Journal of Epidemiology and Community Health 1992; **46**: 323–326

REVIEW ARTICLE

Evaluating prognostic factors: implications for measurement of health care outcome

Martin C Gulliford

Measurement of health care outcomes has become the contemporary philosophers' stone for public health researchers, managers, and politicians alike. Each group recognises that the rational use of scarce resources depends on measuring the effectiveness of different procedures and programmes of care.[1] This point was clearly made by Cochrane in 1972[2] and there has since been increasing interest in developing indicators of health care outcome.[3] Donabedian defined (health care) outcome as "those changes either favourable or adverse in the actual or potential health status of persons, groups or communities *that can be attributed to prior or concurrent care*".[4] Attributing changes in health status to prior or concurrent care is no simple matter and there is continuing uncertainty concerning how this can and cannot be done.[5]

Outcome of disease or outcome of health care?

Health services researchers, following Donabedian, commonly use the term "outcome" as though it is synonymous with "health care outcome"; this is confusing. Changes in health status can be considered more generally as measures of the outcome of disease processes and other influences on health.[6] Health care is only one of the factors which determines the outcome of disease. Age, gender, ethnicity, psychological factors, the social and physical environment, and the nature of underlying and associated conditions also combine to influence the prognosis. A clear distinction should therefore be maintained between the general term "outcome" and the specific term "health care outcome".

It should not be very difficult to suggest changes in health status which might act as indicators of outcome for a given condition. For example, in diabetes the rates of admission to hospital with acute complications such as hyperglycaemia and hypoglycaemia, or the incidence of chronic complications such as chronic renal failure and limb amputation, have been suggested as suitable measures of outcome.[7] A recent review of the measurement of health status provides additional guidance.[8] The major difficulty in describing health status changes as health care outcomes lies in attributing outcomes experienced by patients to prior health care intervention. Obtaining "convincing evidence that health status changes are in fact due to health service interventions is an essential objective of outcome analysis".[3]

Department of Public Health Medicine, United Medical and Dental Schools, St Thomas' Campus, London SE1 7EH, United Kingdom
M C Gulliford

Accepted for publication January 1992

Use of aggregate data

There may be a small number of health care interventions which are so effective that the observed variation in outcome among populations is mainly explained by use of the intervention. Given high standards of medical practice there should be few if any deaths from a small number of conditions which are amenable to health care intervention.[9] Mortality from these conditions can be considered as measures of health care outcome.[10] Although epidemiological studies have yet to demonstrate the validity of "avoidable" death indicators as measures of health care outcome,[11] it may be argued that even if social and environmental factors contribute to variations in the incidence and case fatality of these conditions, effective health services should be able to limit the impact of these influences on area mortality rates.

More often the proportion of variation in outcome explained by health care intervention is rather small when compared with random and systematic variation in the natural history of diseases.[12] Variations in case mix are likely to be even more important in evaluating variations in outcome among health care providers than among geographical areas within countries. As well as being the most important influence on outcome, case severity is of course the most important determinant of the type of health care patients receive. Thus before health care outcomes can be assessed it will usually be necessary to have good measurements of the confounding influences of case severity. Outcomes which can be monitored using aggregate data can rarely be shown directly to be health care outcomes.[13]

Experimental and observational studies

Random allocation is the only certain way of ensuring that there are no systematic differences in the prognostic characteristics of patient groups.[14] Differences in outcome among treatment groups in randomised trials are measures of health care outcome. There are many practical limitations to the use of randomised trials, however, and randomised trials can rarely be used to investigate the quality or effectiveness of routine health care for which there is greatest need of evaluation.[1]

Thus observational study designs, utilising routinely or specially collected data, will inevitably be used to evaluate influences on the outcomes experienced by particular populations

or patient groups.[15] Several recent observational studies which examined the outcome of common and important health care problems have been the subject of considerable debate. The problems investigated included: the effectiveness of alternative therapies in the management of breast cancer[16]; use of the bronchodilator fenoterol and death from asthma[17]; mortality following transurethral resection of the prostate[18]; and use of hospital mortality rates to measure the outcome of health care at hospitals treating Medicare patients.[19] The cause of controversy was an attempt to explain outcome in terms of two groups of variables: health care and case severity. The reported associations of health care characteristics with outcome measures might be accounted for either by incomplete adjustment for confounding of treatment allocation with case severity or in terms of the quality or effectiveness of care.

Definition of severity

Severity is not a concept which is defined in standard epidemiological texts. One definition which has been used is: "severity is defined as risk of death or temporary or permanent impairment".[20] Severity measures should also be characteristics of patients; characteristics of the structure or process of health care cannot be considered as measures of severity even though they may be prognostic factors. Since the risk of an adverse outcome may be modified by health care, severity must be measured before the start of health care intervention.[21] Severity and outcome may both be measured on the same scale. For example, the severity of disability resulting from multiple sclerosis may be measured on the Kurtzke scale and the outcome after intervention measured in terms of the change in position on the scale.[22] Non-fatal outcomes occurring during the course of treatment may be prognostic factors for later outcome events but they cannot be accepted as measures of severity because they may be influenced by prior health care. Similarly, failure to respond to treatment may be associated with an unfavourable outcome but cannot be considered as a measure of severity. In fact separation of prognostic groups according to response to treatment is one of the sources of misleading claims concerning the effectiveness of treatments.[23]

Severity is thus defined by collecting information concerning important prognostic variables which are characteristics of patients measured before the start of health care invervention.[21] If measures of severity are to be used to adjust observed outcomes for case mix it is important that they provide valid predictions of the expected outcome for groups of patients with given case severity.[24]

If a large part of the variation in outcome is explained by few variables then characterisation of severity may be easy. For example the main influences on the outcome of burns patients are the age of the patient and the extent of the burn area.[25] More often many different variables are associated with prognosis, and data collection must be considered for all of them. It cannot be assumed that all important prognostic factors are known. Even if they are, methods of measurement and classification may not be standardised and the status of the variable in an individual may change over time. Thus incomplete ascertainment and misclassification of confounding variables represent the major problems in attempting to characterise severity.

Selection of variables

The values of patient scores on confounding variables are usually combined in a prognostic model. Two problems exist in identifying variables to be included in a prognostic model. First, the factors which influence the prognosis may be incompletely known. Secondly, some selection must usually be made from the many variables which are known to influence the prognosis.

The development of prognostic indices and severity measures has been largely applied to acute illnesses or conditions, such as cancers, in which the clinical course tends to be relatively constant. Many of the influences on health status changes in chronic diseases have yet to be identified. The factors which confer susceptibility to the long term complications of diabetes are largely unknown[26] and there is relatively little evidence to support the attribution of these measures of outcome to prior health care.[27] Thus for common conditions which impose great burdens on health services there is at present only limited understanding of the factors which influence the prognosis and correspondingly little possibility of attributing outcomes either to health care or to case severity.

Some prognostic models have been compiled on the basis of investigators' subjective opinions of the variables which should be included, sometimes without subsequent validation of the index.[28] Multivariable statistical methods are now more commonly used to identify the variables which are independent predictors of outcome.[21] Although automatic variable selection procedures are commonly used to eliminate variables which do not make a significant contribution to the prediction of outcome, it has been argued that such procedures should be used infrequently unless the candidate variables are of no intrinsic interest and this is unlikely to be the case in outcome studies.[29]

Prognostic models may perform excessively sensitively for the data set for which they were developed. Particularly for large data sets, variables may be inappropriately selected on the basis of type I errors. Avoidance of overfitting can be achieved by cross validation of the model; the model should be developed from one random half of the data and validated against the second half.[30 31] External validation is also required when the model is applied to an entirely new set of data, since the importance of confounding variables may differ among populations.[31] For non-condition-specific indices, the validity of the index may vary according to the nature of the condition being examined.[32] Harrell *et al* have emphasised that predictive discrimination ("the ability to separate patients with good and poor outcomes") is the appropriate measure of the validity of a prognostic model.[33] Variables should be included not only on the grounds of statistical significance (whether any contribution is made to prediction) but also on the amount of prognostic information which the variable contributes to the model.[34]

Data quality and misclassification

Many prognostic indices have been developed in the context of particular cohort studies or from case series collected in particular institutions. Such indices may lack general utility, either because the data requirements are excessive for routine use or because the methods of data collection and the definition of variables are unique to the local circumstances. Variables selected for measurement and the methods of measurement should be standardised between studies where possible.[35] The development of prognostic indices with particular attention to the succinct specification of the variables to be included and to the operational definition of the variables has facilitated their more widespread acceptance. Examples include the Glasgow coma scale[36] and the Apgar score.[37] Such simplification is likely to be associated with incomplete adjustment for confounding.

Patients' clinical records remain the major source of information concerning their health care, but these suffer from a number of limitations. Inconsistent and incomplete recording of important items of data can usually be expected when clinical records are used as a data source.[38] There may be little relation between the quality of data and the quality of care.[39] Clinical findings are mainly used as aids to decision making. The aim is not to collect data which will provide the most precise characterisation of the prognosis. Clinicians also use non-quantifiable impressions based on experience to determine what treatment patients receive; the implicit basis of clinical decision making may not be made explicit in clinical records. It has been shown that clinicians' subjective opinions of the severity of the patients condition are powerful prognostic factors in their own right.[40 41] For this reason it would not be surprising to find that adjustment for severity, based on information obtained from clinical records, was incomplete.

When relevant items of data are recorded they may be subject to misclassification. Random misclassification of confounding variables results in a reduced ability to control for confounding.[42] If the sensitivity and specificity of the classification of the confounder are known, then the effect of random misclassification may be evaluated in simple cases. Where there are several confounding variables, the consequences of misclassification may be difficult to predict.[43] Misclassification is often biased. For example, systematic understaging of bladder neoplasms is a well recognised problem.[44] The extent to which patients are investigated varies among institutions, and this can also contribute to differential misclassification.[45] Thus misclassification in the measurement of confounding variables is likely to contribute to the problem of residual confounding and can also result in biased adjustment for confounding.

Risk adjusted measures of outcome

Prognostic models have a number of limitations: data requirements may be excessive for routine use; measurement and recording of important variables may not be standardised; many prognostic variables are not known; the prognostic factors which are important may vary among populations and over time in the same population. Despite these limitations the use of prognostic models to adjust for case severity provides an additional insight into the association of outcome with case severity and health care. Differences between the observed outcomes in a group of patients and those predicted by the prognostic model have been termed "risk adjusted monitors of outcome".[21] When observational data are used to evaluate health care, risk adjusted measures of outcome provide the nearest available approximation to measures of health care outcome but interpretation will always require consideration of case mix as well as health care intervention. It is never possible to be entirely sure that complete and unbiased adjustment has been made for confounding.

In the United Kingdom few routine data allow the development of risk adjusted measures of outcome.[46] In specific studies, risk adjusted measures of outcome have been used to monitor the introduction of new technologies,[47 48] to monitor trends in outcome over time,[49] and to examine contemporary variations in outcome among providers.[50] For example, McArdle and Hole[50] examined variation in practice among 13 surgeons performing surgery for colorectal carcinoma. They found that there were systematic differences among surgeons in rates of postoperative complications, postoperative mortality, and long term survival which were not explained by adjusting for severity. It was reasonable to conclude that some surgeons were more effective than others in treating this disease and to recommend that surgery be concentrated in the hands of specialists. If this study had shown that variation in outcome among surgeons was mainly explained by differences in case severity, then this would have given greater emphasis to the need for more effective treatments for groups of patients with particular prognostic characteristics. The unadjusted association between surgeon and outcome identified a problem; use of adjustment for severity provided an additional insight into the possible causes and the type of intervention which might be helpful in resolving the problem. A randomised trial would be required to provide unequivocal evidence of the effectiveness of the recommended solution.

Conclusions

A number of different changes in health status can be considered as outcomes of disease. Health care is only one of the factors which influence the outcome of disease. A clear distinction should therefore be maintained between the general term "outcome" and the specific term "health care outcome". Unequivocal measurements of health care outcome can only be made in randomised trials. In observational studies the baseline characteristics of patients which characterise case severity must also be taken into account. Incomplete and inaccurate ascertainment of confounding variables limits the extent to which severity can be evaluated. Although it is never possible to be sure that confounding by differences in case mix has been fully evaluated, the estimation of risk adjusted measures of outcome

provides the closest approximation to measurement of health care outcomes that can be obtained using observational data. Routine information systems should record measures of outcome for different conditions together with simple prognostic information relating both to characteristics of patients and to health care intervention. Researchers should aim to identify and evaluate all of the factors which influence the prognosis; such information is needed particularly for conditions with a chronic clinical course.

I thank Dr Peter Burney and Professor Walter W Holland for their comments on earlier drafts of this paper.

1 House of Commons Public Accounts Committee (Tenth report: session 1988–9). *Quality of clinical care in National Health Service Hospitals*. London: HMSO, 1989.
2 Cochrane Al. *Effectiveness and efficiency. Random reflections on health services*. London: Nuffield Provincial Hospitals Trust, 1972.
3 Opit L. The measurement of health service outcomes. In: Holland WW, Detels R, Knox G, eds. *Oxford textbook of public health*. Oxford: Oxford University Press, 1991: 159–72.
4 Donabedian A. *Explorations in quality assessment and monitoring. Vol 3. The methods and findings of quality assessment and monitoring: an illustrated analysis*. Ann Arbor: Health Administration Press, 1985: 256.
5 Harley M. The measurement of health services activity and standards. In: Holland WW, Detels R, Knox G, eds. *Oxford textbook of public health*. Oxford: Oxford University Press, 1991: 145–57.
6 Fletcher RH, Fletcher SW, Wagner EH. *Clinical epidemiology, the essentials*. Second edition. Baltimore: Williams and Wilkins, 1988: 110.
7 Williams DRR. Outcome indicators for diabetes services—what do we have and what do we need? *Community Med* 1989; **11**: 57–64.
8 Bowling A. Measuring health. *A review of quality of life measurement scales*. Milton Keynes: Open University Press, 1991.
9 Rutstein DD, Berenberg W, Chalmers TC, Child CG, Fishman AP, Perrin EB. Measuring the quality of medical care: a clinical method. *N Engl J Med* 1976; **294**: 582–8.
10 Charlton JRH, Hartley RM, Silver R, Holland WW. Geographical variation in mortality from conditions amenable to medical intervention in England and Wales. *Lancet* 1983; i: 691–6.
11 Carr-Hill R, Hardman GF, Russell IT. Variation in avoidable mortality and variations in health care resources. *Lancet* 1987; i: 789–91.
12 Simon R. Importance ofprognostic factors in cancer clinical trials. *Cancer Treat Rep* 1984; **68**: 185–92.
13 Mackenbach JP, Bouvier-Colle MH, Jougla E. "Avoidable" mortality and health services: a review of aggregate data studies. *J Epidemiol Community Health* 1990; **44**: 106–11.
14 Hill AB. *Statistical methods in clinical and preventive medicine*. Edinburgh: Livingstone, 1962.
15 Greenfield S. The state of outcomes research: are we on target? *N Engl J Med* 1989; **320**: 1142–3.
16 Bagenal FS, Easton DF, Harris E, Chilvers CED, McElwain TJ. Survival of patients with breast cancer attending Bristol Cancer Help Centre. *Lancet* 1990; ii: 606–10.
17 Crane J, Pearce N, Flatt A, *et al*. Prescribed fenoterol and death from asthma in New Zealand 1981–3: case-control study. *Lancet* 1989; i: 917–22.
18 Roos NP, Wennberg JE, Malenka DJ, *et al*. Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *N Engl J Med* 1989; **320**: 1120–4.
19 Park RE, Brook RH, Kosecoff J, *et al*. Explaining variations in hospital death rates: randomness, severity of illness, quality of care. *JAMA* 1990; **317**: 1674–80.
20 Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease. A case mix measurement. *JAMA* 1984; **251**: 637–44.
21 Blumberg MS. Risk adjusting health care outcomes: a methodologic review. *Med Care Rev* 1986; **43**: 351–93.
22 Barnes MP, Bates D, Cartylidge NEF, French JM, Shaw DA. Hyperbaric oxygen and multiple sclerosis: short term results of a placebo controlled double blind trial. *Lancet* 1985; i: 297–300.
23 Raghavan D. Chemotherapy for advanced bladder cancer: "Midsummer night's dream" or "Much ado about nothing"? *Br J Cancer* 1990; **62**: 337–40.
24 Gibson G. Indices of severity for emergency medical evaluative studies: reliability, validity and data requirements. *Int J Health Serv* 1981; **11**: 597–622.
25 Feller I, Tholen D, Cornell RG. Improvements in burn care, 1965–1979. *JAMA* 1980; **244**: 2074–8.
26 Borch-Johnsen K, Nissen H, Salling N, *et al*. The natural history of insulin-dependent diabetes in Denmark: 2. Long term survival—who and why? *Diabetic Med* 1987; **4**: 211–6.
27 Stephenson J, Fuller JH. Health of the nation. *BMJ* 1991; **303**: 1478.
28 Gibson G. Indices of severity for emergency medical evaluative studies: reliability, validity and data requirements. *Int J Health Serv* 1981; **11**: 597–622.
29 Cox DR, Snell EJ. The choice of variables in observational studies. *Appl Stat* 1974; **23**: 51–9.
30 Phillips AN, Thompson SG, Pocock SJ. Prognostic scores for detecting a high risk group: estimating the sensitivity when applied to new data. *Stat Med* 1990; **9**: 1189–98.
31 Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991; **10**: 1213–26.
32 Wagner DP, Knaus WA, Draper EA. Statistical validation of a severity of illness measure. *Am J Public Health* 1983; **73**: 878–84.
33 Harrel FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; **3**: 143–52.
34 Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; **247**: 2543–46.
35 Rawson NSB, Peto J. An overview of prognostic factors in small cell lung cancer. *Br J Cancer* 1990; **61**: 597–604.
36 Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; ii: 81–3.
37 Apgar V. A proposal for a new method of evaluation of the newborn infant. *Anesth Analg* 1953; **32**: 260–7.
38 Donabedian A. Evaluating the quality of medical care. *Milbank Q* 1966; **44**: 166–206.
39 Hopkins A. *Measuring the quality of medical care*. London: Royal College of Physicians of London, 1990.
40 Charlson ME, Sax FL, MacKenzie R, Fields SD, Braham RL, Douglas RG. Assessing illness severity: does clinical judgement work? *J Chron Dis* 1986; **39**: 439–52.
41 Pompei P, Charlson ME, Douglas RG. Clinical assessments as predictors of one year survival after hospitalisation: implications for prognostic stratification. *J Clin Epidemiol* 1988; **41**: 275–84.
42 Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; **112**: 564–9.
43 Cox B, Elwood JM. The effect on the stratum-specific odds ratio of non-differential misclassification of a dichotomous variable. *Am J Epidemiol* 1991; **133**: 202–7.
44 Chisholm GD, Hindmarsh JR, Howatson AG, *et al*. TNM (1978) in bladder cancer: use and abuse. *Br J Urol* 1980; **52**: 500–5.
45 Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *N Engl J Med* 1985; **312**: 1604–8.
46 Leyland AH, Pritchard CW, McLoone P, Boddy FA. Measures of performance in Scottish maternity hospitals. *BMJ* 1991; **303**: 389–93.
47 O'Grady JG, Wendon J, Tan KC, *et al*. Liver transplantation after paracetamol overdose. *BMJ* 1991; **301**: 221–3.
48 Bonsel GJ, Klompmaker IJ, van't Veer F, Habbema JDF, Sloof MJH. Use of prognostic models for assessment of value of liver transplantation in primary biliary cirrhosis. *Lancet* 1990; **335**: 493–7.
49 Hopper JL. Pathik B, Hunt D, Chann WWC. Improved prognosis since 1969 of myocardial infarction in a coronary care unit: lack of relation with changes in severity. *BMJ* 1989; **299**: 892–6.
50 McArdle CS, Hole D. Impact of variability among surgeons on post-operative morbidity and mortality and ultimate survival. *BMJ* 1991; **302**: 1501–5.