The quiet revolution: Reporting of health outcomes in general medical journals

Janelle Seymour, David Newell and Alan Shiell

Janelle Seymour is a Senior Research Assistant at the Centre for Health Economics Research and Evaluation (CHERE), University of Sydney. David Newell is a Consulting Statistician. Alan Shiell is Senior Lecturer in Health Economics at CHERE, University of Sydney.

Abstract

This study reviews the extent of evaluation of health outcomes in three general medical journals over the past decade by examining papers published in the original research section of the New England Journal of Medicine (NEJM)*,* The Lancet*, and the* Medical Journal of Australia (MJA) *in 1982 and 1992. Evaluations were identified and classified according to the type of comparison group and the type of outcome measures employed. They were divided into three categories: those employing a comparison group; those employing a before-and-after study design (or own comparison group); and those with no comparison group. The categories of outcome measures were mortality, clinical or intermediate measures of health state, and final outcome measures (quality of life). Results show that the proportion of papers evaluating a health services intervention remained stable over the period. However, the* MJA *published considerably fewer evaluations than the other journals. In the* NEJM *and* The Lancet*, 75 per cent of evaluations incorporated comparison groups, in the* MJA*, less than 40 per cent. Overall, the proportion of papers reporting final outcome measures increased significantly between 1982 and 1992 ($p = 0.04$) but the change in each journal individually did not reach statistical significance. This study indicates that the reporting of health outcomes evaluations has remained constant but there has been some change in the use of comparison groups and final outcome measures over time.*

Introduction

A recent editorial published in *The Lancet* noted that 'Health outcomes and, *pari passu*, a questioning of traditional medical practices are high on the menu of researchers in Europe and the USA' (Anon 1992). Doctors are being increasingly asked: 'Does it work?' (Anon 1993).

In the United States, concern over variations in health care intervention rates has led Congress to direct the newly established Agency for Health Care Policy and Research to study the outcomes of treatment. This has resulted in the establishment of 11 Patient Outcome Research Teams, each with a brief to examine clinical practice and patient outcomes for important health conditions (Clinton 1991). In the United Kingdom, the reform of the National Health Service has encouraged interest in health outcomes as purchasing authorities seek ways of monitoring the quality of service provision, and providers compete for funding. This has prompted the development of a clearing house to disseminate research findings and evidence on the effectiveness of treatments (Long, Bate & Sheldon 1992).

Such international interest in the measurement of health outcomes is also reflected in Australia. The publication of the Commonwealth National Health Goals and Targets document (Nutbeam et al. 1993) and the New South Wales Department of Health's initiative on outcomes has helped to place the measurement of health outcomes high on the political agenda (Frommer, Rubin & Lyle 1992). This, in turn, has contributed to the inclusion of health outcomes in the new Medicare Agreement, though the precise form this will take is as yet unspecified (Leeder 1993). A clearing house, similar to those established in the United States and the United Kingdom, has also been set up by the Australian Institute of Health and Welfare (1994).

If these initiatives on outcomes are to be translated into better medical practice, the evaluation of health care must move from being an exclusively academic interest to become part of the conventional wisdom of practising clinicians. The aim of this paper is to consider whether the new emphasis on the evaluation of health care and the measurement of health outcomes is reflected in the medical literature. Comparison is an essential feature of health care evaluation and is '...crucial in reaching conclusions about what is normal or abnormal in determining whether a treatment improves the course of a disease' (Grisso 1993, p 157). Therefore, good study design is required if changes in health outcomes are to be attributed to medical practice. To this end, articles in three general medical journals were reviewed to assess whether there has been a change in the number of published studies which used comparison groups and which reported final outcome measures over the past decade.

Methods

Three general medical journals which reflect international coverage and high quality research were chosen for review. These were the *New England Journal of Medicine (NEJM)*, *The Lancet* and the *Medical Journal of Australia (MJA)*. Each journal contains material of general medical interest and is peer-reviewed. With one exception, this selection matches the journals reviewed by Fletcher and Fletcher between 1946 and 1976 (Fletcher & Fletcher 1979a, p 180). The exception is that we have substituted the *MJA* for the *Journal of the American Medical Association (JAMA)*. The *MJA* is of obvious local relevance and importance, while the *NEJM* and *The Lancet* have an extensive international readership and are among the most commonly cited journals (Garfield 1979).

Sixteen issues of each journal were reviewed, eight from April and May 1982 and eight from the same two months in 1992. In the case of the *MJA*, the sample period was extended to include March and June of both years to allow for the fact that the journal is published fortnightly rather than weekly. These equinoctial months were selected to reduce any possible influence of seasonal factors (Christmas and summer holidays) on the editorial process. Random selection of articles from a longer period was felt to be unnecessary as the variable delay between submission and publication would, to some extent, secure the same end.

The review was limited to research papers appearing in the 'Original Research' section of each journal and the 'Clinical Practice' section in *The Lancet*. Each paper appearing in these sections was read by one of the authors (JS), who recorded the reference details, the objective of the study and the research design employed.

Papers reporting on the effectiveness of a health care intervention (evaluations) rather than basic clinical science, epidemiology of disease or health policy were further classified according to the type of comparison group used and the type of outcome measures employed. The term 'evaluation' typically involves a comparison of two or more interventions, but we included studies which described an intervention in isolation in order to assess the extent to which the use of control groups had changed. The studies were divided into three categories: those employing a comparison group; those employing a before-and-after study design (or own comparison group); and those with no comparison group. The range of research designs included randomised controlled trials, non-randomised controlled trials, uncontrolled prospective studies, before-and-after studies, case-control studies, cross-sectional studies, case series and case studies.

Three categories of health outcomes were used: mortality; clinical or other intermediate measures of health outcome; and final outcomes or quality of life.

Intermediate referred mainly to clinical signs and symptoms, but also included changes in risk factors (smoking behaviour, for example) where these were used as proxy indicators for the effectiveness of the intervention. Final outcomes included self-perceived health status and functional capacity, as well as social and emotional health state.

What distinguishes final from intermediate outcomes is their immediate relevance to the patient's sense of well-being (Hollandsworth 1988; Evans & Stoddard 1990; Bowling 1991). The distinction is between measures of goal attainment and measures which predict goal attainment (Read 1993). The objectives of health care are to improve life expectancy and quality of life. Subjective changes in health are defined as 'final outcomes' because they reflect the final aim of health care. Biochemical changes and changes in other clinical signs and symptoms are, at best, intermediate indicators of changes in final health and so were labelled intermediate outcomes. For some conditions, the relationship between the two is close (Veldhuyzen van Zanten et al. 1993). For other conditions, especially the control of hypertension, there is little relationship between clinical dimensions of outcome and the patient's subjective well-being (Kawachi & Malcolm 1991).

However, the distinction between intermediate and final outcome is not always easy to make in practice. Pain, for example, is obviously subjective but is occasionally ranked in evaluative studies by the clinician, with only indirect reference to the patient. For the purposes of this study, where it was not otherwise clear, the categorisation depended on who rated the outcome of interest and how it was scored. So, pain or itching recorded as present or not present by the physician was categorised as a sign or symptom and was therefore considered an intermediate outcome. However, if the severity of pain or the degree of discomfort from itching was rated by the patient, then this was classified as a final outcome.

Classification of the outcome measures used in each study was made independently by two of the authors (JS and either AS or DN). Where there was disagreement or difficulty in classifying the measures used, the opinion of the remaining author was sought.

Differences in the number of papers making comparisons and in the use of final outcome measures across the three journals and over time were tested for statistical significance using chi-square tests of association, or Fisher's exact equivalent where the number of papers was small (Armitage & Berry 1987).

Results

In total, 255 papers were reviewed, 106 from 1982 and 149 from 1992. The number of papers published in the sections considered in this review has increased by over 40 per cent, but the proportion of papers evaluating a health services intervention remained stable at 33–34 per cent (see Figures 1 and 2). Case series and case studies have declined in proportion, while studies reporting on the epidemiology of disease have doubled in number and increased by almost 40 per cent in proportion.

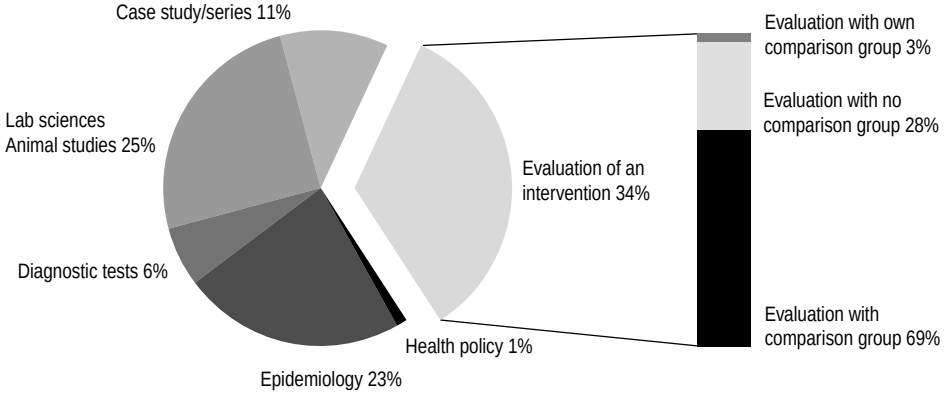

Figure 1: Classification of original research papers, 1982

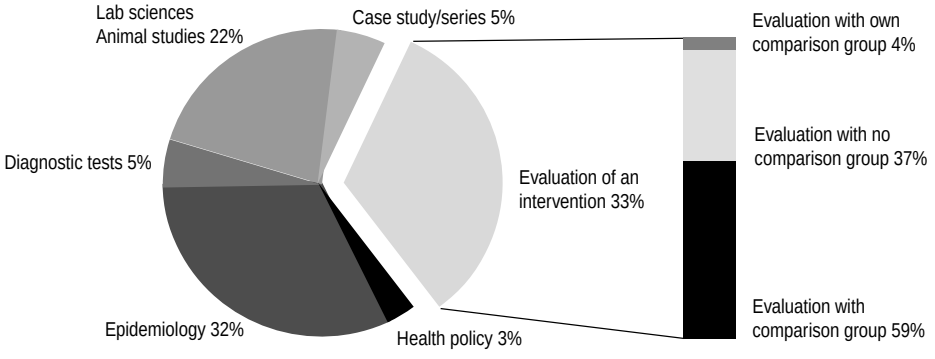

Figure 2: Classification of original research papers, 1992

Table 1: Number of studies by year, journal and type of evaluation

	With comparison group n (%)	With no comparison group n (%)	With own comparison group n (%)	Non-evaluations n (%)	*Total* n (%)
NEJM 1982	8 (33)	2 (8)	(0) (0)	15 (60)	*25 (100)*
NEJM 1992	12 (29)	3 (7)	1 (2)	26 (62)	*42 (100)*
The Lancet 1982	14 (30)	3 (6)	1 (2)	29 (62)	*47 (100)*
The Lancet 1992	13 (27)	6 (12)	1 (2)	29 (69)	*49 (100)*
MJA 1982	3 (9)	5 (15)	0 (0)	26 (76)	*34 (100)*
MJA 1992	4 (7)	9 (16)	0 (0)	45 (78)	*58 (100)*
Total 1982	*25 (24)*	*10 (9)*	*1 (1)*	*70 (66)*	*106 (100)*
Total 1992	*29 (19)*	*18 (12)*	*2 (1)*	*100 (67)*	*149 (100)*

Table 2: Number of studies measuring outcomes by year, journal and type of outcome measure employed[1]

	With comparison		No comparison		Total	
	1982	**1992**	**1982**	**1992**	**1982**	**1992**
NEJM						
Mortality	5	5	0	2	5	7
Intermediate	8	10	2	3	10	13
Final	1	5	0	0	1	5
The Lancet						
Mortality	5	3	1	0	6	3
Intermediate	12	12	3	6	15	18
Final	2	6	1	2	3	8
MJA						
Mortality	0	1	2	4	2	5
Intermediate	3	4	5	5	8	9
Final	1	2	1	2	2	4
Total						
Mortality	*10*	*9*	*3*	*6*	*13*	*15*
Intermediate	*23*	*26*	*10*	*14*	*33*	*40*
Final	*4*	*13*	*2*	*4*	*6*	*17*

[1] Note that some studies have measured more than one outcome.

Table 1 shows that there were marked differences among the journals. In 1982, evaluations constituted 40 per cent of original papers in the *NEJM*, 38 per cent in *The Lancet,* but only 24 per cent in the *MJA*. The position was relatively unchanged in 1992, with evaluations constituting over one-third of papers in the *NEJM* and *The Lancet,* but less than one-quarter of papers in the *MJA*.

There were also marked differences among the journals in the number of evaluative studies which employed a comparison group. In 1982 over three-quarters of the evaluations published in the *NEJM* and *The Lancet* employed a comparison group, but fewer than 40 per cent in the *MJA* did so. The results for 1992 were very similar.

Table 2 shows the types of outcome measure employed. As there were only three before-and-after evaluations, they have been omitted. Three papers (each appearing in the *MJA* during 1992) which reported no outcome data and described only the structure and process of services were also excluded. Overall, there has been little change in the number of studies reporting intermediate outcomes and mortality data.

Table 3: Number of evaluations (with comparison groups) measuring intermediate and final outcomes by year and journal

	With comparison		P values[1]
	1982	1992	
NEJM			
Intermediate only[2]	7	6	p = 0.18
Final[3]	1	5	
The Lancet			
Intermediate only	10	7	p = 0.20
Final	2	6	
MJA			
Intermediate only	2	2	p = 1.00
Final	1	2	
Total			
Intermediate only	19	15	p = 0.04
Final	4	13	

[1] Fisher's exact test
[2] Studies reporting intermediate and not final outcomes.
[3] Studies reporting final and or both intermediate and final outcomes.

Table 2 shows that the proportion of studies reporting final outcomes has almost doubled from 6 per cent to 11 per cent. Table 3 shows that the shift from intermediate outcome measurement over time is significant, though not on a journal-by-journal basis. The change, however, is measured from a very small base and it is therefore impossible to make meaningful comparisons among the journals.

Discussion

Two points stand out from this review. The first relates to study design. It is evident that most evaluations appearing in the *NEJM* and *The Lancet* have employed comparative research designs. Case studies and other uncontrolled research designs no longer predominate (Fletcher & Fletcher 1979a, p 180; Garfield 1979; Najman & Levine 1981). The number of randomised trials remains small but does not necessarily imply any serious criticism. A randomised controlled trial remains the ideal way of evaluating medical practice, but such trials are expensive and, in many circumstances, it is unethical to randomise. Ethical difficulties arise, particularly when the study is concerned with an existing intervention which has never been formally evaluated. In these circumstances, it is more difficult to justify withholding treatment from patients who would form the control group, and more opportunistic methods of evaluation then have to be employed.

Yet it is in the evaluation of existing interventions that the new emphasis on outcome measurement offers the greatest benefit. The need to evaluate new technologies is largely undisputed. The greatest challenge is to assess the effectiveness of existing expensive technologies against zero or placebo interventions. If randomisation is not possible, the real test of the quality of evaluative studies is how the knowledge gained compares with what was known to begin with, not what might be known were it possible to carry out a randomised trial.

What is disappointing is how seldom comparison groups are used in evaluations appearing in the *MJA*. It may be that the better designed studies being carried out in Australia are sent elsewhere for publication, leaving more space in the journal for descriptive studies or basic science. A search for papers written by Australian-based authors in *The Lancet* and the *NEJM* in April and May 1982 and in the same two months in 1992 found one paper published during the period. There is insufficient evidence based on this search to support the hypothesis that better designed Australian-based studies are being published in overseas journals, rather than in Australian journals.

The number of uncontrolled studies reported in the *MJA* may also indicate a higher propensity to exploit existing cross-sectional data. It would be of concern if the low use of comparison groups reflected a greater propensity among evaluators to use poor study design or possibly even a less critical editorial policy. One of the main forces behind the outcomes movement is the desire to make better use of scarce health resources. Equally, therefore, the resources allocated to clinical research must be used to their best effect (Anderson & Evered 1986; Carpenter 1993).

The results of this study suggest a small but not insignificant increase in the number of studies reporting final outcomes. This result needs to be set into its historical context. Fletcher and Fletcher (1979a, p 180) reviewed original articles appearing in the *NEJM*, *The Lancet* and the *JAMA* between 1946 and 1976. That study focused specifically on evaluation methods, but also considered the type of outcome measures employed using a similar classification to the one employed here. In 1976 only 4 of 155 (3 per cent) published articles had used final outcome measures, a substantial fall over the preceding 30 years from the 17 studies out of 151 (11 per cent) found in 1946. The increase in both the number and the proportion of studies using final outcomes between 1982 and 1992 may therefore reflect a return to the levels being reported in 1946, but may equally be an artefact of the smaller number of papers published then and so says little about changes in the perceived importance of outcomes measurement.

One obvious limitation with this study is that it focuses on only three journals. Over the 10-year period reviewed, there has been a substantial increase in the number of specialist medical journals, non-medical journals and journals focusing specifically on quality of life or evaluation. Thus the scope to publish the results of outcomes studies has increased (Krantz 1979). Whether or not evaluations employing good research designs (that is, those involving comparison groups and the use of final outcome measures) have been diverted away from generalist journals reviewed here is an empirical question (Fletcher & Fletcher 1979b, p 1293).

Some evidence in favour of this displacement hypothesis is provided by surveys of the number of articles which cite 'quality of life' or 'health status' as keywords. Between 1975 and 1979, 23 articles with quality of life as a keyword were cited in Medline, seven of which appeared in generalist medical journals (Najman & Levine 1981). Between 1980 and 1984, the number of articles citing quality of life had increased threefold to 69. However, of these, only eight appeared in generalist journals (Hollandsworth 1988). Thus there has been a substantial increase in the reporting of subjective or final health outcomes in other journals.

It is difficult to predict future trends in the publication of health outcomes-based research in Australia; however, there is reason to be optimistic. The New South Wales Health Department's Health Outcomes Initiative is striving towards implementing an outcomes-based approach to planning, delivery and evaluation of health services in New South Wales (Frommer, Rubin & Lyle 1992). A key element of the initiative has been to support the development of health outcome indicators by funding a series of demonstration projects. This should result in the publication of well-designed outcomes-based studies which will not only provide valuable information to assist with the planning of health services in New South Wales, but may encourage further outcomes research.

The new emphasis on measurement of health outcomes has been labelled by Relman (1988) as the third revolution in medicine. Judged by the evidence presented here, the revolution is proceeding very quietly. If outcomes management is to be the force which drives health care systems towards the provision of better quality of care, the measurement and reporting of health status must be more widespread than is evident here (Harvey 1991; Patrick & Bergner 1991). It is expected that some current initiatives in Australia will generate health outcomes information in the next few years. An increase in the number of studies reporting health outcomes in an evaluative setting would be a sure sign that the revolution was truly under way.

References

Anderson J & Evered D 1986, 'Why do research on research?' *The Lancet*, vol 8510, pp 799–800.

Anon. 1992, 'Outcomes and PORTs' (editorial) *The Lancet*, vol 340, p 1439.

Anon. 1993, 'Doctors are asked: Does it work?' (editorial) *Journal of the American Medical Association*, vol 270, p 153.

Armitage P & Berry G 1987, *Statistical methods in medical research*, Blackwell Scientific Publications, Oxford.

Australian Institute of Health and Welfare 1994, *Health outcomes bulletin*, 1.

Bowling A 1991, *Measuring health: A review of quality of life measurement scales,* Open University Press, Philadelphia.

Carpenter L 1993, 'Is the study worth doing?' *The Lancet*, vol 342, pp 221–3.

Clinton JJ 1991, 'From the agency for health care policy and research,' *Journal of the American Medical Association*, vol 266, no 15, p 2057.

Evans GR & Stoddard GL 1990, 'Producing health, consuming health care,' *Social Sciences and Medicine*, vol 31, no 12, pp 1347–63.

Fletcher RH & Fletcher SW 1979a, 'Clinical research in general medical journals. A 30-year perspective,' *New England Journal of Medicine*, vol 301, pp 180–3.

Fletcher R & Fletcher S 1979b (comment) *New England Journal of Medicine*, vol 301, p 1293.

Frommer M, Rubin G & Lyle D 1992, 'The NSW health outcomes programme,' *NSW Public Health Bulletin,* vol 3, no 12, pp 135–7.

Garfield E 1979, *Citation indexing – its theory and application in science, technology and the humanities*, John Wiley, New York.

Grisso JA 1993, 'Making comparisons,' *The Lancet*, vol 342, pp 157–60.

Harvey R 1991, *Making it better. Strategies for improving the effectiveness and quality of health services in Australia*, National Health Strategy Background Paper No 8, Melbourne.

Hollandsworth JG 1988, 'Evaluating the impact of medical treatment on the quality of life: A 5-year update,' *Social Sciences and Medicine,* vol 26, no 4, pp 425–34.

Kawachi I & Malcolm L 1991, 'The cost effectiveness of treating mild to moderate hypertension: A reappraisal,' *Journal of Hypertension*, vol 9, pp 199–208.

Krantz K 1979 (comment) *New England Journal of Medicine*, vol 301, no 23, pp 1292–3.

Leeder SR 1993, 'The recent Medicare agreement, goals and targets, and health outcomes,' *Australian Journal of Pubic Health*, vol 17, pp 87–8.

Long A, Bate B & Sheldon TA 1992, 'Establishment of a UK clearing house for assessing health services outcomes,' *Quality in Health Care*, vol 1, pp 131–3.

Najman J & Levine S 1981, 'Evaluating the impact of medical care and technologies on the quality of life: A review and critique,' *Social Sciences and Medicine*, vol 15F, pp 107–15.

Nutbeam D, Wise M, Bauman A et al. 1993, *Goals and targets for Australia's health in the year 2000 and beyond*, Australian Government Publishing Service, Canberra.

Patrick DL & Bergner M 1991, 'Measurement of health status in the 1990s', *Annual Review of Public Health,* vol 11, pp 165–83.

Read JL 1993, 'The era of quality of life assessment,' *in* SR Walker & RM Rosser (eds) *Quality of life assessment. Key issues in the 1990s*, Kluwer Academic Publishers, United Kingdom.

Relman AS 1988, 'Assessment and accountability: The third revolution in medical care,' *New England Journal of Medicine*, vol 319, pp 1220–2.

Veldhuyzen van Zanten SJ, Tytgat KM, Pollak PT et al. 1993, 'Can severity of symptoms be used as an outcome measure in trials of non-ulcer dyspepsia and helicobacter pylori associated gastritis?' *Journal of Clinical Epidemiology*, vol 46, pp 273–9.