1 The measurement of outcomes of health care

ANTHONY HOPKINS*

There are a number of reasons for the increasing interest in outcomes of health care interventions over the last few years. Foremost is the realization that in all health care systems, resources are limited, and should be directed towards those interventions that are of proven effectiveness, and producing outcomes that are valued by patients. As will be discussed below, deciding upon the outcome to be achieved is a necessary prerequisite for determining whether any treatment is effective.

The next stimulus to outcomes research is the realization that there are large variations in practice, not only throughout the world, but within the same health care system, and indeed in neighbouring cities. For example, Wennberg's comparison between Boston and New Haven, two university cities on the East coast of the United States, showed that in Boston people had twice the chance of having a carotid endarterectomy compared to New Haven, but only half the chance of having coronary bypass surgery (Wennberg *et al.* 1987). These differences were apparent even when corrections were made for the age and sex distribution and other variables between the two local communities. Another striking example is that rates of hysterectomy correlate more closely with the number of gynaecologists per head of population than they do with the number of women within a population. Such variations indicate that much of what doctors do is a matter of practice style, is imprecise, and is not related to procedures of proven effectiveness. Belatedly, the precision that has been brought to bear in biomedical research is now being brought to health services research. However, as in many other areas of research measurement of the variable concerned is not straightforward, and careful attention to methods is necessary if gross errors of measurement and interpretation are not to be made. The third pressure to develop outcome measurement is the increasing strength of consumer movements. Consumer pressures now affect more strongly not only the medical profession, but teaching, the law, and public service. The public now demands that all professions are more accountable for their work. The public is far more likely to leave self-regulation to health professionals if they can be assured that reliable systems of outcome measurement are in place, monitored by clinical audit.

* Dr Hopkins died suddenly on 6 March 1997, after this book went to press.

Definition

An outcome is a change in state that is attributable to a process; in terms of health, a change in health status (however defined) that is attributable to a health care intervention. Sometimes the intervention prevents a change in state. For example, pertussis immunization changes the recipient's immune state, and prevents whooping cough.

The population

Much of this book will be devoted to measurement of health status but first, we must consider the target population whose health states we are measuring. This may range from an individual to a nation.

For a nation, perinatal mortality has traditionally been taken as an overall measure of its health, because it reflects important components of the system, namely antenatal care, including nutritional support to pregnant women, good obstetric care, and good neonatal care. Furthermore, there is no doubt about the outcome measure – the baby is either alive or dead. An improvement of perinatal mortality has therefore been taken as a proxy for a measure of the health of the nation as a whole. It may indeed be a better proxy than expectancy of life at birth, which in developed societies may depend more upon the genetic pool than on health care interventions. For example, the expectancy of life for Japanese men is longer than for caucasians in the UK. This difference may be, but is not necessarily, due in large part to the lower incidence of coronary artery disease in the Japanese.

Years of life free from disability

There is increasing concern that extending the expectancy of life into older age will add years to life, but provide a poor quality of life in very old age, with the last years being spent in increasing infirmity. Put another way, have the gains in longevity increased the number and percentage of very ill, frail people who require protracted and expensive medical care and whose wellbeing is severely compromised? As a portmanteau measure for the outcome of the health care system of a nation, therefore, years of life free from disability is gaining increasing prominence. To measure this, of course, requires an operational definition of freedom from disability, but several are available (Grimley Evans 1993). Disability-free life years as a measure is particularly relevant to neurological practice, insofar as non-lethal impairments, such as persisting disability after stroke, significantly impact upon this measure. Research efforts should concentrate on delaying the onset of such diseases if we are not to have longer life but worsening health.

Inter-agency working

Although perinatal mortality and expectancy of life at birth and at different ages might be taken as good proxies for the outcome of health care systems, it must be remembered that both these indices are dependent upon other structural qualities in society, such as the provision of adequate nutrition, good sanitation, safe roads, government policies related to smoking, and so on. The United Kingdom is prominent amongst all developed nations in recognizing the need for inter-agency collaboration between the different Departments of State in order to improve health. However, it is to be much regretted that such collaboration does not extend to more rigorous financial disincentives to smoking and excessive consumption of alcohol.

Regional and local services

Descending from the national level, outcomes are also of importance in regional planning. For example, there are significant regional variations in mortality from stroke in the UK (Department of Health 1992). Although knowledge of these outcomes cannot benefit those who have already suffered a stroke, or who are dead, such measurements of health outcome should alert public health physicians working in localities with high rates to redouble their work to ensure adequate efforts in primary care to detect and treat hypertension, and to develop local incentives to encourage people to stop smoking, such as the establishment of smoking clinics, as two examples. There is now in the UK a national effort to develop a number of health care indicators, the so-called phase 3 indicator project of the Department of Health. The potential use of these and other indicators when considering comparability between outcomes of different service units must also be considered, but neurologists and neurosurgeons alike must be aware that there is now a developed public interest in the aggregated outcomes of the work of their service. All of us in clinical practice therefore need to understand something of the field, both so that we can influence policy when irrational 'league tables' are published, but, more particularly, so that we can use them as indicators for exploration within our own service for improvement of the quality of the care that we are delivering.

Outcomes for the individual patient

For many practising neurologists, however, the principal focus of outcome measurement is centred upon the consultations taking place between individual patients and their doctors and the supporting team of nurses, therapists and so on. For neurosurgeons, perhaps, the focus is more upon the outcomes of technical procedures, but the principle is the same. In this context an outcome of a consultation, a therapy or a procedure can be rephrased as follows: 'What is it that I, as a neurologist or neurosurgeon, am hoping to achieve with my management of this particular patient?' Rephrasing the term 'outcome' in this way has at least three benefits. It underlines the duality of

the interaction between doctor and patient, so that patient values are pre-eminent; it provides a focus for action; and it should encourage before and after measurement, however informal in the first instance.

Attributability, efficacy and effectiveness

All neurologists have had occasional initially satisfactory consultations with patients who say that their epileptic seizures have stopped, the neurologist attributing this to his carefully chosen therapy, only to be disabused by the patient sheepishly confessing that he or she has not taken the tablets anyway! This raises an important point about outcome measurement; its *attributability*. In order to be sure that a health care intervention is effective, one has to have a sound epidemiological basis of the natural history of disease. It is impossible, except on probabilistic terms, to generalize from epidemiological studies to the course of an individual patient's illness, but, without a crystal ball, that is the best that we can do. Clearly, health care systems should only be interested in funding interventions to which an effect can be attributed. Research funding has in the last 40 years often supported randomized controlled trials in order to determine the *efficacy* of different therapies. In most trials, however, the outcome measure defined at the beginning of the trial is usually comparatively straightforward – in the case of tuberculosis, survival, or eradication of the bacillus from the sputum; in the case of hypertension, reduction in blood pressure to a previously defined range. Such trials can determine reliably the *efficacy* of interventions, but there is increasing realization that what may be efficacious in a randomized controlled trial, with the research team chasing up and supervising the therapy of individual patients, may not be *effective* when translated into routine health care delivery. For example, some antituberculosis regimes may be ineffective in developing countries unless systems are introduced, in parallel with the provision of the drug, to ensure supervised taking of the drug. With regard to the example of therapy for hypertension, there is no doubt from randomized controlled trials that beta-adrenergic blocking drugs reduce blood pressure. However, in practice many patients are non-compliant with medication because of the adverse effects of impotence, cold extremities and so on. In theory, one of two drugs might be more efficacious in randomized controlled trials, but less effective in practice because of lesser compliance.

Clinical importance of an intervention

The example of therapy for hypertension can be used to illustrate another point: that is, *clinical importance*. For example, very large-scale trials of drugs used in the management of hypertension may show that the reduction in diastolic blood pressure by drug A is, on average, 2 mm greater than drug B, and that the difference is highly statistically significant. The fact that the difference is statistically significant is not necessarily important if, for example, drug A is ten times the price of drug B, or accompanied by

Table 1.1. *Effect of combination of clinical features upon rate of recurrence at 1 year after a first seizure in 304 adults (late entry group excluded)*

Age < 50 Seizure between midnight and 8.59 am Family history of epilepsy/febrile convulsions	% recurred	Estimated probability of recurrence (95% confidence interval)
28 had none of these features, 5 recurred	18	0.18(0.075–0.28)
177 had one of the features, 54 recurred	31	0.30 (0.25–0.36)
84 had two of the features, 35 recurred	42	0.43 (0.35–0.51)
15 had all three features, 9 recurred	60	0.56 (0.42–0.70)

more unwanted effects. Furthermore, although true in population terms that drug A is more efficacious than drug B, it becomes difficult to conceptualize, in relation to the prognosis of an individual patient rather than of large-scale trial populations, as to what a difference in 2 mm of mercury actually means. To bring this example back into the context of neurology, Hampton calculated on the basis of the MRC trial of therapy for mild hypertension that more than 800 person years would have to be treated in order to prevent one stroke (Wilcox *et al.* 1996). Other similar examples are given by Laupacis *et al.* (1988). It is increasingly clear therefore that randomized controlled trials can provide highly important evidence about efficacy, but can only be a rough guide as to what is fruitful to put into everyday practice, when health care expenditures and patient preferences become pre-eminent.

Case severity and co-morbidity

Many clinicians feel that those who are interested in aggregating outcome measures into indicators of performance fail to recognize the impact of case severity and coexisting illnesses upon outcome. To take an obvious example, no neurosurgeon would be interested in a study about the outcome of surgery, radiotherapy and chemotherapy for cerebral glioma unless the study took into account markers known to predict poor outcome, such as disability at presentation and histological grade (Davies 1996). Table 1.1 illustrates another example from my own work with Garman and Clarke on the risk of recurrence after a first epileptic seizure.

First seizure patients with different attributes have different risks of relapse. Almost certainly there are other risks yet to be discovered. Unless such factors or 'case-severity' measures are taken into account, it is meaningless to compare outcomes of care, so there is a tremendous research effort in all branches of medicine to determine factors that predict good and poor outcomes in the natural history of any disease.

Unless case severity is taken into account, aggregated outcomes from tertiary or

university hospitals may appear to be worse than the outcomes from what may be termed 'ordinary' secondary care, until it is realized that the most complex and difficult cases tend to end up in the most academically and technically advanced centres.

Many older patients suffer from diseases that coexist with the primary diagnosis for which the outcome is being measured. For example, the outcome of a patient with a stroke who has extensive vascular disease, manifest by coronary artery disease and peripheral vascular disease, and who also has diabetes is likely to be significantly worse than someone of the same age who has stroke without any of these co-morbidities.

There are two dangers here: first, as already mentioned, the complexity of case severity and co-morbidity, and the difficulties in measuring the impact of these, will be inadequately recognized by those who construct aggregate outcome measures; the second is that units knowing that their outcomes are likely to be looked at, will turn down cases for therapy or for operation, simply because taking on these cases will make their results 'look bad', even though individual patients within the aggregate may be strikingly helped by an intervention. In the United States, case severity is often estimated by commercial software systems that require the clinical record to be reviewed to collect data. Examples include MEDIS groups and Systemmetrics. However, retrospective record review is expensive. Nonetheless, striking examples of the effectiveness of an academic analysis of case severity are the APACHE system in intensive care, which measures case severity by a number of physiological variables in the first 24 hours (Rowan *et al.* 1993), and the CRIB system in neonatal intensive care, which does much the same (The International Neonatal Network 1993). There is now good evidence that the outcomes of survival in both these intensive care situations can be predicted with a high degree of accuracy by such measures of case severity. Furthermore, it is proposed that, having corrected for case severity, survival can be used as an audit measure for the success of a unit. For example, no district general hospital achieved a survival rate corrected for case severity using the CRIB scale for neonatal intensive care that was as good as any teaching hospital. This is circumstantial evidence that care in the latter is better.

Retrospective record review is expensive, and there is now good epidemiological evidence of factors that do predict a poor outcome in many disorders. For example, in neurological practice in relation to stroke, there are a number of prognostic scores (e.g. Allen 1984). It might be feasible to audit the effectiveness of rehabilitation units by considering the functional status of those with stroke, modified by the prognostic score. That is to say, a patient who failed to walk again after a stroke, having been predicted to walk again on the basis of a measure of case severity, might be assumed, other things being equal, to have had inadequate rehabilitation.

Although the case severity and comorbidity, if adequately measured, can sometimes be considered as satisfactory 'explanations' for less than satisfactory outcomes, age and, in particular, ethnicity should not be accepted as 'excuses' without scrupulous self-appraisal. For example, old people with coronary artery disease may have poor outcomes simply because they are not offered effective interventions such as angio-

plasty or coronary bypass surgery as often as they ought to be (Krumholz *et al.* 1993); in the field of neurology, old people may not be considered to be 'worth' spending the money on effective rehabilitation. With regard to ethnicity, poor outcomes may be too readily attributable to racial (i.e. genetic) differences, whereas in truth the differences are due to less good care. A particularly striking example is a careful analysis of outcomes of renal transplantation in Afrocaribbean and Caucasian Americans. Although graft survival was worse in Afrocaribbean individuals, analysis of the population studied showed that these recipients received less well matched kidneys than Caucasian recipients, and if this and other factors were controlled for, then the survival of both recipient populations was more or less the same (Butkus *et al.* 1992).

Multiplicity of outcomes, and patient values

So far I have considered straightforward outcomes, such as death, or eradication of bacteria, or a reduction in blood pressure. However, a central aspect of out thinking in the measurement of outcomes, must be the multiplicity of possible outcomes, and their valuation by the individual patient.

There is often a potential conflict between the outcome valued by a patient and the outcome that the neurologist values. For example, neurologists know from epidemiological experience that those who have little use of the hand 24 days after a hemiplegic stroke stand little chance indeed of having useful function of that hand in the future (Heller *et al.* 1987). However, they know from the scientific literature that, if hypertensive, a reduction in the patient's blood pressure will reduce the chances of a subsequent stroke. Neurologists also know from experience that handicap can be successfully minimized by suitable attention to the patient's environment and the provision of appropriate aids and appliances. A neurologist's successful outcome therefore will be, with the aid of an occupational therapist, to help a hemiplegic woman back to self-care, to work or to look after her own home. The patient's perspective on outcome, however, will be that the hand has not got better, and that as far as she is concerned, treatment has been a failure. All doctors must sit down with their patients when they plan management and inform them what outcomes can be realistically expected, such disclosure being tempered by what is thought to be kind, and supportive of the individual patient. It is likely that resources are wasted upon continuing physiotherapy for stroke patients for whom there is no likelihood of useful further recovery simply because the neurological team have not had the courage to explore with the patient both what their difficulties are, and what can be realistically offered in the way of improvement.

The case of a hemiplegic stroke can be used to illustrate another point about the multiplicity of outcomes. In spite of the protestations of speech therapists, the weight of research evidence suggests that speech therapy does not much help the recovery of language after stroke. An important UK study showed that recovery, in terms of one respected measure of communication (the PORCH index), was the same whether the

'therapy' was given by speech therapists or by volunteers, who had received some basic training and supervision by a speech therapist (David *et al.* 1982). There are a number of possible interpretations of the recovery in communication. It is possible (were it not for other studies [David *et al.* 1982]) that volunteers can acquire, in a few hours training, the skills that it takes speech therapists three years to acquire. More likely, the limited recovery reflects innate spontaneous cerebral recovery. But this example would not encourage any health provider to put further resources into the training of speech therapists for the treatment of dysphasia following stroke.

Speech therapists might respond by stating that they accepted that the evidence was slight that their efforts could improve language, but that patients valued their help in *coming to terms with their difficulties, and in being taught various coping strategies*. If the words in italics are defined as the outcome, then it would be necessary to mount a new trial. Perhaps volunteers would be as effective in these domains as well. Using this example as a model, social research should discover what achievable outcomes are valued by patients, and further discover the most cost-effective way of achieving them. To switch examples, if a specially trained nurse, on a lower salary than a junior doctor, more successfully harvests saphenous vein grafts for coronary bypass surgery, then there are good arguments for moving to such a system.

The problem of outcome definition may perhaps more strikingly be brought into prominence by considering the case of a woman aged 38 years who has had bitemporal headaches for the last 2 or 3 years, worse in the evening and worse at times of menstruation, not relieved by analgesics. They occur on a background of confessed anxiety about her husband's fidelity. All clinical supposition is that she has tension headaches. Two weeks before this neurological consultation, a distant acquaintance was reputed to have a brain tumour, and the headache patient has taken it into her head that she might also have a tumour. Such anxieties amongst patients with headaches are common, and are usually relatively easily laid at rest (Fitzpatrick & Hopkins 1981). The patient feels that she needs a scan of some sort to exclude a tumour. The outcome that both patient and neurologist will wish to achieve is reassurance about the absence of a tumour, which in itself may go some way towards encouraging the patient to cope with her headaches, even if they do not necessarily resolve. From a technical diagnostic point of view, an imaging study in this circumstance is a waste of the resources of the health care system, as the chances of it showing a tumour or some other important treatable lesion are extraordinarily low (Lavson *et al.* 1980). However, if the focus of the interaction between doctor and patient is on reassurance, rather than on the sensitivity and specificity of the investigation, then the imaging study may be an effective and an appropriate intervention. All clinical neurologists will recognize this dilemma, and hopefully most attempt to spare resources by relying upon relatively cheap counselling and supportive therapy. However, the conflict in perspective between what is considered appropriate by the health care system from a population perspective, determined to husband its resources for effective technical health care, and what the patient considers to be appropriate has not been resolved (NHS Management Executive 1993). The unwritten understanding that doctors control access to

investigations and procedures for safety's sake becomes strained at the edges, particularly when an investigation costs a lot of money, but, as far as we know, the procedure itself carries no conceivable risk to the patient as is the case with magnetic resonance imaging. The conclusion of Brett and McCullough (1986) that no patient in such a situation should not have the investigation if they wanted it and were prepared to pay for it is probably just, but in practice, the burden of payment is often shifted onto health care insurers, whose utilization review tends to be more directed towards interventional procedures. Furthermore, Brett and McCullough do not recognize the diversion of capital resources and the time of trained staff away from what most physicians would consider more appropriate health care.

This apparent digression into the ethics of 'unnecessary' imaging studies does, however, introduce the notion that the outcomes towards which doctors should work must be the outcomes desired by the patient. Other chapters in this book review the reliability, sensitivity to change and so on of various measures of functional status, which are certainly one important measure of the outcome of rehabilitative care. However, it may be that neurologists and physiotherapists too readily perceive disabled people from their own stance of locomotor perfection, and concentrate excessively upon functional aspects of daily living to the extent that they fail to address how best to help patients achieve their own targets and goals, which may in part be emotional.

To illustrate further the multiplicity of outcomes that must be considered in good care, consider a simple procedure from general surgical rather than neurosurgical practice, such as a herniorrhaphy. First of all, there is the unlikely event of perioperative mortality. There are potential adverse outcomes other than mortality, including wound infection and of course recurrence of the hernia at a later date. There is the outcome of freedom from dragging pain in the groin, and from an embarrassing unsightly lump. Then there is the satisfaction of the patient with postoperative pain relief, the cosmetic acceptability of the scar, with the courtesy of the surgeon who talked to him kindly before and after the operation, and with the depth of the advice that he received in relation to future activities, such as return to work, to sexual intercourse, and to lifting. All neurologists and neurosurgeons could translate such a scenario to procedures in their own practice. All of us would acknowledge that each of the dimensions just recorded in relation to herniorrhaphy were all aspects of good care. The truth of the matter is that it is very difficult to capture such data for clinical audit, with the exception of gross adverse outcomes such as perioperative mortality; even recurrence rate may be confused by case mix, such as previous herniorrhaphy, obesity, occupation, and by co-morbidities such as chronic bronchitis which, by causing repeated coughing, makes a recurrence of the hernia more likely.

Measures of quality of life

Faced with the complexities of individual patient characteristics and of real life, there have understandably been a number of attempts to record in a single number a patient's overall 'quality of life.' Some of the methods of doing this are reviewed

elsewhere (Bowling 1991; Hopkins 1992) but there are two principal ways of going about it. First of all, researchers can decide, after prolonged exploration with the public, what dimensions of existence are important (mobility, freedom from pain, mood, sexuality and so on). Scales can then be derived for each of these dimensions, and standards set by applying the scales to a normative population. The Nottingham Health Profile (Bowling, 1991) and the Sickness Impact Profile (Bergnev *et al.* 1976) are both examples of these. There is no doubt that these scales are sensitive at detecting changes in health status which accord with clinical reality and patient judgement. However, their very multidimensionality militates to some extent against their easy use. For example if, after an intervention, a patient scores more favourably on freedom from pain, and less favourably on freedom of mobility, how do we judge the success of our intervention? To take another example from everyday neurological practice, trials of treatment for headache are bedevilled by patients who say sometimg along the lines that 'I have fewer headaches, but those that I do have are more severe.' Similarly, trials of drugs for migraine and epilepsy are bedevilled by whether or not one should weight the severity of headaches or seizures. Faced with these difficulties, therefore, another school of research tries to integrate all aspects of a patient's wellbeing and quality of life on a scale of 0 to 1. The original and highly imaginative work of Rachel Rosser (1978) was to ask members of the population to rate on two axes of impairment and distress various health states as briefly described. Critics of this work pointed out that the raters were unusually medically orientated, being healthy staff largely in and around one hospital, but that criticism has been laid to rest by extensive surveys of valuations of health states amongst a more representative population carried out by the Institute of Health Economics at York. From such valuations of health states, and from the duration of survival in those states, quality adjusted life years (QALYs) can be calculated. That is to say, ignoring discounting the future, a year of life in a perfect health state (value 1) is equivalent to two years of life in a health state valued at 0.5. If the costs of interventions and subsequent support are adequately calculated, then, in theory at least, league tables can be constructed to show what resources buy the most QALYs (Williams 1985). To illustrate this point, calculations purport to show that a hip replacement costs £800 per QALY gained, and a neurosurgical intervention on a malignant glioma over £100,000 (at current prices) per QALY gained.

There is growing concern about the use of such calculations, which do not take into account many research and ethical issues (Hopkins, 1992). From the research point of view, there is no evidence to suppose that the valuations by the general public of what it is like to be in a certain health state bear any relationship to what it is *really* like to be in that health state. Unfortunately, one cannot ask the patients in that health state, as they have no experience of other health states with which to compare their present situation. To give an example of the ethical concerns, what of the care given to people with learning disabilities? I know of no evidence that humane care improves cognitive function and memory in those with severe learning disabilities, but any ethical society would expect to look after those so severely handicapped in a humane and caring way.