Summary points

- In presenting and discussing the results of an observational study the greatest emphasis should be placed on bias and confounding
- Confidence intervals should be relegated to a small part of both the results and discussion sections as an indication, but no more, of the possible influence of chance imbalance on the result
- 90% Confidence intervals should be accepted as readily as 95% confidence intervals
- The term "statistical significance" should not be used and P values should not be published in observational studies

Recommendations

On the basis of these arguments we make the following practical suggestions regarding the use of probability statistics in observational studies.

(1) In presenting and discussing the results of an observational study the greatest emphasis should be placed, firstly, on questions of bias within the study and, secondly, on the issue of confounding.

(2) Confidence intervals should be relegated to a small part of both the results and discussion sections as an indication, but no more, of the possible influence of chance imbalance on the result. The move away from using confidence intervals as a surrogate for statistical significance would be helped if 90% confidence intervals were accepted as readily as 95% confidence intervals.

(3) The structure of any discussion would then be in terms of: (a) to what extent might flaws in the study design have biased the study result?; (b) if the result is thought to be free of bias, to what extent might other causes have confounded the observed association?; and

(c) if the result appears unbiased and known confounders have been accounted for, what might have been the extent of chance imbalance as represented by the confidence interval?

(4) The term "statistical significance" should not be used and P values should not be published in observational studies.

We thank Drs P Hannaford, S Khan, A Silman, D Symmons, the statistical referee, and Dick Keystone for helpful comments.

- 1 Hill AB. The environment and disease: association or causation? *J R Soc Med* 1965;58:295-300.
- 2 Jolley D. The glitter of the t table. *Lancet* 1993;342:27-9.
- 3 Morrison DE, Henkel RE. *The significance test controversy*. London: Butterworths, 1970.
- 4 Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421-9.
- 5 Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 1979;32:51-63.
- 6 MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the MRC vitamin study. *Lancet* 1991;338:131-7.
- 7 Davey-Smith G, Phillips AN, Neaton JD. Smoking as an 'independent' risk factor for suicide: illustration of an artefact from observational epidemiology? *Lancet* 1992;340:709-12.
- 8 Smithells RW, Shephard S, Schorah CJ, Seller MJ, Nevin NC, Harris R, et al. Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:647.
- 9 Stone DH. Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:647.
- 10 Smithells W, Seller MJ, Harris R, Fielding DW, Schorah CJ, Nevin NC, et al. Further experience of vitamin supplementation for prevention of neural tube defect recurrences. *Lancet* 1983;i:1027-31.
- 11 Rose G, Cooke ID, Poland PE, Wald NJ. Vitamin supplementation for prevention of neural tube defect recurrences. *Lancet* 1983;i:1165-6.
- 12 Greenland S. *Evolution of epidemiological ideas*. Chestnut Hill, Massachusetts: Epidemiology Resources, 1987.
- 13 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;292:746-50.
- 14 Rothman K. A show of confidence. *N Engl J Med* 1978;299:1362-3.
- 15 Poole C. Beyond the confidence interval. *Am J Public Health* 1987;77:195-9.
- 16 Lanes SF. The logic of causal inference. In: Rothman KJ, ed. *Causal inference*. Chestnut Hill, Massachusetts: Epidemiology Resources, 1988:59-75.

(Accepted 6 June 1994)

Some observations on attempts to measure appropriateness of care

Nicholas R Hicks

There are a growing number of published studies that suggest that much health care is delivered inappropriately. There are calls for measures of appropriateness to be used by purchasers and others to regulate or influence the delivery of health care. This paper explores assumptions inherent in results generated by a leading measure of appropriateness and concludes that there are considerable uncertainties about the measure's meaning, the magnitude of bias that it contains, and the degree to which its application can be generalised. Some of these uncertainties could be resolved if the tacit assumptions inherent in the generation of the criteria could be made explicit. Existing measures of appropriateness are not yet sufficiently robust to be used with confidence to influence or control the delivery of health care. They may have a use as an aid rather than as a constraint in clinical decision making. A randomised controlled trial could resolve whether patients achieve better outcomes if their care is influenced by appropriateness criteria.

A leading article by Brook published in the *BMJ* recently identified appropriateness as "the next frontier" in the development of clinical practice.¹ It argued that, firstly, there is too much information about medical practice for any doctor to assimilate all the information relevant to their practice. It is therefore impossible to "practise good medicine without

additional help." Secondly, for this (and other reasons) many patients receive care that is "inappropriate" (contributing to overuse of health care) and many others are not offered "appropriate" care (underuse of health care). Thirdly, the appropriateness of care can be measured, and, finally, the application of measures of appropriateness can reduce or eliminate both overuse and underuse of medical interventions.

These claims, if true, have huge implications for medical practice, given that some studies estimate that 20% to 60% of care is less than appropriate.² These findings have led to calls for the profession, patients, and purchasers of care to use measures of appropriateness to regulate the delivery of care. Before such measures are used to influence care in the United Kingdom, it seems reasonable to explore the meaning of appropriateness scores to ask if the results could be biased and to consider how well judgments about appropriateness can be generalised. It is even pertinent to ask what the phrase "appropriate care" means. Brook and colleagues at the Rand Corporation and the University of California at Los Angeles (UCLA) have developed and pioneered the use of one of the leading tools for measuring appropriateness of care.³ I explored the question of whether measures of appropriateness are sufficiently robust to apply to everyday practice in the United Kingdom by examining the process by which Rand appropriateness scores are generated.

Department of Public Health and Health Policy, Oxfordshire Health Authority, Headington, Oxford OX3 9DZ
Nicholas R Hicks, consultant public health physician

BMJ 1994;309:730-3

Patients have a legitimate role in judging appropriateness of care

ULRIKE PREUSS

Rand approach to measuring appropriateness

The Rand method of assessing appropriateness was developed in the mid-1980s to test the hypothesis that geographical and institutional variations in the rates of use of specific procedures—for example, hysterectomy—could be explained by variations in the proportions of appropriately delivered care.⁴ The Rand research team developed a systematic method for generating explicit criteria for appropriateness that could be applied evenhandedly to interventions performed in different institutions. Their method has been described in detail elsewhere.^{5,6} It entails a review of the literature and the generation of catalogues of all conceivable indications for using a particular procedure. A panel of nine expert clinicians is appointed. Each panellist is sent a copy of the literature review and the catalogue of potential indications and asked to rate the appropriateness of performing the procedure for each potential indication on a nine point scale (1 being extremely inappropriate, 9 extremely appropriate). The panel then meets. Each panellist is reminded of the way they rated each indication and given an anonymous breakdown of the other panellists' ratings for each indication. After discussion of areas of disagreement, panellists anonymously rerate the entire set of indications. For each indication a mean score and a measure of the panel's agreement is calculated. Where the mean score for an indication is 1 to 3 and there is broad agreement among the panel the indication is classified as inappropriate. When the mean is 7 to 9 and there is agreement the indication is classified as appropriate. If the mean is 4 to 6 or if there is disagreement among the panel then the indication is classified as equivocal.

The innovative Rand method of assessing the appropriateness of care was successfully applied in the study for which it was designed. It allowed the researchers to draw the important and now widely believed conclusions that substantial proportions of health care are inappropriate and that geographical variations in rates of intervention cannot be explained by variations in the proportions of care that are delivered appropriately.⁴ Subsequently, although the Rand method has been criticised for ignoring patients' preferences,⁷ overestimating rates of inappropriateness,⁸ generating criteria that may rely on consensus that cannot be supported by scientific evidence,⁹ and for ignoring the subjective, "visceral" judgments that doctors reach during a consultation,¹⁰ it has been used by many insurance companies in the United States which, as a condition for paying a doctor's fee, require that doctors obtain previous approval before operating

on any patients insured with them. It has also been promoted in the United States to doctors' employers and preferred provider organisations as a means of assessing the appropriateness of the practices of individual doctors.

Observations and questions about the Rand approach

DEFINITIONS OF APPROPRIATENESS

A reliable and valid measurement of appropriateness requires a clear and precise definition of appropriateness. The Rand method defines care as appropriate when "for an average group of patients presenting to an average US physician... the expected health benefit exceeds the expected negative consequences by a sufficiently wide margin that the procedure is worth doing... excluding considerations of monetary cost." A British group has criticised the Rand definition for omitting two important determinants of appropriateness: resources and the individuality of the patient.¹¹ The same British group proposed an alternative, longer definition of appropriateness.¹¹ Although there remains no universally accepted definition of appropriateness, the Rand definition is probably the most widely used of the various definitions that are available.

AIMS OF CARE

A simpler definition of appropriate is offered by the Oxford English dictionary. It defines appropriateness as "suitable or proper to or for [a particular purpose]." The inclusion of the words to or for implies that appropriateness of care should be judged in the light of knowledge of the intended outcome(s) of intervention. The Rand definition of appropriateness does not make the intended outcome(s) of care explicit, and the Rand process for generating criteria does not require the panellists to make the aims of intervention explicit as they rate each indication. This is important because different people may legitimately have different aims and expectations for care, even in identical clinical circumstances. One cannot necessarily assume that when an indication is rated by nine different panellists that every panellist has in mind the same intended outcome. This makes it hard for users of the appropriateness ratings to understand the meaning of the ratings.

RISKS AND BENEFITS: SELECTION

The Rand process asks panellists to judge the net benefit of intervention for each indication by balancing their assessment of the risks and benefits of intervention. Neither the definition nor the process, however, makes explicit which risks and benefits panellists have taken into account or ignored. Different panellists will probably take into account different benefits and different risks.

RISKS AND BENEFITS: ESTIMATING MAGNITUDE OF EFFECTS OF HEALTH CARE INTERVENTION

Panellists' assessment of the appropriateness of intervention will depend on their beliefs about the size of the effect of intervention. For some of the indications that the panellists are asked to rate, their estimates of effect size may be based on the literature review provided by the Rand researchers. For most indications, however, the literature is uninformative about the size of effect. Even when presented with evidence doctors vary greatly in their assessment of the size of an effect. Variations in judgments seem to have both random and systematic components. In particular, doctors may commonly underestimate the risks of intervention.¹³ Estimates of appropriateness are likely to be flawed if they are based on inaccurate estimates of how big an effect is.

RISKS AND BENEFITS: WEIGHTING RELATIVE IMPORTANCE OF DIFFERENT OUTCOMES

Panellists are asked to assess the net benefit of intervention by balancing the risks and benefits of intervention. It seems unlikely that different panellists will judge the relative importance of particular risks and benefits equally. Indeed, the Rand group itself has shown that doctors practising in different specialties rate the appropriateness of care in systematically different ways—for example, surgeons are more likely to rate surgical interventions as appropriate than are either general or specialist physicians.¹⁴ The implicit weights that panellists attach to different dimensions of outcome are important determinants of their assessment of appropriateness of care. The extent to which the appropriateness scores are generalisable is influenced by the degree to which these weights are shared by users of the scores.

WHAT IS WORTH DOING?

The Rand definition instructs panellists to judge something appropriate when the net benefit is sufficiently large that it is “worth doing.” Again, the definition and process give no guidance as to what magnitude of net benefit is worth doing, and again there is no reason to suppose that different panellists should necessarily agree about what is worth doing.

The developers of the Rand method have indicated that they were more concerned with improving the quality of care than reducing health care costs. This is reflected in the Rand definition of appropriateness, which asks panellists to “exclude considerations of monetary cost.” Many people, however, believe that it is impossible to exclude financial costs in making judgments about the appropriateness of care, particularly in a cash limited system such as the NHS.

WHAT IS THE COMPARISON TREATMENT?

The Rand definition does not make clear what alternative treatment, if any, panellists should consider in making their judgments of appropriateness. Is a procedure appropriate only if it is the best possible treatment for a given indication or is it considered appropriate if it is better than no treatment at all? Under what circumstances is it appropriate to offer a second or third best option to a patient?

WHOSE VIEWS ARE RELEVANT?

The Rand method uses a panel of doctors of high reputation to judge appropriateness. The method does not incorporate a lay view of any sort. Whereas doctors' views are important, most would accept that others—for example, patients, relatives, and society—also have a legitimate role in judging the appropriateness of care.¹¹

TO WHOM DO THE CRITERIA APPLY?

Rand panellists are asked to judge the appropriateness of performing a procedure for an “average group of patients [who meet the medically defined criteria of particular indications] presenting to an average US doctor.” There is no method for adjusting the Rand scores for differences in individual patients' circumstances, hopes, and fears. How well can one doctor extrapolate his or her practice to those of others? How generalisable do the panellists expect their ratings to be? Do they expect them to apply to every patient who meets the detailed clinical criteria of the indications? How can users of the appropriateness scores judge how similar any individual patient is to the hypothetical patient that the panellists had in mind?

WHICH FACTORS DO PANELLISTS TAKE INTO ACCOUNT WHEN RATING APPROPRIATENESS OF CARE?

It is also far from clear what factors panellists take

into account when rating the appropriateness of care. Are they scoring the strength of evidence for a particular outcome; their own confidence that a particular outcome will be achieved; or their estimate of the size of net benefit. These are separate and independent attributes. Confidently held beliefs can be wrong. How should a panellist rate the appropriateness of intervening for a small but certain net benefit compared with the appropriateness of intervening for a potentially large but uncertain net benefit? Without knowing which factors panellists have taken into account in reaching their judgments of appropriateness it is hard for users of the scores to know what the scores are measuring.

Possible sources of bias

LITERATURE REVIEW

It is a strength of the Rand method that the process of developing appropriateness scores includes a review of the literature. Unsystematic reviews, however, can produce unreliable conclusions.¹⁵ It is important that users of the Rand method should be able to assure themselves that the reviews on which panellists have been asked to base their judgments are systematic and well documented.

PANEL SELECTION

The appropriateness criteria produced by the Rand method are summaries of the collective opinions of nine panellists. Different panellists reach different conclusions. Their differences may be random or they may be systematic—that is, biased. Strong evidence that different panels studying the same issues can produce systematically different judgments was provided when two Rand type panels—one made up of doctors from the United States and one of doctors from the United Kingdom—considered the appropriateness of angiography and of coronary artery bypass grafting. The two panels disagreed on 226 (47%) of the 480 indications rated for coronary artery bypass surgery and on 150 (50%) of the 300 indications rated for coronary angiography. Of the 376 indications on which the panels disagreed, in only 14 (3.7%) did the British panel consider intervention more appropriate than did the American panel.¹⁶ The differences between the panels were systematic, showing that the process is highly sensitive to the selection of the panel members.

THE ROLE OF PANEL CHAIRPERSON

I have been fortunate to meet several people who had participated in or observed one or more Rand panels. They emphasised the importance of the role of the panel's chairperson. The chairperson can guide discussion and focus the panel on particular issues and must make quick decisions on details that are important to the consensus making process such as proposing and sanctioning changes to the structure and content of the indication catalogue. Opinions were divided as to how sensitive the final ratings are to the performance of the chairperson, but several people thought that he or she may have a very important effect. The influence of the panel chairperson on the final product has yet to be investigated.

Conclusions

The detailed examination of the Rand approach to measuring appropriateness illustrates several points about the nature of appropriateness and begins to suggest how existing scores might be refined.

Most importantly, appropriateness is seen to be an abstract concept whose assessment necessarily entails value judgments. The assessment of appropriateness is highly dependent on the context in which care is

delivered and the judgment made. Attempting to produce a valid, reliable, and widely generalisable definition of appropriateness that can be used as the basis for measuring appropriateness of care in various clinical settings may thus be impossible. A judgment about the appropriateness of care, however, will be more interpretable and therefore more useful to others if the details surrounding the judgment are made as explicit as possible.

These comments should not be interpreted as implying that the Rand method produces no useful information. It was an innovative and well thought out tool for the purpose for which it was originally designed—that is, retrospective comparison of care received by groups of patients in different institutions. Examination of the detail of the Rand process, however, suggests that there are uncertainties about what the appropriateness scores measure and about the biases that they might contain. The legitimacy of the scores can also be challenged as they exclude the views of patients, carers, and other relevant lay groups. In addition, because panellists have to make many tacit assumptions in their judgments of appropriateness it is impossible for users to be confident that the ratings are relevant to their own practice. The uncertainties surrounding the interpretation of the Rand scores suggests that one should be cautious before adopting them as guides for care to be given to individual patients. This conclusion is reinforced by theoretical considerations which suggest that existing measures of appropriateness are likely to have considerable false positive rates and so substantially overestimate rates of inappropriate care.⁹ The implication is that as yet there is no case for purchasers using existing guidelines for appropriateness to determine whether care for individual patients should be funded or not. What may be worth exploring, however, is whether the information contained within appropriateness criteria can be harnessed to improve clinical decision making.

There are at least two routes forward. One might experiment with existing appropriateness guidelines as part of a randomised trial in which the hypothesis to be tested is that patients whose doctors have access to information about ratings of appropriateness of care before care is delivered achieve better outcomes than patients whose doctors do not have information about appropriateness ratings. Alternatively, it might be argued that the uncertainties surrounding the meaning, biases, and ability to generalise existing measures are so great that they are not suitable for any prospective use without refinement. A first step to reducing these uncertainties would be to make explicit some or all of the assumptions that are currently tacit. In particular it would be helpful to specify the following:

- (1) Who made the judgment? This allows the user of the rating to decide whether another's opinion is relevant.
- (2) When was the judgment made? This allows the user to decide whether it is likely that there is new and important information that has not been taken into consideration.
- (3) Whose interests and views were they representing? Does the surgeon speak for the physician? Is a relevant lay view represented?
- (4) What was the intervention in question? This should be easy to specify.
- (5) What were the clinical indications? The Rand method identifies clinical details well.
- (6) What were the most important hoped for outcomes? Some patients may place quality of life before length of life.
- (7) What risks and benefits were taken into account? Not everyone routinely considers all the dimensions of

outcome that may be relevant to patients—for example, how many cardiac surgeons discuss non-specific neurological risks of cardiac surgery with their patients preoperatively?

(8) What was the estimated probability and magnitude of particular risks and benefits? Doctors' assessments of the likelihood of particular outcomes of care vary widely.¹² It is these probabilities that underpin doctors' advice to patients. If we are to understand the basis of the judgments of others it will help us to know their beliefs about the probabilities of risk and benefits.

(9) What is the strength of evidence against which the judgment was made? Is a judgment a sincerely held but scientifically unsupported opinion or is it supported by high quality scientific research? The most sincerely held opinion of the most eminent physician can still be wrong if it is not supported by science.

In summary, although measures of appropriateness may seem to be objective, the process by which they are produced, although systematic, remains highly subjective. The Rand appropriateness criteria are consistent with the literature and summarise a body of expert medical opinion. They would be easier to interpret if the tacit assumptions that underlie their generation could be made explicit. It is a plausible and testable hypothesis that patients would achieve better outcomes if doctors knew whether the care they were considering offering was rated appropriate or not. In the absence of such information and given the uncertainty of the meaning of the scores, they are as yet unsuitable to be used to form the boundaries of clinical practice.

Much of the material in this paper is based on work that began in 1991-2 when I spent several months in the United States based at the Rand Corporation, Santa Monica, California, as a Harkness Fellow funded by the Commonwealth Fund of New York. I am grateful to the Commonwealth Fund for financial support and to Dr R H Brook and his colleagues at Rand who generously gave their time to discuss the measurement of appropriateness and related issues.

- 1 Brook RH. Appropriateness: the next frontier. *BMJ* 1994;308:218-9.
- 2 Winslow CM, Koseoff J, Chassin MR, Kanouse DE, Brook RH. The appropriateness of performing coronary artery bypass surgery. *JAMA* 1988;260:505-9.
- 3 Bernstein SJ, McGlynn EA, Siu AL, Roth C, Sherwood MJ, Keesey J, et al. The appropriateness of hysterectomy: a comparison of care in seven health plans. *JAMA* 1993;269:2398-402.
- 4 Chassin MR, Koseoff J, Park RE, Winslow CM, Kahn K, Merrick N, et al. Does inappropriate use explain geographic variations in the use of health care services? *JAMA* 1987;258:2533-7.
- 5 Brook RH, Chassin MR, Fink A, Solomon DH, Koseoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986;2:53-63.
- 6 Kahn KL, Koseoff J, Chassin MR, Flynn M-F, Fink A, Pattaphongse N, et al. Measuring the appropriateness of the use of a procedure: can we do it? *Med Care* 1988;26:415-22.
- 7 Kassirer JP. The quality of care and the quality of measuring it. *N Engl J Med* 1993;329:1263-5.
- 8 Phelps CE. The methodologic foundations of studies of the appropriateness of medical care. *N Engl J Med* 1993;329:1241-5.
- 9 Mulley AG, Eagle KA. What is inappropriate care? *JAMA* 1988;260:540-1.
- 10 Tanenbaum SJ. What physicians know. *N Engl J Med* 1993;329:1269-71.
- 11 Hopkins A for a working group for the Director of Research and Development of the NHS Management Executive. What do we mean by appropriate health care? *Quality in Health Care* 1993;2:117-23.
- 12 Eddy DM, Billings J. The quality of medical evidence: implications for quality of care. *Health Affairs* 1988;7:19-32 (Spring).
- 13 McClellan M, Brook RH. Appropriateness of care: a comparison of global and outcome methods to set standards. *Med Care* 1992;30:565-86.
- 14 Park RE, Fink A, Brook RH, Chassin MR, Kahn KL, Merrick NJ, et al. Physician ratings of appropriate indications for six medical and six surgical procedures. *Am J Public Health* 1986;76:766-72.
- 15 Chalmers I, Hetherington J, Elbourne D, Keirse MJNC, Enkin M. Materials and methods used in synthesising evidence to evaluate the effects of care during pregnancy and childbirth. In: Enkin M, Keirse MJNC, Chalmers I, eds. *Effective care in pregnancy and childbirth*. Oxford: Oxford University Press, 1989:38-59.
- 16 Brook RH, Koseoff JB, Park RE, Chassin MR, Winslow CM, Hampton JR. Diagnosis and treatment of coronary heart disease: comparison of doctors' attitudes in the USA and UK. *Lancet* 1988;i:750-3.

(Accepted 6 June 1994)