Risk prediction in medicine and surgery: ethical and practical considerations

D. G. SEYMOUR, MD, BSc, MRCP(UK), *Senior Lecturer, University Department of Geriatric Medicine, Cardiff Royal Infirmary*
M. GREEN, MSc, PhD, *Lecturer in Statistics, University of Dundee*
F. G. VAZ, MB, MRCP, *Consultant in Geriatric Medicine, South Warwickshire Hospital, Warwick*
E. C. COLES, MB, MTech, FBCS, MFCM, *Head of Department of Medical Computing and Statistics, University of Wales College of Medicine, Cardiff*

ABSTRACT — Risk prediction is a subject of increasing clinical interest, and publications in this area are likely to have an important influence on patient care in the near future. A multiplicity of risk prediction systems, many of them computer-based, will raise a number of ethical and practical questions. These questions need to be addressed by the originators of systems, the editors of journals, practising clinicians, and the lay public.

Conventional statistics (eg chi-square, t-tests, correlation coefficients and linear regression) have been in common use in medicine for three decades. Nevertheless there is clear evidence that these relatively simple techniques are often misused or misunderstood, a fact that has caused a number of leading medical journals to re-examine their statistical requirements [1–7]. Altman [8] has made the point that poor statistical techniques are not only academically undesirable but also unethical. In clinical trials, for instance, bad statistics may result in the wrong treatment being recommended, useful therapies being overlooked, or research time being wasted [8].

If practical and ethical problems can arise with conventional statistics they become almost inevitable where the goal of the data analysis is risk prediction. Here statistical techniques and/or computer methods are complex and errors are difficult to detect. Furthermore, studies that set out explicitly to quantify risk are particularly likely to be used as a reason for giving or withholding therapy in individuals. If published predictive equations are not valid, and there is reason to believe that many of them may not be [9], patients are likely to suffer. The ethical problems are not confined to patients with a clinical illness, as epidemiological studies may raise ethical questions about the modification of risk factors in the general population [10].

The ethical aspects of risk prediction have attracted surprisingly little comment in the literature, although a recent article on the ethics of computer-assisted diagnosis has done much to rectify this [11]. The link between 'statistical fact and ethical imperative' has also been raised in the field of obstetrics by Silver and Minogue [12]. The present article discusses the ethical and clinical implications of risk prediction studies, and suggests a number of questions that should be asked when such studies are published. These questions have been stimulated by our interest in post-operative outcome in the elderly surgical patient [13–15], and by our recent attempts to predict this outcome using the techniques of Spiegelhalter and Knill-Jones [16,17]. There may be a reluctance to operate on elderly surgical patients on the grounds of age alone, so a predictive equation that erroneously identified an individual as being at 'high risk' might well result in necessary surgery being withheld. The other stimulus for the present paper has been the pioneering diagnostic and prognostic work carried out over the past 15 years on the clinical problems of acute abdominal pain [18], gastrointestinal disease [19,20], head injury [21], post-operative deep venous thrombosis [22–24], post-operative myocardial events [25], and survival patterns in intensive care [26,27].

Questions of interest to the clinician when considering a system of prediction

1. Were the computational techniques appropriate and were they applied accurately?

If we extrapolate from the work of Altman [8], it is unethical to publish a predictive equation (or propose the use of a computer system) that has been derived by an inappropriate method. However, a major problem encountered in risk prediction is that what constitutes 'appropriate' analysis is often a matter of dispute. There may be fundamental philosophical differences between those who advocate 'probabilistic' or 'statistical' methods and those who champion the 'knowledge-based' approach of artificial intelligence. Even

Address for correspondence: Dr D. G. Seymour, University Department of Geriatric Medicine, Cardiff Royal Infirmary (West Wing), Newport Road, Cardiff CF2 1SZ.

within the 'probabilistic' school there tend to be subdivisions into those using Bayes's theorem and those who advocate 'non-Bayesian' techniques such as regression analysis or discriminant function analysis. Spiegelhalter and Knill-Jones [20] have recently suggested that it is possible to combine useful elements of many of these techniques, and developments of this type have much to offer clinical medicine in the near future [16]. Looking at present techniques, Wasson *et al.* [9] have suggested some broad methodological standards for studies involving clinical prediction rules, and it is to be hoped that individual medical journals will introduce their own standards in the future, just as they have done recently for conventional statistics.

2. Has there been close medical involvement at all stages of the statistical analysis?

It is now well recognised that clinicians and statisticians should co-operate closely at all stages of a clinical study, particularly at the time when variables are being selected for inclusion in a predictive equation. It is also becoming increasingly clear that simply entering a large number of clinical variables into a multivariate analysis, and relying solely on a statistical technique (such as stepwise selection) to choose a 'good' predictive equation, may be much less successful than an approach that also takes into account clinical knowledge [28]. In the highly respected Glasgow study of the effects of head injury, the use of clinical knowledge to group together several risk factors *prior* to analysis resulted in a better predictor than would have been otherwise obtained [21]. The clinician may also need to exercise clinical judgement in relation to other aspects of variable selection, for example by advising against the inclusion of variables that are statistically useful but require unacceptably invasive procedures. Since the ultimate aim of a predictive system is to improve patient care, the clinician may also favour variables that are potentially modifiable by therapy over those that are not.

3. Is the predictive equation likely to prove medically acceptable to clinicians?

The answer to this question depends mainly on the number and nature of variables chosen for inclusion in the predictive equation and is therefore related to question 2. In day-to-day clinical practice, decision making is often based on a small number of key clinical findings. Predictive equations that mirror this by containing a small number (say five or less) of key variables are more likely to be accepted than those that are more complex. Mathematical considerations also favour keeping the number of predictor variables small, since equations containing large numbers of variables often prove to be 'overfitted' and to perform poorly in new data sets [29]. Overfitting is particularly likely when the original data sets are small, such that random variations tend to distort any underlying patterns. Here one useful rule of thumb is that the original data set should contain at least five individuals with the chosen outcome for every variable in the equation; for example, if the predictor uses three variables to predict death, there should have been at least 15 patient deaths (and 15 survivors) in the original data set [9].

Another factor that will determine whether a predictor is likely to be acceptable to clinicians is whether or not the variables included in it concur with current medical theories about aetiology (ie whether they make 'medical sense'). Here again the role of the clinician working with the statistician is of great importance. When variables make both statistical and clinical sense, the ethical problems of applying a predictive equation are likely to be reduced. However, dialogue is vital; there will be occasions where it is the 'accepted medical opinion' that is in error.

After the process of variable selection has been completed, clinical judgement remains important. For instance, the clinician may feel ethically compelled to 'redraw' the line of demarcation chosen by the statistical analysis, judging that it is preferable to include a few more misclassified cases rather than miss cases of treatable disease [21]. Alternatively, when a predictive equation fails to perform as well as was expected, the clinician may be able to offer pathophysiological reasons and suggest alternative courses of action. For example, in elderly surgical patients it has been possible to achieve accurate prediction of post-operative respiratory complications by applying multivariate methods to a small number of easily obtained pre-operative variables [16,17]. However, prediction of post-operative cardiac failure has proved much more difficult, probably because several of the major risk factors (such as pre-operative ventricular function and the severity of coronary artery narrowing) are difficult to estimate by simple clinical means. In such circumstances the clinician might suggest a refinement of the clinical method prior to analysis. An alternative would be to abandon the statistical approach altogether, using instead invasive methods to monitor cardiovascular function throughout the operative period [30,31]. Yet another possibility would be to combine high-technology and multivariate statistical methods. Problems such as this require close co-operation between statistician and clinician if they are to be solved.

In the field of artificial intelligence, and particularly in the area of 'expert systems', the diagnostic or predictive system may initially be founded entirely on clinical opinion, although a mixture of clinical opinion and hard data is preferable from the outset [20], and essential where systems are likely to influence patient care.

4. Has the predictive equation been shown to be valid in new sets of patients?

This is the crucial question for the clinician and patient alike. The concept of 'training' and 'test' data sets is well recognised in statistical circles [9, 16, 21, 28, 32]. The 'training' data set is that set of patients from which the predictive equation (or predictive system) was derived. In the 'training' set, the variables likely to be useful in prediction, and the outcome that it is desired to predict, are all known. It will readily be seen that the predictive equation, designed as it is to fit the individuals in the 'training' data set, is highly unlikely to fit another data set quite as well. The acid test of a predictive equation is therefore not the quality of prediction that it achieves in the 'training' data set but its predictive ability in another, 'test' data set of patients.

Ideally, more than one 'test' data set should be used, perhaps starting with patients attending the same institution as patients in the 'training' data set, and then using a data set from another centre [9]. Even though patients from another centre may be chosen to be broadly similar to the original data set (eg patients over 65 presenting to general surgeons), population characteristics are likely to differ from centre to centre, providing a sterner test of the general validity of a predictive system. It is perhaps significant that the risk prediction methods that have been adopted most widely [18, 21] have all been repeatedly validated in a variety of test data sets. A predictive equation based on a 'training' data set alone is at best suspect, and it may need to be withdrawn subsequently. Yet, in their review of three major American journals and the *British Medical Journal*, Wasson *et al.* [9] found that two-thirds of predictive studies failed to validate their predictions in a 'test' data set.

The cardiac risk index of Goldman *et al.* [25] has been in clinical use for more than a decade, and has stimulated considerable interest in the potential for using numerical methods in day-to-day clinical diagnosis. However, doubts about the general validity of this index are still being raised, primarily because the first attempts to validate it in a test data set did not appear until at least 6 years later, despite the comments by the originators of the index [25] that validation was necessary before the index entered general clinical medicine. Recently, Detsky and his colleagues [33, 34] have proposed a modified version, but this in turn has been criticised [35].

As Wasson *et al.* [9] point out, in an ideal world predictive equations would not just be validated on a test data set but would also be shown to produce clinical benefits when applied. However, this was attempted in only 6% of the reports they reviewed.

The separate collection of training and test data sets is time-consuming, and ingenious methods have been devised to extract as much information as possible from a single data set. Examples of 'jack-knife' (one left out) and 'bootstrap' methods are given by Wasson *et al.* [9]. However, since these methods will not eliminate the effects of biases in patient selection or data collection [9], a complete separation of training and test data sets is preferable, particularly where direct effects on patient care are likely to result from the predictive equation.

5. Is it ethical to publish a predictive equation that has not been validated in a new (test) data set?

The principle that a predictive equation should be validated in a test data set before being put on 'general release' is now so widely accepted in statistical circles that it may now be *unethical* to propose a risk prediction system for clinical use until validation has been satisfactorily carried out. To safeguard patients from the risks of non-validated equations, it might be preferable if journals were to refuse to regard papers containing non-validated risk indices as completed work. Perhaps they should be listed instead as preliminary communications, to be subsequently upheld or retracted when the results on a test data set became available. While this might delay initial publication, the risk of patients coming to harm from the premature release of a non-valid equatioıı ;vould be greatly reduced.

If validation is to be demanded of predictive equations of a 'probabilistic' or 'statistical' nature, it seems only logical that 'knowledge-based' expert systems should similarly be required to prove their worth before they are adopted into general clinical practice (see discussion following paper of Spiegelhalter and Knill-Jones [20]).

6. How should the clinician use information obtained from a predictive equation or a computer?

This is one of the major questions considered by de Dombal [11] in his review of the ethical aspects of computers in medicine. Assuming that a predictive system has passed the barriers referred to above, how much weight should a clinician give to a single prediction in an individual patient? With a newly proposed system it is probably best to regard the prediction obtained as 'just another test' to be weighed against more conventional pieces of clinical information [11]. Later it may be possible to put more 'trust' in the results of the system, but de Dombal [11] argues persuasively that the final decision should always lie with the doctor who can bring to bear concepts such as 'error' and 'harm' which are unknown to most computer programmes. The doctor can also allow for circumstances that lie outside the 'range of experience' of the computer system. The legal position of a doctor who overrides a computer system that subsequently turns out to be correct is also considered by de Dombal.

A difficult problem may arise when a newly introduced system pronounces a patient to be at 'high risk'. This is especially important in surgery where one of the therapeutic options is to abandon potentially useful surgery if the operative risk is thought to be too high. An illustrative approach to this problem was adopted by Babu *et al.* [30]. In their study of elderly patients undergoing vascular surgery, Babu and his colleagues used invasive monitoring rather than predictive equations to identify high-risk patients. However, they approached their study with the aim of offering surgery to every patient if at all possible. A patient assessed as being at high risk was therefore given vigorous medical therapy to try to reduce that risk and was monitored very closely. Using this approach, surgery proved possible in 74 out of 75 cases. The approach of Babu *et al.* [30] has much to commend it, and we suggest that the usual clinical response when a *newly introduced* predictive system forecasts 'high risk' in an individual should be an attempt to reduce the reversible elements of that risk, rather than to exclude the patient from potentially useful therapy on the grounds of the predictive equation alone. Later, when a system has become more thoroughly validated, the use of the predictive equation to exclude patients from therapy is easier to justify on an ethical basis.

7. *Is the predictive equation clearly presented and does it mean what it says?*

Even when a predictive equation has been thoroughly validated, the way in which it is presented to the general reader has a major bearing on whether or not it will be adopted clinically. Many systems are offered to the potential user in the form of a complex formula. This tends to have one of two unfortunate effects on the non-mathematical reader. The most likely effect is that the predictive system is ignored, and this might explain why so few of the systems have entered day-to-day clinical practice [20]. The opposite response is for the non-mathematician to equate complexity with truth and to accept a system without questioning or understanding. The need to present a prediction system with clarity therefore becomes a matter of ethics as well as one of aesthetics. Poor presentation may cause patients to suffer because a useful method of prediction or diagnosis is ignored, or a dubious system is embraced.

Imaginative schemes to make systems of differential diagnosis or clinical prediction more acceptable to clinicians have been developed and deserve wider use. The de Dombal system for diagnosing the 'acute abdomen' has proved highly acceptable using a microcomputer [36], and this approach also reduces the risk of computational error. The importance of using pictorial representation and explaining the reasoning behind a prediction is also well recognised [20]. Finally, as Spiegelhalter [37] puts it, if the prediction is expressed as a probability it should mean what it says. Thus, if the system states that the probability of disease in an individual is 90%, then nine times out of ten a patient with that individual's characteristics should get that disease.

A predictive system or equation, like any other measurement instrument in medicine, will not give perfect results. To aid the doctor who might be considering using the system, the likely range of error arising from imperfections in the system itself should be stated if at all possible. There would be merit in presenting the final prediction along with a confidence interval, as the general medical reader is becoming increasingly familiar with this style of data representation.

8. *Who should update a predictive system?*

Even in the most well validated systems, probabilities are likely to alter with time. This implies that predictive equations should be re-assessed from time to time and, as in the case of the initial validation [9], the researchers best placed to perform this updating are probably those who originated the system. This principle has been well established by de Dombal and his group in Leeds who have been extending and refining their system for more than 15 years.

One of the attractions of some of the newer computer software is that it has the potential to update itself by 'learning from' the new cases that it processes.

9. *How should predictive systems affect the doctor–patient relationship?*

Few would deny that the patient should be closely involved in decision-making, and it seems to follow that the doctor should share with the patient any additional information obtained from a prediction system. While there may be practical difficulties in explaining degrees of risk to individual patients [38], the underlying ethical principle nevertheless appears clear. de Dombal [18] has pointed out that, with the development of information science and computers in medicine, the traditional skills of taking a history and performing an accurate and relevant examination have become more rather than less important, since such symptoms and signs have to be rigorously defined before they can be coded or computerised. There appears to be no reason why the doctor–patient relationship or the 'autonomy' of patient and doctor should be eroded by a well validated system of clinical risk prediction [11].

Conclusion

The potential clinical benefits of systems of risk prediction are considerable, yet such systems have been little used to date. As systems become more accurate, convenient and acceptable to clinicians, a number of ethical problems will arise. Such problems are the legitimate concern not only of the research workers who devise the systems but also of the journals which publish them, the clinicians who are their potential users, and the patients who stand to benefit from their use.

Acknowledgements

This work was supported in part by a grant from Research into Ageing (formerly the British Foundation for Age Research). We are also grateful to our colleague Dr Robert Newcombe for his helpful comments on the manuscript.

References

1. Altman, D. G., Gore, S. M., Gardner, M. J. and Pocock, S. J. (1983) Statistical guidelines for contributors to medical journals *British Medical Journal*, **286**, 1489–93.

2. Gardner, M. J., Altman, D. G., Jones, D. R. and Machin, D. (1983). Is the statistical assessment of papers submitted to the British Medical Journal effective? *British Medical Journal*, **286**, 1485–8.
3. Bailar, J. C. (1986) Science, statistics and deception. *Annals of Internal Medicine*, **104**, 259–60
4. Gardner, M. J. and Altman, D. G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*, **292**, 746–50.
5. Bulpitt, C. J. (1987) Confidence intervals. *Lancet*, **i**, 494–7.
6. Little, J. M. (1987) Presenting statistics. *Australia and New Zealand Journal of Surgery*, **57**, 417–21.
7. Pocock, S. J., Hughes, M. D. and Lee R. J. (1987) Statistical problems in the reporting of clinical trials: a survey of three medical journals. *New England Journal of Medicine*, **317**, 426–32.
8. Altman, D. G. (1982) Statistics and ethics in medical research. In *Statistics in practice* (ed. S. M. Gore and D. G. Altman) pp 1–24. London: British Medical Association.
9. Wasson, J. H., Sox, H. C., Neff, R. K. and Goldman, L. (1985) Clinical prediction rules: applications and methodological standards. *New England Journal of Medicine*, **313**, 793–9.
10. Brett, A. S. (1984) Ethical issues in risk factor intervention. *American Journal of Medicine*, **76**, 557–61.
11. de Dombal, F. T. (1987) Ethical considerations concerning computers in medicine in the 1980s. *Journal of Medical Ethics*, **13**, 179–84.
12. Silver, R. K. and Minogue, J. (1987) When does a statistical fact become an ethical imperative? *American Journal of Obstetrics and Gynecology*, **157**, 229–33.
13. Seymour, D. G. and Pringle, R. (1983) Post-operative complications in the elderly surgical patient. *Gerontology*, **29**, 262-70.
14. Vaz, F. G and Seymour, D. G. (1989) A prospective study of elderly general surgical patients. I. Pre-operative medical problems. *Age and Ageing*, **18**, 309–15.
15. Seymour, D. G. and Vaz, F. G. (1989) A prospective study of elderly general surgical patients II. Post-operative complications. *Age and Ageing*, **18**, 316–26.
16. Seymour, D. G., Green, M. and Vaz, F. G. (1990) Making better decisions: the construction of clinical scoring systems by the Spiegelhalter and Knill-Jones approach. *British Medical Journal*, **300**, 223–6.
17. Seymour, D. G. (1988) Prediction of risk in the elderly surgical patient. MD thesis, University of Birmingham.
18. de Dombal, F. T. (1985) Analysis of symptoms in the acute abdomen. *Clinics in Gastroenterology*, **14**, 531–43.
19. Knill-Jones, R. P. (1985) A formal approach to symptoms in dyspepsia. *Clinics in Gastroenterology*, **14**, 517–29.
20. Spiegelhalter, D. J. and Knill-Jones, R. P. (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society*, *A***147**, 35–77.
21. Titterington, D. M., Murray, G. D., Murray, L. S. *et al.* (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society*, *A***144**, 145–75.
22. Clayton, J. K., Anderson, J. A. and McNicol, G. P. (1976) Preoperative prediction of postoperative deep vein thrombosis. *British Medical Journal*, **2**, 910–2.
23. Crandon, A. J., Peel, K. R., Anderson, J. A. *et al.* (1980) Postoperative deep vein thrombosis: identifying high-risk patients. *British Medical Journal*, **281**, 343–4.
24. Crandon, A. J., Peel, K. R., Anderson, J. A. *et al.* (1980) Prophylaxis of postoperative deep vein thrombosis: elective use of low-dose heparin in high-risk patients. *British Medical Journal*, **281**, 345–7.
25. Goldman, L., Caldera, D. L., Nussbaum, S. R. *et al.* (1977) Multifactorial index of cardiac risk in noncardiac surgical procedures. *New England Journal of Medicine*, **297**, 845–50.
26. Shoemaker, W. C. (1987) Physiology, monitoring and therapy of critically ill general surgical patients. In *Diagnostic methods in critical care: automated data collection and interpretation* (ed. W. C. Shoemaker and E. Abraham) pp 47–86. New York: Marcel Dekker.
27. Chang, R. W. S., Jacobs, S. and Lee, B. (1988) Predicting outcome among intensive care unit patients using computerised trend analysis of daily APACHE II scores corrected for organ system failure. *Intensive Care Medicine*, **14**, 558–66.
28. Spiegelhalter, D. J. (1986) Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, **5**, 421–33.
29. McCullagh, P. and Nelder, J. A. (1983) *Generalised linear models*. London: Chapman and Hall.
30. Babu, S. C., Sharma, P. V. P., Raciti, A. *et al.* (1980) Monitor-guided responses: operability with safety is increased in patients with peripheral vascular diseases. *Archives of Surgery*, **115**, 1384–6.
31. Del Guercio, L. R. M. and Cohn, J. D. (1980) Monitoring operative risk in the elderly. *Journal of the American Medical Association*, **243**, 1350–5.
32. Copas, J. B. (1983) Regression, prediction and shrinkage. *Journal of the Royal Statistical Society*, *B***45**, 321–54.
33. Detsky, A. S., Abrams, H. B., McLaughlin, J. R. *et al.* (1986) Predicting cardiac complications in patients undergoing non-cardiac surgery. *Journal of General Internal Medicine*, **1**, 211–9.
34. Detsky, A. S., Abrams, H. B., Forbath, N. *et al* (1986) Cardiac assessment for patients undergoing noncardiac surgery: a multifactorial clinical risk index. *Archives of Internal Medicine*, **146**, 2131–4.
35. Hochman, H. and Lumb, P. (1987) Cardiac risk in noncardiac surgical procedures (Letter and authors' reply). *Archives of Internal Medicine*, **147**, 1001–4.
36. Adams, I. D., Chan, M., Clifford, P. C. *et al.* (1986) Computer aided diagnosis of acute abdominal pain: a multi-centre study. *British Medical Journal*, **293**, 800–4.
37. Spiegelhalter, D. J. (1985) Statistical methodology for evaluating gastrointestinal symptoms. *Clinics in Gastroenterology*, **14**, 489–515.
38. O'Brien, B. (1986) *What are my chances, doctor? A review of clinical risks*. London: Office of Health Economics.