische: dokters kunnen eenvoudig in individuele gevallen geen goede schatting maken van priorkansen, laat staan dat ze die samen met de testkenmerken snel kunnen omzetten in voorspellende waarden. Gegevens uit de klinische epidemiologie zijn zinvol voor het maken van standaarden, voor het fundament van het medisch handelen van huisartsen. Zittend tegen over mijnheer Jansen moet en zal de dokter het toch zonder deze hulpmiddelen doen. Het ware verstandig als wetenschappers zich realiseren dat de werkelijkheid complexer is dan veelal uit hun onderzoeksgegevens blijkt.