Outcomes Measurement in Psychiatry

A critical review of patient based outcomes measurement in psychiatric research and practice

by

Dr Simon Martin Gilbody

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

University of York
Department of Health Sciences and Clinical Evaluation

26th September 2001

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

Background

Outcomes are measured to establish *what works*, in the context of evaluative research, and to improve the *quality of care* that is offered. Traditional outcomes focus upon biomedical endpoints, but there is an increased interest in *patient based outcomes*, which measure the impact of illness or healthcare interventions on the individual and how they live their day to day life. There are reasons to expect that the application of patient based outcomes would be especially relevant to the discipline of psychiatry.

Aims

To explore the measurement of outcome in psychiatric research and practice, with particular reference to patient based outcomes.

Methods

1. A critical literature review of the *outcomes movement* in health care.
2. A survey and systematic review of the methods used to measure outcome in evaluative psychiatric research (randomised trials and outcomes research)
3. A survey of the use of outcomes measures by UK psychiatrists in their day-to-day practice.
4. A systematic review of the effectiveness of routine outcomes measurement in improving the quality of care for those with common psychiatric disorders.

Results

An outcomes movement has emerged in healthcare, which can be understood in social, political and economic terms.

Outcomes measurement in psychiatric research is dominated by the measurement of psychiatric symptoms, with little reference to patient based measures.

Practising UK psychiatrists rarely measure outcomes. There are substantial practical and attitudinal barriers to the use of outcomes instruments in NHS mental health services.

There is little evidence to support the potential for routine outcomes measures to improve the quality of mental healthcare.

Discussion

Current mental health policy places great emphasis on the measurement of outcomes, and is likely to fail. The potential for patient based outcomes to be adopted in psychiatric research and practice has yet to be realised. The need for important research into the suitability and value of patient based outcomes measures in mental health research and practice is identified.

Acknowledgement

The work described in this thesis was conducted during the tenure of a training fellowship in health services research funded by the Medical Research Council, to whom I am indebted. I am grateful for the time and support given by my supervisors; Professors Trevor Sheldon and Allan House. I am also indebted to: Professor Jos Kleijnen and the staff of the NHS Centre for Reviews and Dissemination and Professor Ian Russell and the staff of the Department of Health Sciences and Clinical Evaluation. Lastly, I thank Cathy and my boys.

Abbreviations

AHCPR ..... Agency for Health Care Policy and Research

BDI ..... Beck Depression Inventory

BPRS ..... Brief Psychiatric Rating Scale

CORE ..... Centre for Outcomes Research

DSM ..... Diagnostic and Statistical Manual

FFS ..... Fee for Service

FSQ ..... Functional Status Questionnaire

GAF ..... Global Assessment of Functioning

GAS ..... Global Assessment Scale

GHQ ..... General Health Questionnaire

HIE ..... Health Insurance Experiment

HoNOS ..... Health of the Nation Outcome Scale

HRQoL ..... Health Related Quality of Life

HDRS ..... Hamilton Depression Rating Scale

MMSE ..... Mini Mental State Examination

MOS ..... Medical Outcomes Study

NHS ..... National Health Service

NNT ..... Number Needed to Treat

PANSS ..... Positive and Negative Symptom Scale

RCT ..... Randomised Controlled Trial

PORT ..... Patient Outcomes Research Team

PRN ..... Practice Research Network

PTSD ..... Post Traumatic Stress Disorder

RR ..... Relative Risk

SF36 ..... Short Form 36

SIP ..... Sickness Impact Profile

QALY ..... Quality Adjusted Life Year

Publications

The following peer reviewed publications have resulted directly from the work presented in this thesis:

Gilbody, SM, House, AO, Sheldon, TA (2001) Routinely administered questionnaires for depression and anxiety: a systematic review *British Medical Journal* **322**, 406-409.

Gilbody, SM, House, AO, Sheldon, TA (2001) Routine outcomes measurement and needs assessment for schizophrenia and related disorders (Cochrane Review). In: *The Cochrane Library,* Oxford, Update Software (in press).

Gilbody, SM, House, AO, Sheldon, TA (2001) Routine outcomes measurement for depression and anxiety (Cochrane Review). In: *The Cochrane Library,*. Oxford, Update Software (in press).

Gilbody, S. M. & Whitty, P. A. (2001) Improving the delivery and organisation of mental health services: beyond the conventional RCT. *British Journal of Psychiatry*, (in press).

Gilbody, SM, House, AO, Sheldon, TA (2001) Outcomes research in mental health - A systematic review *British Journal of Psychiatry* (in press).

Gilbody, SM, House, AO, Sheldon, TA (2001) UK psychiatrists don't use outcomes measures – A national survey [editorial] *British Journal of Psychiatry* (in press).

Gilbody, SM & House, AO (1999). Variations in Psychiatric Practice: neither unacceptable nor unavoidable, but only under researched [editorial]. *British Journal of Psychiatry* **175**: 303-305.

Gilbody, SM & Petticrew, M (1999). Rational decision making in mental health: the role of systematic reviews in clinical and economic evaluation. *Journal of Mental Health Policy and Economics* **2: 99-107**.

Table of contents

List of tables

List of figures

Overview

The past 30 years have seen a rise in interest in the measurement of the outcomes of medical care – to the extent that an 'outcomes movement' has been described, which has been labelled 'a [third] revolution in healthcare' (Relman, 1988). A feature of this outcomes movement is an interest in the measurement of outcome from the patient's perspective, with attempts to measure the impact of healthcare and illness on the individual, in terms of how they live their day-to-day life. Clinicians are urged to measure these outcomes of patient care in the context of their day-to-day practice – in a quest to improve the quality of this care.

The purpose of this thesis is to explore the measurement of outcome within the speciality of psychiatry, with particular reference to patient based outcomes, and the use of outcomes measurement in the context of **both** clinical practice and evaluative research. The research presented within this thesis forms four distinct, but inter-related sections that are summarised below.

Section 1 – Outcomes measurement in healthcare

The measurement of outcome, and in particular patient-based outcome has been studied by a variety of disciplines and from a number of perspectives. The reasons for the increase in emphasis in outcome measurement are social, political and economic. A pre-requisite to a study of the measurement of outcome in psychiatry is an understanding of these perspectives and an examination of the often-confusing terminology used in the measurement of outcome in general.

The first section aims to:

- Clarify the terminology and main theoretical perspectives that are used in discussing outcome and patient based outcome.
- Provide a brief review of the origins of the outcomes movement in wider healthcare.

The method adopted in this section of the thesis is a narrative overview or scoping review, summarising a diverse and disparate literature examined in preparation of the thesis.

Section 2 – Outcomes measurement in contemporary psychiatric research

The *outcomes movement* in healthcare has laid emphasis on the methods used to measure outcome, such that it is more patient-centred and reflects not just biophysical markers of disease, but wider 'quality of life'. The degree to which patient-based outcomes measures have been used in psychiatry is not known, and neither is the value of this approach when compared to traditional outcomes measurement.

The second section aims to:

- Describe methods used to measure outcome in contemporary psychiatric research, with special reference to clinical trials and outcomes research.

The methods adopted in this phase of the thesis are empirical survey and systematic review.

Section 3 – Outcomes measurement in psychiatric practice

Whilst the previous sections focus on outcome measurement in evaluative research, a cornerstone of the outcomes movement is the encouragement of clinicians to measure the outcome of their patients in the context of routine day-to-day care, in order to improve the quality of this care. This emphasis has been made explicit in recent UK mental health policy statements (Secretary of State for Health, 1999). The degree to which this approach has been adopted by clinicians has yet to be charted in UK mental health services. Similarly, the measurement of outcome for each and every patient represents a health care intervention or 'technology'. The adoption of this approach by the UK National Health Service will consume substantial resources, and should therefore be justified by an empirical demonstration of some benefit, to both the clinician and the patient; i.e. it should be both clinically and cost effective.

The third section aims to:

- Examine the actual use of standardised outcomes measure by practising UK clinicians, and to identify barriers and advantages to their use.
- Establish the research evidence that exists to support the use of routine outcome measurement in the day-to-day care of patients with mental health problems.

The methods adopted in this phase of the thesis are cross sectional survey and systematic review:

- A cross sectional survey of practising UK psychiatrists will be presented.
- A systematic review of randomised studies will be presented to examine the value of the implementation of routine outcomes measurement in improving the quality of healthcare for those with mental health problems.

Section 4 – Discussion of the thesis

The final section will draw together the questions and research findings of sections 1-3. The findings of the thesis will be discussed with reference to research, practice and policy within psychiatry.

Section 1 – Outcomes measurement in healthcare

Chapter 1 Introduction to the review of outcomes measurement in healthcare

The past 30 years have seen a rise in interest in the measurement of the outcomes of medical care – to the extent that an 'outcomes movement' has been described, which has been labelled 'the third revolution in healthcare' (Relman, 1988). A feature of this outcomes movement has been an increased interest in the measurement of outcome from the patients' perspective, with attempts to measure the impact of healthcare and illness on the individual, in terms of how they live their day-to-day life. Terms such as 'health status' and 'quality of life' have entered medical vernacular and instruments have been developed with which to measure these constructs as actual outcomes of medical care ('patient based outcomes measures'). This thesis sets out to explore in detail the measurement of outcome within psychiatry, with particular reference to how and in what way psychiatry has adopted a more *patient based* approach to outcome measurement in clinical practice and research. However, any examination of this topic within psychiatry requires a knowledge and understanding of the core concepts, areas of controversy and debate and methods used in the measurement of outcome in wider healthcare. Section 1 of this thesis presents such an overview.

Aims of the review

The aims of the introductory overview were to outline:

- The historical origins of outcomes measurement in healthcare;
- The methods that have been employed in measuring outcome in healthcare, with special reference to patient based outcomes measurement;
- The uses to which outcomes measures have been put in healthcare;
- Possible social and political explanations for any increased interest in outcomes measurement.

Chapter 2 Review method

The introductory overview that follows summarises a broad body of literature surveyed in a scoping or conceptual review.

The *scoping review* borrows and adapts methods that are used in the conduct of a systematic review – such as the use of electronic literature searches. However, the term systematic review implies the use of specific techniques to identify and summarise an empirical literature in order to answer specific research questions (or hypotheses) (NHS Centre for Reviews and Dissemination, 2000). The use of a scoping review in this context and the methods employed will now be discussed in more detail.

Conceptual review versus systematic review

Reviews of the literature now increasingly adopt the methods of 'systematic review' – where a systematic review has been defined as one which seeks to *'locate, apprise and synthesise evidence from scientific studies in order to provide empirical answers to scientific questions'* (NHS Centre for Reviews and Dissemination, 2000). This method will be described and used in subsequent sections of the present thesis, however, the primary purpose of the introductory review was to produce a broad overview of outcome measurement in wider healthcare. The method of systematic review was not appropriate for this topic for three main reasons: Firstly, because a different literature from that normally synthesised in systematic reviews was the main focus of this study - i.e. textbooks, opinion pieces and traditional narrative review articles. Secondly, the topic of study was not one that can be considered a 'scientific question' to which 'empirical answers' can be sought. Rather the purpose of the study was to produce a coherent narrative that summarises a large body of literature and provides (and demonstrates) a greater understanding of a topic. Thirdly, it was impossible to review all the literature that has ever been written on 'outcome measurement' in all areas of healthcare within the time frame of the study (see on for a fuller discussion of this issue).

Core issues addressed in the review

A core body of research was consulted in deciding upon the scope of the review. This consisted of approximately 200 key articles and key textbooks, known to the author and supervisors. The following topics were identified as being key to gaining an understanding of how and why outcome is measured in wider healthcare, and in understanding what methods might be used in measuring outcome in psychiatry (the topic of the thesis).

- What (if any) are the historical precedents in the measurement of outcome?
- Why has there been a rise in the prominence of outcome measurement in recent years?
- What is meant by outcome, when this term is used in the medical literature?
- What is meant by 'measurement', when this term is applied to outcomes?
- What are 'patient based' outcome measures?
- What is the origin of the measurement of patient based outcomes?
- What is meant by common synonyms for patient based outcome measurement – such as quality of life and health status?
- What specific types of patient based outcome measurement instrument are there?
- To what uses are outcome measures put?

Having stipulated the core areas of the scoping review – further literature was sought to examine each of these areas in more depth. An inclusion criterion for articles was broadly defined as being directly relevant to any one of the topics outlined above. The bulk of this literature was expected to be review articles and opinion pieces from journals. Primary empirical research was only ordered and judged for inclusion if it was likely to discuss the measurement of outcome in some depth and some substance – rather than just reporting how outcome was measured within a specific study.

Literature search methods

Systematic reviews explicitly set out to *'locate, apprise and synthesise evidence from scientific studies in order to provide empirical answers to scientific questions'* (NHS Centre for Reviews and Dissemination, 2000). In the first of these functions, locating research, electronic literature searches are often employed as the most efficient method of locating literature.

It was decided at an early stage that the explicit electronic search methods of a systematic review were not appropriate for the literature review presented in the introductory section (section 1). The primary reason for this being that the volume of literature on the topic of 'outcome measurement' and 'patient based outcome measurement' which is identified through electronic literature searches is so vast as to make it impossible to view all relevant literature and screen it for inclusion within the review. Similarly, the ubiquitous nature of likely search terms (such as outcome) in primary research articles would have resulted in large volumes of literature being identified that are of no relevance to the review – i.e. electronic searches will lack 'specificity' or 'precision' (McKibbon & Walker-Dilks, 1995),.

For example, a Medline search for the years 1966-99 using the term 'outcome' yields 192,570 references, which are likely to contain large volumes of literature, which are of no direct relevance to the topic. Literature searches using more specific terms, such as 'quality of life' yield similar volumes of literature (36,905 individual references). Consultations with an experienced information officer were similarly unable to refine search strategies such that they were able to produce manageable volumes of literature, which could be usefully screened within the time frame of the research programme. Thus more specific methods to identify key papers of relevance to the scoping literature review were adopted from the outset and are identified below.

'Bootstrapping' from reference lists

From the reference lists of known papers and books, a further 246 references of direct relevance as source material were identified. This was the most productive in terms of identifying important literature for the review.

Searches using author names

Some authors (for example John Ware, author of the Short Form 36, and Avedis Donabedian, author of influential theoretical articles on outcome measurement) have published extensively on the topic of outcome measurement. A search using 10 key authors, yielded a further 60 articles for inclusion, in addition to those identified above.

Hand searching of key journals

Some journals have published relatively large numbers of articles discussing the measurement of outcome. These include the Journals, Medical Care, Quality in Health Care; Quality of Life; Journal of Clinical Epidemiology (formerly Journal of Chronic Diseases). These journals were hand searched over the past 20 years, and a further 72 articles, in addition to those identified above were identified.

Synthesis of the literature

A narrative overview of the main themes and issues revealed in the literature is presented. This document does not summarise or include every argument or theoretical viewpoint expressed. Rather it is intended to summarise the major arguments and positions that are advanced. The narrative nature of the literature does not lend itself to a quantitative summary in the way in which research is often presented in systematic reviews.

Chapter 3 Historical precedents in the measurement of outcomes

The term 'outcome', in its contemporary use can be traced back to Donnabedian, who presented a tripartite evaluation of healthcare: *structure, process* and *outcome*. He defined *health outcome* as;

'....a change as a result of antecedent healthcare' (Donabedian, 1966)

And identified the need for,

'the improvement of methods for identifying key features of medical care that are associated with favourable outcomes, so that these features can be preserved despite the constraints imposed by an increasingly cost conscious healthcare environment'

Several writers have commented that this focus was nothing new (e.g. Brookes, 1995; Lohr, 1988; Scroeder, 1987). What Donabedian was in fact reflecting was a resurgence of attention to the *results* of medical care. For example, Davies, *et al.*, (1994) suggest that;

'For generations we have used indicators of mortality, morbidity and expenditure when describing and evaluating the performance of individual clinicians, provider groups, hospitals and healthcare organisations, and the healthcare system in general. We have measured, tracked, reported and often attempted to alter rates of death, disease and expenditure.'

Important historical contributions to the measurement of health outcome and quality improvement from Florence Nightingale working in the Crimea, Ernest Codman in Boston , J Allison Glover in the UK, are identified by a number of authors (Brookes, 1995; Donabedian, 1989; Eisele, *et al.*, 1956; Rosser, 1983; Rosser, 1993; Scroeder, 1987).

Nightingale, upon her return from the Crimea described her concern with the quality of care in hospitals, the need for hospital data and the importance of cost effectiveness and accountability. In her Notes on Hospitals (Nightingale, 1863) she writes,

'I am fain to sum up with an urgent appeal for adopting this or some uniform system of publishing the statistical records of hospitals. There is a growing concern that in

all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life. In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purpose of comparison. If they could be obtained, they would enable us to decide many other questions besides the ones alluded to. They would show subscribers how their money was being spent, what amount of good was being done with it or whether the money was not doing mischief rather than good'

She recommended that for individual patients, their outcome of care be classified as *dead*, *relieved* or *unrelieved*. Further, Florence Nightingale was the first to achieve changes as a result of her outcome measurement, for example she influenced hospital design and her 'accounting methods' were adopted by many teaching hospitals and some continued to use them until the Hospital Activity Analysis was introduced in the 1950s (Rosser, 1983).

Ernest Codman, a Boston surgeon working at the beginning of this century advocated ideas that predate Donabebian's idea of outcome, which he termed the 'end result' (Codman, 1914; Donabedian, 1989). Further, he advocated a process that we would now recognise as 'quality assurance' and 'audit'. The 'end result' idea was that a hospital should follow every patient long enough to determine whether treatment had been successful or not, and to question the adequacy of care given to those with an unsuccessful 'end result'. His system involved analysis of each patient's diagnosis, treatment and results in the years subsequent to inpatient intervention. This allowed representative 'efficiency boards' to redirect policy, organisation, and operation of the hospital into more efficient channels. Positive outcome or 'end result' was specified as a 'satisfied or relieved patent'. He believed that his end result data would be useful in monitoring quality, advocating clinical science, establishing accountability and allocating resources.

In a similar vein, J Allison Glover (Glover, 1938; Glover, 1948), a community medical officer working in the UK in the 1930's observed massive variations between doctors and between geographical areas in the use of *tonsillectomy* for school children. Aside from documenting important differences in clinical practice, he was able, through the use of population morbidity statistics to demonstrate that higher operation rates made little impact on the natural history and outcome of childhood

middle ear disease. Glover's work predates that of Jack Wennberg in the USA (Wennberg & Gittelsohn, 1973; Wennberg & Gittelsohn, 1982), who studied *small area variations in clinical practice* and developed 'outcomes research' methods to study the consequences of these variations (Wennberg, 1991; Wennberg, *et al.*, 1980). The importance of Wennnberg's work as a cornerstone of the 'outcomes movement' in the US will be discussed in more detail below.

More recently, but still nearly half a century ago, Paul Lembke, an early US health services researcher who pioneered the use of audit in the evaluation of surgical care stated that 'the best measure of quality is not how well or how frequently a medical service is given, but how closely the result approaches the fundamental objectives of prolonging life, relieving distress, restoring function and preventing disability'. (Lembke, 1952)

The conclusion that must be drawn from these selective historical examples is that the measurement of outcome (as it is defined by Donnabedian) extends back further than the past 20 years. An appreciation of the primacy of outcome over, for example, measures of process is seen. Further, outcome is measured with a purpose which we will see mirrors many contemporary themes - such as the improvement of the quality of healthcare; increasing the accountability of those who provide healthcare and increasing the relevance with which outcome is measured.

What is less clear is why outcomes measurement came to popularity and developed the status of a 'movement' in the latter part of the twentieth century, rather than in the times of Codman and Nightingale. The ideas of Nightingale had only a limited effect on health planning and policy in the UK, since it was not until the 1950s that a more complex system of recording hospital outcomes was introduced (Rosser, 1983). However, it is clear that her innovations were not further developed, or instituationalised. Similarly, Codman's ideas were not adopted in the US. In fact his pursuit of the 'end result' made him an outcast in the medical circles of turn of the century Boston and for his efforts, he was expelled form his post as chief of staff (Donabedian, 1989).

A wider explanation of this recent rise in terms of social and political forces is therefore required. The reasons why an increased interest in outcome measurement has come about more recently will now be considered in some detail.

Chapter 4 What has stimulated the rise in outcome measurement?

Given the historical precedents that are seen in the measurement of outcome in medical care, it is sensible to ask why the increased interest in outcome measurement has come about now, rather than, say earlier this century. Several overlapping drivers of the 'outcomes movement' are seen - which reflect social, political and economic changes within society as a whole and within healthcare in particular. This section provides an overview of some of these changes, both from a UK and North American perspective.

Effectiveness and efficiency

The past 25 years has seen a revolution in the way in which healthcare has been evaluated. The publication of Archie Cochrane's *'Effectiveness and Efficiency'* (Cochrane, 1972) heralded a growing interest in the importance of establishing 'what works' (through the use of rigorous evaluative studies such as the randomised controlled trial) and the use of only the most effective treatments, such that maximum health benefit can be obtained within given resources. The clear message of Cochrane is that in order to determine effectiveness, we need realistic measures with which to judge the success or otherwise of healthcare interventions and programmes. Although Cochrane does not explicitly refer to outcomes, the work is a plea for the need to measure outcomes (Opit, 1990).

The drivers of *effectiveness and efficiency* have been political and economic (Doessel & Marshall, 1985; Epstein, 1990; Opit, 1990). The most conspicuous of these forces have been within the US healthcare system and include: (i) attempts to address the rising costs of healthcare,(ii) the demonstration of massive variations in clinical practice, and, (iii) the drive to improve the quality of healthcare (Brookes, 1995). Further, there have been several landmark high cost initiatives that have facilitated the move towards outcome measurement. These will be outlined below with illustrative examples and parallels will be drawn between the US and UK healthcare system.

Rising costs of United States healthcare and healthcare reform

The US healthcare industry is amongst the most costly in the developed world. Escalating costs during the 1960s brought little extra benefit in terms of

corresponding improved healthcare, while significant proportions of the population continued to be excluded from receiving healthcare, through making it prohibitively expensive (Aday, *et al.*, 1998; Fuchs, 1974; Milio, 1983). The healthcare system was officially declared to be in crisis during a presidential address in 1969 (Ellwood, 1988). In response to this, various healthcare reforms were proposed, which aimed to make US healthcare more universal in its coverage, more effective and more affordable (Thier, 1992).

Efforts to contain the rising costs of healthcare were openly proposed and the measurement of outcome had several functions within this process. Epstein (1990) identifies three important factors that led to an increased emphasis on outcomes. *Firstly*, payers of healthcare costs were determined to find out 'what works' (demonstrate effectiveness), such that they might cease to reimburse procedures that were ineffective, and to eliminate unnecessary and unexplained variations in medical practice. *Secondly*, changes in health coverage and reimbursement mechanisms generally meant a reduction in the volume and scope of medical care (in addition to that above) available to certain portions of society. *Outcomes measures* were required to provide a monitoring system aimed not so much at improving the quality of care, but to identify and monitor the adverse consequences of reforms. *Thirdly*, the setting up of collective providers of healthcare (such as Health Maintenance Organisations - HMOs) within the market based US healthcare system produced competition. HMOs were required to be able to demonstrate the clinical and cost effectiveness of their services to prospective purchasers of their healthcare (individuals and collective employee insurance schemes).

A useful illustration of the major research initiatives prompted by the healthcare reforms of the 1970s is the US Health Insurance Experiment (HIE) carried out by the RAND Corporation (Brook, *et al.*, 1983). The HIE was conceived in the 1970s to evaluate the potentially adverse consequences of cost containment strategies, such as user charges and other health payment systems (Wright, 1994b). Population samples of 4000 people were enrolled at six sites and were followed up for two years. The study showed that those enrolled in co-payment schemes made a third less health visits and were hospitalised a third less often than those receiving healthcare free at the point of delivery. In order to evaluate whether reduced healthcare utilisation was detrimental to health, a series of *outcome instruments* were developed which measured self assessed general health (physical functioning,

role functioning, mental health, social contacts, & health perceptions). The main results of the study were that for the majority of people, free care did not improve health status. However, co-payment mechanisms deterred a significant portion of the most disadvantaged patients from seeking care, to the detriment of their health – as evidenced by the comprehensive range of outcomes measures which were used to measure this. Of significance was the development of a battery of outcomes measures, which were subsequently refined and further developed in the Medical Outcomes Study (MOS) of the 1980s, and from which developed the now widely used Short Form 36 (SF-36) health status instrument (Ware & Sherbourne, 1992).

Improving the 'quality' of healthcare in the United States

The outcomes movement has also been inextricably linked with the drive to appraise and improve the quality of medical care (Aday, *et al.*, 1998; Epstein, 1990). One manifestation of this has been the initiation of programmes such as 'Quality Assurance', the aims of which are succinctly described by Brook & Lohr (1985) as being *'to improve healthcare in terms of outcome, functional ability, patient well-being and consumer satisfaction and the use of resources by shaping health policy and practice'.*

The measurement of outcome in the pursuit of 'quality' represents a significant advance, since quality had hitherto been judged by the activities of clinicians, without reference to how these impacted on their patients. For example, Makover, working in the 1940s and early 1950s attempted to determine the quality of care offered by clinicians working for the the Health Insurance Plan (HIP) of New York. Makover sought *'to determine the quality of the end product – the actual medical services rendered – on the basis of clinical performance.'* (Makover, 1951). However, Makover's conception of clinical performance was the frequency with which procedures were carried out, rather then the impact of the procedure on the individual.

The work of Makover dominated the methods by which healthcare was evaluated for the best part of the 1950s and early 1960s. According to Doessel & Marshall, (1985), the methods of Makover represented an *'indifference to the importance of medical care outcomes'.*

A contemporary of Makover's states that *'valid instruments for measurements of end results of medical care, such as amelioration of suffering and promotion of well-being are unavailable'* (Rosenfeld, 1957). Thus in the absence of such measures, quality was largely measured by (what Donabedian would later call) process.

The contribution of Donabedian in separating *structure, process and outcome* should be seen in the context of a number of theorists who in the 1960s and 1970s decided to reformulate how quality should be conceptualised and measured. One such approach was that of 'Health Accounting' advocated by Williamson (1978), which adopted the methods of 'financial accounting' in pursuit of an improved 'end result'. Another key theorist was Robert H Brook (Brook & Appel, 1973), who called for an *epidemiology of medical care*; meaning the systematic investigation of the linkages between the components of medical care and patients' outcomes (Brook & Lohr, 1985). However, the work of Avedis Donabedian has been the most influential in facilitating the use of outcomes to evaluate the effects of healthcare practice and policy. Donabedian (1966) first offered a categorisation of the quality of medical care in terms of *structure, process and outcome*.

Structure refers to the resources that are made available for medical care: in physical terms these are medical and other personnel, hospitals, clinics and technologies of all kinds. Structure also encompasses the ways in which resources are provided, including the organisation and the differing aspects of finance, and the skills and training of individuals. *Process* concerns the ways in which the structure is used in diagnosis and treatment, including the patient's own activities in seeking treatment, and the sorts of interventions that are offered and the way in which they are delivered. *Outcome* (as defined by Donabedian) focuses on the results of medical care processes for the health status of individuals and populations as the consequences of using the structure of resources. The influential framework proposed by Donabedian has had a number of consequences, including the raising of the profile of outcomes, when quality had previously been judged solely by process, or even structure in terms of equipment available and staff ratios etc.

An illustrative example of the influence of Donabedian in the measurement of quality within the reformed US healthcare system is the way in hospital performance was measured. The US Joint Commission on Accreditation of Hospitals (JCAH) has for a number of years been charged with the maintenance of quality in hospitals

providing care for HMOs (including the federal Medicare programme), and has traditionally provided indicators of quality based upon indices of hospital activity. In the 1980s the JCAH took the controversial much-criticised step of rejecting these *process* based measures in favour of publishing *outcomes* in the form of mortality data in individual hospitals (O'Leary, 1987; Scroeder, 1987).

Within the same vein, there have been initiatives to judge the 'appropriateness' of healthcare. Appropriateness has been variously defined and the following example given by the RAND Corporation illustrates the centrality of outcome:

'Appropriate care means that the expected health benefit (i.e. increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (ie mortality, morbidity, anxiety of anticipating the procedure, misleading or false diagnosis) by a sufficiently wide margin that the procedure is worth doing.' (Kahn, *et al.*, 1988)

Clearly the measures of benefit and dis-benefit implicit in judging approriateness in the above consideration require broad-based measures of outcome. Again the RAND Corporation, which has undertaken the Health Insurance Experiment (HIE) and Medical Outcomes Study (MOS), has also pioneered techniques to judge the 'appropriateness' of various interventions (Brook & Lohr, 1985).

Addressing variations in medical practice

Linked to efforts to improve the quality of US healthcare has been the demonstration of massive regional variations in the rate and indications for various common medical procedures (Aday, *et al.*, 1998). The implicit assumption is that unexplained variation in practice represents a poor quality and inappropriate healthcare. Wennberg and his colleagues have pioneered the *small area variation* study and have demonstrated regional variation in many common medical and surgical procedures (Wennberg & Gittelsohn, 1973; Wennberg & Gittelsohn, 1982; Wennberg, 1990). Most famously, Wennberg and colleagues demonstrated massive and inexplicable variations between New Haven and Boston in the frequency with which coronary artery bypasses and carotid endarterectomies were offered.

Their supposition is that both over and under use of some procedures must represent 'inappropriate' care for a large portion of patients; ie errors of omission (failing to do necessary things) and errors of commission (doing unnecessary things). Wennberg's research has had a number of direct implications in terms of raising the profile of 'outcomes measurement', particularly in the US, and deserves further discussion.

Firstly, Wennberg has asserted that it is not sufficient to demonstrate practice variation, but it is also necessary to study the (potentially adverse) outcomes of these variations on patients – both in terms of traditional indices of outcome, such as mortality, but also in terms of the impact of illness on the individual in terms of the quality of life and health status (Wennberg, 1991; Wennberg, *et al.*, 1993). Secondly, *outcomes research* has used large-scale insurance claims databases as a source of primary data. This has in turn influenced the type of data that is routinely collected - making it more 'patient centred' (Wennberg, 1991). Lastly, Wennberg's work captured the political imagination of the time and culminated in the creation of the Agency for Health Care Policy and Research (AHCPR) and the funding of large scale 'outcomes research' into common medical conditions (Patient Outcome Research Teams - PORTs) (Wennberg, *et al.*, 1993). In 1989, the US Congress passed the Patient Outcome Research Act, which called for the establishment of a broad based, patient centred outcomes research programme. The research programme was allocated resources of $6 million in its first year, rising to $63 in 1991, with the purpose of using routine outcomes data to determine *'outcomes, effectiveness and appropriateness of treatments'* (Anderson, 1994)..

The work of Wennberg was cited as central in the movement towards outcomes measurement, distilled in Paul Elwood's Shattuck lecture and manifesto of 'outcome's management' (Ellwood, 1988) that came to be called the 'third revolution in healthcare (Relman, 1988) - an era where there is '*consensus on the need for accountability and for the assessment of outcome'.*

Ellwood (1988) describes the US healthcare system as:

'...an organism guided by misguided choices; it is unstable, confused and desperately in need of a central nervous system that can help it cope with the complexities of modern medicine.'

The central nervous system he proposes is the measurement of outcome and the integration of these measures into health management strategies - 'outcomes management'.

UK healthcare system

Whilst the US healthcare system has provided much of the impetus behind the current interest in outcome in general, the UK healthcare system provides some interesting parallels and differences to that outlined above. Whilst Archie Cochrane's Effectiveness and Efficiency (Cochrane, 1972), was cited by Wennberg & Gittelsohn (1973) as of importance and influence to US thinking in the 1970s, it is clear that this text essentially evolved from the tradition of socialised medicine in the UK (Opit, 1990).

The socialised model of healthcare enshrined within the UK National Health Service did not appear to suffer so much the problems of escalating cost and ensuing the financial crises that bedevilled the US healthcare system in the 1960s and 1970s. The primary reason being that it was (and is) a cash limited system (Klein, 1995). However, there has been a growing awareness that the system can be more efficient i.e. we can do not just 'more', but we can deliver 'better' healthcare, within given resources. Further, the arrival of various high cost innovations and treatments has forced a debate about how the NHS should respond and how the value of these treatments to patients and society should be determined (Sheldon & Faulkner, 1996). The language of the UK NHS has been less about cost containment and more about how to measure health gain and cost effectiveness (Klein, 1995). Aside from the political differences between the UK and US, the practical and organisational differences between the two healthcare systems have meant that the raw data which has enabled the US 'outcomes movement' to grow, would not be readily available to researchers or developers in the UK. For example, the absence of large claims databases in the UK would make PORTs and outcomes research, as proposed by workers such as Wennberg impossible to apply generally (Black, 1999). However, the rising awareness of the importance of 'outcome measurement' has none the less come about within the UK as evidenced by a number of initiatives

The government's 1989 white paper *Working for Patients* (Department of Health, 1989) explicitly called for the collection of 'outcomes', particularly within the remit of medical audit. The political emphasis of *Working for Patients* was one that

highlighted concepts such as *choice, quality* and *standards*. Associated initiatives, such as the *Health of the Nation Strategy* and the *Patients' Charter* gave a greater explicit role for patient and consumer involvement in healthcare decision making. This emphasis is in line with the rise in 'consumerism' within society as a whole (Klein, 1995). A role for outcome was also proposed in the 'market healthcare' emphasis of the white paper. For example, the purchaser-provider split allowed health authorities to think more about the needs of populations and how they should respond to these needs in their purchasing decisions (Jordan, *et al.*, 1998). In common with the US healthcare system, a value for money rhetoric was adopted. It was proposed that providers of healthcare would seek to demonstrate the effectiveness and value of their services thorough the measurement of outcome, and that 'purchasing by outcome' would come about (Beckingham, 1994).

Various initiatives can be seen to have arisen in response to the developing UK 'outcomes agenda' - some of which predate the white paper and reflect an already emerging resurgence in outcome measurement. For example, confidential enquiries have been established into peri-operative deaths and suicides. These initiatives would fall under the broad approach of the examination of 'outcomes'. The Clinical Accountability Service, Planning and Evaluation (CASPE) study was initiated at the Freeman Hospital in Newcastle (Bardsley & Coles, 1992). The Department of Health commissioned a national centre - the UK Clearing House for the Assessment of Health Outcomes – which was charged with raising of the profile of 'outcome measurement' in general and US developed measures (such as the Short Form 36) in particular (Long, *et al.*, 1993). Various Royal Colleges responded *to Working for Patients* and the rise in profile of 'outcomes measures' by setting up workshops on how and why 'outcome' should be measured (Hopkins & Constantin, 1990). The Colleges have also encouraged research units to develop measurement instruments with which to measure outcome, for example of the Health of the Nation Outcome Scale (HoNOS), developed by the Royal College of Psychiatrists (Wing, 1994). The Department of Health commissioned a series of working groups, under the auspices of a Central Outcomes Unit (Lakahni, 1994), charged with identifying outcomes indicators for use in routine care settings for ten common conditions - including severe mental illness (Carlwood, *et al.*, 1999).

Audit as an activity

One of the major mechanisms that have been proposed by which to measure outcome in the context of routine care has been 'clinical audit'. The white paper, *Working for Patients,* explicitly included the measurement of 'outcomes' within the remit of audit, where *processes* had previously been measured. The purpose of audit, as proposed within the white paper was to encourage a move towards the 'measurement of outcome' in the hope of assessing the benefits of healthcare and to help generate testable hypotheses on causal connections between *process* and *outcome* (Shanks & Frater, 1993). Comparisons of outcome between one clinician and another or one hospital and another would be possible in order to generate hypotheses for further research and to identify areas that might warrant further investigation. Further, standards of expected outcome might be set, against which performance or effectiveness would be judged.

Audit has been heavily resourced and has contributed to raising the profile of outcomes measurement in the UK (Davies, 1997), despite concerns relating to its effectiveness as a way of improving the quality of healthcare (e.g. Walshe, 1995). Michael Power, Professor of Accounting at the London School of Economics, provides an interesting critique of the rise and role of all forms of audit within society, and places this recent preoccupation within a general movement towards greater accountability (Power, 1997). He describes an audit 'explosion', with wholesale importation of (US) accountancy-based notions of 'value for money' into diverse areas of human activity, including healthcare, policing and education. He goes on to describe 'rituals of verification', where there are checks and balances on all activities without forethought about the benefits or costs involved in this activity. The origins of this explosion date back to the demands for better governance and accountability, and a loss of trust in public services.

Concluding comments

The rise outcome measurement has come about through two interrelated but distinct influences – the quest to find what works (evaluation) and the quest to improve the quality of healthcare (accountability). Various social and political drivers to this rise have been described, which are both specific and common to the US and UK healthcare systems.

There is an implicit assumption that the measurement of outcome is a good thing, in that it will improve clinical practice, and therefore worth the effort and cost. At the micro level, clinicians will be keen to incorporate these data into their day-to-day practice, and at the macro level, that that they will be used in some meaningful way by hospitals. These issues, within the context of psychiatry, will be explored in more detail within this thesis (Sections 2 and 3).

The term outcome is, however, one that is used imprecisely and with various meanings throughout the literature. The next chapter will highlight the imprecise nature of this term, and will seek to clarify the way in which this term will be used in the empirical studies of outcome in this thesis.

Chapter 5 What is meant by outcome?

The preceding discussion highlights some of the diverse influences that have contributed to the rise of 'outcome' as an issue and topic of study. However, the term 'outcome' is used in different ways and with a diversity of meanings, which is in part a reflection of these various influences. Further, it has been endowed with a in the health, social and political spheres.

Table 1 provides examples of the diversity of definitions for 'outcome'. These reflect two related influences on outcome measurement outlined in the previous section; namely the quest to find out what works (evaluation) and the urge to improve the 'quality' of medical care. They also reflect notions of 'change' over time, which may or may not be attributed to antecedent healthcare. Additionally, it is seen that implicit within many of these definitions are varying conceptions of how and what should be measured as an outcome. Only some of these include 'health' in its broadest sense. In some there are implicit assumptions about what constitutes health – where this can be negatively framed *('death disease and disability'* Lohr, 1988) or can be positively framed *('the impact that changes in health have on quality of life'* Seymour, *et al.*, 1993). The following sections will explore in more detail what is meant by 'outcome' and will clarify in what sense the word outcome will be used in the rest of the present thesis.

Table 1: What is meant by outcome?

'....a change as a result of antecedent healthcare" (Donabedian, 1966; Donabedian, 1980).
'.... five Ds: death disease, disability, discomfort and dissatisfaction' (Lohr, 1988)
'(The) results of health care processes.' (Baumberg, *et al.*, 1995)
'All the possible results that stem from preventative or therapeutic interventions; all identified changes in health status arising as a consequence of handling of a health problem' (Last, 1994)
'Outcome to the individual, essentially comes down to how comfortable, how accessible and how appropriate will be the care that is offered between the onset of mortal illness and death.' (Best, 1988)
'A measure of the quality of medical care, the standard on which is made the assessment of the expected end result of the intervention employed.' (Glanze, 1990)
'The attributable effect of an intervention or its lack on a previous health state.' (Department of Health, 1992)
'The inter-relationship between health, health services and other factors in the social system are unclear. An improvement in 'health' may not be the outcome of health services. However, in this context it is measures of those aspects of health which are likely to be affected by changes in health services which are required as indicators of outcome. This concept of outcome implies both a measure of change in health status and an association with health service use/provision.' (Hall, *et al.*, 1984)
'Outcomes are the effects of the utilisation of health services on the health status of the population. This definition implies both a measure of health status and the imputing of a change to the intervention of health services, i.e. a cause-effect relationship.' (Hall, *et al.*, 1986)
'The end results of medical interventions and processes. These can be assessed in terms of mortality, morbidity, physiological measures and, increasing, more subjective patient-based assessments of health.' (Jenkinson, 1994)
'An outcome is a natural or artificially designed point in the care of an individual or population suitable for assessing the effect of an intervention, or lack of intervention, on the natural history of a condition.' (McCallum, 1993)
'a change in the health of an individual, group of people or population which is attributable to an intervention or series of interventions.' (NSW Health Department, 1992)
'Outcome is a relative value. It is a measure of change, the end point is compared with the situation at the start of the study period.' (Pynsent, *et al.*, 1993)
'An important issue in health planning is the measurement and valuation of health outcomes. The former is concerned with description and the latter is concerned with the subjective perception of the impact that changes in health have on quality of life.' (Seymour, *et al.*, 1993)
'In education planning, this refers to any change in health status in a group or population that results from health promotion or health care utilised as measured at one point in time; a cross-sectional epidemiological study of health concerns or health status. A measure of the results of health activation, health education and health promotion.' (Timmreck, 1992)
'The dictionary definition of outcome is 'result or visible effect'. To be concerned with outcomes is simply to be concerned with the causal relationships between antecedent and subsequent conditions or events. But in the context of health and illness, outcome is usually defined in terms of the achievement of or failure to achieve desired goals. Relative to these goals, from a defined starting point, outcomes can be either positive or negative, ranging from complete health to death (or worse).' (Wilkin, *et al.*, 1992)

Attributing outcome to healthcare processes

The frequently cited definition of health outcome given by Avedis Donabedian (1966) as a '*change as a result of antecedent healthcare'* provides a useful starting point in seeking to examine what is meant when the term 'outcome' is used in the medical literature. From Donabedian's definition of 'outcome', there is an emphasis on both change and the attribution of this change to some healthcare intervention. Various criticisms have been levelled at this definition, particularly the assumption that any change can be attributed to healthcare its self, since there are many determinants of health, including cultural, environmental, economic and social factors (Shanks & Frater, 1993).

In order to bring clarity to the idea of outcome in terms of the attribution that is implied, Shanks & Frater (1993) offer the following four distinctions:
Outcome - a result
Health outcome - an effect manifest as a change in health status
Health care outcome - a result which is attributable and responsive to health care
Health outcome of healthcare - a result evident in terms of health status which is attributable to and responsive to healthcare.

The distinction between 'outcome' and 'health outcome' is conceptually clear, but difficult to apply in practice. It is true that many things are measured as important 'outcomes' of healthcare, but which are not a direct measure of the patients underlying health state, such as aspects of service use, employment etc. Likewise change (or lack of change) is commonly attributed to a healthcare intervention, despite no clear establishment of causality. In using the word 'outcome', no clear link between outcome and healthcare process will be assumed. Where a causal link is explicitly demonstrated or inferred, this will be pointed out.

What is measured as outcome.

The definitions also reflect changing ideas about what should be measured as outcomes - from negatively framed biomedical endpoints ('*death and disease'* - Lohr, 1988), through to more recent trends in the measurement of positive aspects of health and the use of 'patient based' measures, which assess the patients' perspective in terms of the impact of disease processes on the individual (Greenfield

& Nelson, 1992). Some of the specific ways in which outcome has come to be measured, particular through the measurement of 'health status' and 'health related quality of life', will be studied in detail in the following chapter. In many cases, when the term 'outcome measure' is used, it is implied that this means the measurement of these broader 'patient based outcomes'. For conceptual clarity, the term 'outcome' when used in this thesis will not presuppose what is measured as an outcome. When necessary, this will be explicitly stated and the use of terms such as 'patient based outcome' will be used where appropriate.

'Temporal change' and outcome.

The idea of 'temporal change' is central to the notion of 'outcome' - and the ability to detect change is a necessary (though often ignored) attribute of any instrument that is used to measure outcome (McDowell & Jenkinson, 1996).

Change can only be measured over time and ideally requires the serial application of a measurement instrument or index. This instrument or index must be capable of measuring this change. The notion of 'temporal change' will be a key feature in considering when an instrument is being used appropriately as a 'measure of outcome' throughout this study. However, stipulating the serial application of a measurement instrument to infer change will exclude many examples of 'outcome measurement'. Where measurement only takes place once, in a scientific sense, it should only be considered a 'measure of outcome' when it is made during or at the end of some healthcare process and there is some implicit relationship inferred between the preceding process and this outcome. In this way, important categorical event-like 'outcomes', such as death, discharge from hospital or relapse can be recorded.

What is meant by 'measurement'?

Operationalisation of 'outcome' in order that it can be measured requires a consideration of what is meant by the term 'measurement'. This necessarily requires a brief overview of measurement theory and the basic tenets of psychometrics (see Nunnally, 1967, for an overview). These are provided below.

Streiner and Norman (Streiner & Norman, 1995) outline two divergent traditions in the conceptualisation of 'measurement'; the *categorical* versus the *dimensional*. *Categorical* measurement stems from the medical tradition, where the world is construed in terms of diagnoses and treatments: either the patient has the disorder or does not and is prescribed a treatment or is not. Diastolic blood pressure and depression might vary in magnitude or severity, but individuals are classified as being normotensive/hypertensive or not depressed/depressed, and hence requiring treatment or not. Conversely, the *dimensional* tradition is exemplified by 'psychometrics' in psychological and educational research, where a phenomenon under study differs only quantitatively at different severities. The science of psychometrics takes the writings of Stevens (1951) as received wisdom (Streiner & Norman, 1995). Stevens has provided a widely accepted definition of measurement *as 'the allocation of numbers to things according to rules'* and has introduced the notion of *'levels of measurement'*, which categorises variables into *nominal, ordinal, interval, and ratio*. The basic idea is that the more finely we can measure something, the better.

Having examined the notion of outcomes and how these can be measured, with reference to the basic tenets of measurement theory, the following section will explore how measurement has been adapted in the pursuit of *patient based outcomes*.

Chapter 6 Patient based outcome measurement

During the 20th century the developed world has seen a rise in life expectancy and a consequent increase in prominence of chronic diseases. Where previously mortality and morbidity rates were collected and were informative about the burden of illness and the quality of healthcare for the population at large - this is now less clear cut (Ebrahim, 1995). Particularly for chronic diseases, there has (necessarily) been a change in the way in which health and healthcare are measured and evaluated (Ware, 1995). Treatments and outcomes in these cases depend not just on quantity but on quality of life.

In healthcare, there has been a shift from the reliance on population based measures of mortality and morbidity to what can be called 'patient based' measures of health and illness (McDaniel & Bach, 1994; McDowell & Newell, 1996). They are 'patient based' in that they incorporate the patients' subjective experience of illness over more traditional biophysical measures that have previously dominated medicine in the evaluation of healthcare (Fitzpatrick, *et al.*, 1984). Where more 'patient based' measures are used to evaluate changes in health status and antecedent healthcare - then we have 'patient based measures of outcome' (Jenkinson, 1994). Some areas of medical speciality have readily incorporated or adopted patient based measures of outcome - for example oncology (Selby, 1993) and rheumatolgy (Liang & Katz, 1992).

The focus of the current thesis is to examine the measurement of *patient based outcomes*. The term cannot easily be defined (McDaniel & Bach, 1994), but the common denominator of all instruments that can be termed 'patient based outcome measures' is that they are said to address some aspect of the patient's 'subjective' experience of health and the consequences of illness. Such instruments ask patients to report views, feelings, experiences that are necessarily perceived by the respondent (Mor & Guadagnoli, 1988). One of the key features of patient based outcomes measures is the recognition of the fact that the *patients' perspective* is worthy of measurement in its self (Fitzpatrick, *et al.*, 1984). The patients' perspective will provide useful information that might not otherwise be obtained from 'hard' (physical or laboratory based) parameters. This approach is based on theories of the 'subjective experience of illness', which assume that individuals

experience illness in ways that cannot be measured well through objective tests and that these feelings and perceptions influence health outcomes (Fitzpatrick, *et al.*, 1984). Respondents are asked about experiences such as satisfaction, difficulty, distress or symptom severity that are unavoidably 'subjective phenomena'. It is taken as given that such experiences cannot be objectively verified (Albrecht, 1994).

A number of synonyms are used for patient based outcome; particularly quality of life, health status and health related quality of life. The terms *quality of life* and *health status* have crept into common usage and instruments designed to measure *patient based outcome* variously describe themselves as measures of *health status*, *quality of life* or *functional status*. Few authors take the trouble to define these terms or explicitly describe what in fact they are measuring (Farquar, 1995; Gill & Feinstein, 1994). It is useful therefore to review the origins of these terms and their subsequent development and appropriation within medical vernacular. The theoretical underpinnings of quality of life and health status will be reviewed and the degree to which these terms might be considered synonymous will be considered. This discussion is not merely an academic one, but will outline debates of genuine relevance to how outcome is now measured and will inform a subsequent glossary of the confusing terminology that is used.

The origins of 'quality of life'

The first recorded uses of the term *quality of life* are discussed by various authors and reflect the origins of the term within the economic and social sciences. Albrecht (1994) cites Pigou's (1920) *The Economics of Welfare* where he discusses government support for the poor in terms of personal well-being and the national dividend. The Oxford English Dictionary first notes the use of quality of life in J. B. Priestly's work, Daylight on Saturday, *'The plans are already ...maturing that would give all our citizens more security, better opportunities and a nobler quality of life'.* In the United States, J.K. Galbraith uses the term throughout his influential thesis *The Affluent Society* (Galbraith, 1958).

The first recorded use of the term within the medical literature can be found in the American publication *Annals of Internal Medicine*, which published an editorial with the title '*Medicine and Quality of Life'* (Elkington, 1966), discussing the problems of transplantation medicine and how its benefits might be measured.

'Quality of life' has both an academic pedigree, in that various theoretical positions and traditions have shaped our understanding of the construct, and has come to be used in a non technical sense in everyday conversation. Patrick & Erickson (1993) provide a review of some early efforts to measure quality of life such as the US Eisenhower Commission on National Goals, which noted the a variety of social and environmental influences on quality of life and spurned research initiatives to operationalise, investigate and measure the concept of *quality of life* (Oliver, *et al.*, 1996). Patrick & Erickson (1993) also trace distinct theoretical bases that have contributed to our subsequent understanding of quality of life. These include the functionalist theories of sociology and anthropology, and theories of positive well-being and quality of life from psychology. Methods by which quality of life has come to be measured have also been influenced by utility theory form economics and decision sciences and by psychophysical theory from psychology.

Those who choose to use measures of quality of life are far from making an atheoretical measurement. The choice of measurement instrument and the items that it contains are implicitly influenced by the theoretical standpoint of those who constructed the instrument. Differing instruments form differing theoretical standpoints may produce different answers when applied to the same phenomenon. For example, functionalist theories, such as those of espoused by Talcot Parsons involve the analysis of social and cultural phenomena in terms of the functions that they perform in a sociocultural system. Parsons, in The Social System, defined illness as:

'*A state of disturbance in the normal functioning of the human individual including both the state of the organism as a biological system, and his personal and social adjustments'.* (Parsons, 1951)

This sociological basis has been the basis for many health and quality of life indicators, and has resulted in measures which focus on the individuals capacity to perform the major social roles – such as work, caring for others or ones own personal needs. This has in turn influenced the items that are included in measures of quality of life – such as ability or capacity to meet these social roles.

Similarly, The economic perspective on quality of life draws much of its theoretical foundation from the classical theories of utility espoused Jeremy Bentham (Bentham, 1789), whose utility principle holds that all individuals and society, as an aggregate of individuals, are directed towards a single end – to increase pleasure and to decrease pain, and that these preferences can be measured. Developments of these theories by, for example, Von Neumann and Morgenstern, who extended notions of uncertainty into preference judgements (von Neumann & Morgernstern, 1944) have directly influenced the methods that are used to measure and value health states. For example the standard gamble, as a dimensional measure of quality of life draws directly from these theories of expected utility and decision making under uncertainty (Drummond, *et al.*, 1997; Torrance, 1987).

The ubiquitous nature of the term quality of life and the differing theoretical stand points which are assumed when it is discussed are reflected in the various differing definitions which are offered. For example, Calman (1984) has presented a widely used definition of quality of life as *'the gap between the patients expectations and achievements'*. Thus, the smaller the gap, the higher the quality of life. Others use the term or 'measure' the construct, but decline to define it (Farquar, 1995; Gill & Feinstein, 1994). The lack of definitional consensus is exemplified by Campbell (1976) who stated that *'quality of life is a vogue and ethereal entity, something that many people talk about, but which nobody clearly knows what to do about'* (Campbell, 1976, cited in Bowling, 1995).

In many cases, quality of life has been very broadly defined and operationalised, through the recognition of the fact that there are many components of quality of life and influences which determine the quality of life for an individual. Health is just one (albeit important) component and determinant of subjective quality of life (Bowling, 1995; Fitzpatrick, *et al.*, 1992a; Patrick & Bergner, 1990). The intellectual influences of economics and welfare have recognised that quality of life is also dependent upon factors such as housing, financial security, employment and opportunity. Much subsequent disagreement, particularly when quality of life has been measured in relation to healthcare, has centred on the definition and relative importance of health and social influences on subjective quality of life, and whether these should be measured (Faden & Leplege, 1992; Greenfield & Nelson, 1992; Guyatt, *et al.*, 1991; Guyatt, *et al.*, 1989). Critics of the use of 'quality of life', as it has come to be

operationalised and measured within much medical research, have commented that wider influences upon quality of life and health are ignored (Gill & Feinstein, 1994).

Health Related Quality of Life

In the face of these criticisms, the idea of *health related quality of life* (HRQoL) has emerged (Patrick & Erickson, 1993; Ware, 1985). HRQoL is a concept that attempts to encompass broader ideas of health than diseases or their absence, by incorporating both personal health status and social well being in assessing the health of individuals and populations (Guyatt, *et al.*, 1993a). One widely quoted definition of HRQoL is that offered by Patrick & Erickson (1993):

'Health related quality of life is the value assigned to duration of life as modified by the impairments, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment or policy"

The use of the term *value* should be noted in this definition. Any measure of HRQoL involves some explicit or implicit value judgement. Measures of HRQoL contain a value judgement on the part of those who develop them, in terms of what constitutes HRQoL and hence what should be measured. HRQoL measurement instruments, such as utility measures, explicitly set out to measure the valuation put on various health states (Torrance, 1987).

Authors generally concur about the components of health which should be included in any measure of HRQoL, and these include psychological, social and physical health; duration of life; impairments; functional status; health perceptions and opportunities (Testa & Nackley, 1994). These are health related, in that they are influenced by disease, injury, treatment or health policy (Patrick & Bergner, 1990). Such items reflect states that are felt to be universally desirable (Faden & Leplege, 1992). Other widely valued aspects of human existence that might be included in some measure of 'quality of life' are not generally domains of HRQoL. These include safe environment; adequate housing; guaranteed income and freedom. Such global concerns may adversely affect or be affected by disease, injury, treatment or policy, but are often unrelated to or distant from health or medical concern. Health Related Quality of Life generally distinguishes the social, familial and behavioural factors and processes that influence it, particularly when health is viewed as an outcome. It is from the outcome perspective that most health status

measures are developed and applied (Patrick & Bergner, 1990). Table 2 produces a comprehensive summary of some of the wide variety of components that have been included in operationalised measures of HRQoL. Not all instruments will measure each and every one of these dimensions, and the scope and comprehensiveness of the domains that are studied will in part depend upon the perspective and purpose of the measure of HRQoL.

Table 2: Concepts and domains of health related quality of life

Adapted from Patrick and Erickson, 1993

Concepts and domains	Definitions/indicators
OPPORTUNITY	
Social or cultural handicap	Disadvantage because of health
Individual resilience	Capacity for health; ability to withstand stress; 'reserve'
HEALTH PERCEPTIONS	
Satisfaction with health	Physical, psychological and social function
General health perception	Self rating of health; health concern/worry
FUNCTIONAL STATUS	
Social	Work and daily role
Psychological	Distress (anxiety, depression, loss of behavioural and emotional control)
Cognitive	Memory, alertness, reasoning
Physical	Activity restrictions, fitness
MORBIDITY	
Signs	Objective clinical findings directly observable
Symptoms	Subjective evidence indirectly observable
Self reports	Patient self reports of symptoms and conditions
Physiologic	Laboratory measures and pathology
Diagnosis and severity	
DEATH & DURATION OF LIFE	Survival, longevity, years of life lost.

Health status

A construct which is related to and sometimes used interchangeably with *quality of life* and *health related quality of life* is that of *'health status'* (Ware, 1987). The term was popularised and explicitly adopted by influential researchers such as John Ware - author of the SF-36 health status questionnaire (Ware & Sherbourne 1992) and Marilyn Bergner - author of the Sickness Impact Profile - SIP (Bergner, *et al.*, 1981)).

The preceding discussion highlights the sociological and economic origins of quality of life and HRQoL. However, *health status* begins from a theoretical position more grounded in health and healthcare. Exponents of the term *health status* begin with a consideration of what constitutes or defines 'health' and seek to operationalise this concept. Again the definition of health that most researchers have relied upon is that of the World Health Organisation (1948), which describes health as:

'a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity'

Bergner (1985) states that in order to operationalise such a concept *'the factors that affect health must be distinguished for the factors that are health'*. Accordingly, in developing measures of health status, she starts with the notion that *'health status ends at the skin'*. Health status measures contain those elements that are an integral part of the person, but excludes those that exist and behave independently of the person (Ware, *et al.*, 1981). Hence, wider economic and social influences on health - which might be included in some definition of 'quality of life' - are generally excluded form measures of HRQoL.

In a similar vein, Ware (1987) suggests that '*....the use of 'quality of life' nomenclature is likely to cause confusion because it is too encompassing. Jobs, housing, schools, and the neighbourhood are not attributes of a person's health, and they are well outside of the purviews of the health care system....The goal of the health care system is to maximise the quality of life, namely health status*'

A further distinction between health status and HRQoL is that although health status measures are meant to be based upon definitions of health which include positive aspects of health *(....not just the absence of illness*), in effect they have generally

been based principally on negative aspects and definitions of health (Bergner, 1985; Bowling, 1995). This distinction raises problems. Proponents of ideas of *positive health*, believe that negative conceptions of health (i.e. the absence of disease) may be sufficient for measuring health status in ill patients (who constitute approximately 20% of the population at one time (Stewart, *et al.*, 1989; Ware & Young, 1979)), but are of little value in defining the experience of the other 80-90% of the population.

Proponents of *health status* have reflected on the lack of conceptual clarity or definition that surrounds the concept of quality of life (and inter-alia health related quality of life) (Bergner & Rothman, 1987) and the fact that those who purport to measure quality of life rarely define it (Gill & Feinstein, 1994). According to Ware (1995),

'..it has become fashionable to lump all measures that define health beyond traditional indicators of biologic functioning into a single category of quality of life measures. This practice offers a shorthand method for referring to a collection of concepts both more broad and more qualitative than traditional measures of clinical endpoints. However, quality of life as traditionally defined is a much broader concept than health. Quality of life encompasses standard of living, quality of housing and neighbourhood, job satisfaction, health, and other factors.....using quality of life nomenclature, without qualification is likely to cause some confusion.'

Debates about the distinction between health status and quality of life continue, but some authors have contested that the distinction is no longer relevant since the most commonly used measures of HRQoL and health status are in effect measuring the same dimensions (Ware, 1995). A useful summary of the core components of HRQoL and health status are provided in table 3, and when a health status measurement instrument measures all of these things it is in effect a measure of HRQoL.

Table 3. Components of Health Status and HRQoL

Adapted from (Ware, 1987; Ware, 1995)

Physical health Ability to perform physical activity and self care activities (e.g. eating, bathing & dressing), and the limitations caused by illness and pain.
Mental health Intensity of symptoms of psychological distress and behavioural dysfunction consequent upon this. Includes not just psychological distress, but psychological well-being
Social functioning Social contacts and other activities (e.g. visits with friends and relatives), and social ties or resources (e.g. close friends and relatives that can be relied upon for tangible and intangible support).
Role functioning The performance (or ability to perform) usual role activities, including formal employment, school work, etc.
General perceptions of health and well-being How people evaluate overall health and well-being. Representing an individual appraisal or overall evaluation of the above factors.
Cognitive capacity and function Orientation, memory, comprehension, abstract reasoning and problem solving

In summary, the term Quality of Life is conceptually different from health status and HRQoL and when used should be defined or qualified, such that users of such scales are aware of exactly what is being measured (Gill & Feinstein, 1994). Mosteller attempts to explain the persistence of 'quality of life' (Mosteller, 1989) as being due to the intrinsic appeal of the term, which is *'such a winner of a title that none of us want to give it up'*. A more useful approach than the argument over terminology might be the explicit statement of what is in fact being measured by an instrument. Thus, we can recognise that HRQoL and health status measures are the same, when we know what they include. Similarly, the oversimplification of HRQoL when it measures only one domain of health (such as depression or anxiety or physical functioning) does not cause problems when the scope of what is measured is made explicit (Pope & Tarlov, 1991).

Chapter 7 Taxonomies of measurement instruments

One of the most influential and widely adopted taxonomic classification systems to have been proposed in recent years focuses upon a description of the area of application of a specific instrument. The broad distinction of measurement instruments into: (i) those that are **generic** and can be applied across all populations and disease groups and severities of disease, and (ii) those that are **specific** and focus upon problems associated with individual diseases, patient groups or areas of function. The distinction has developed as a way of classifying various patient based measures of outcome - particularly health status and quality of life measures (Bergner, 1985; Bergner, 1989; Guyatt, *et al.*, 1989; Patrick & Bergner, 1990; Patrick & Deyo, 1989; Stewart, *et al.*, 1989).

Generic HRQoL instruments are intended to provide a common metric with which to compare health, illness and outcome between different populations and conditions. Their purpose and requirements neatly summed up by Kane (1987), who described generic measures as:

'the clinical equivalent of the Swiss Army knife - something small and easily taken into the field with enough blades and attachments to fit any number of circumstances which may arise'.

Thus, the ideal generic outcome measure will be an instrument that covers sufficient facets of life and health that are universally important to people – irrespective of their age, sex, disease or health-state. Similarly, they will be sufficiently brief and easy to complete such that patients will not be inconvenienced by their application. There are broadly two approaches to the measurement of generic HRQoL - *profiles* and *utility measures*. These are outlined below:

Health profiles

The measurement of health through the use of health profiles adopts an explicit 'psychometric' approach. Psychometrics originally derived from the science of pscyhophysics – which attempted the measurement of human perceptions of different natural stimuli, such as heat and light (Nunnally, 1967). Psychometrics extended the methods of psychophysics into more subjective psychological

dimensions such as intelligence, attitudes and health perception (McDowell & Newell, 1996).

Health profiles seek to measure important dimensions of health related quality of life. As noted above, these tend to be some combination of the following: *physical functioning; social functioning; role functioning; mental well-being; and general health perceptions* (Ware, 1995). Health profiles provide detailed information about multiple domains of HRQoL, which is useful for specifying the pattern of functioning and well being of the individuals being studied. Scores on each of these dimensions are either reported separately (a true health profile) or are occasionally combined to provide an overall summary score (a summary health index). Two widely used health profiles include the SF 36 (Ware & Sherbourne, 1992) and the Sickness Impact Profile (Bergner, 1985; Bergner, *et al.*, 1976), which measure some combination of the above domains.

Utility measures

Despite the fact that health is multi-dimensional, there are practical reasons why it might be useful to obtain some overall summary score of 'health' that combines these dimensions. The main reason why this approach might be preferred to a *health profile* is that health profiles make it difficult to make decisions about health gain and comparisons between different interventions (Brazier, 1993). In psychometric terms, the scores generated by a health profile are not in units comparable either between its own dimensions, or with other consequences. For example, an intervention might produce health gain in one dimension (e.g. social functioning) and decrements in another (e.g. mental well-being), and it is difficult to know whether this represents an overall 'health gain'. Similarly, one intervention might produce benefit in one domain, whilst another intervention might improve a different domain. Individuals, patients or wider society might value a gain in one dimension as being more desirable than gain in another (Kaplan & Coons, 1992). When decision-makers come to allocate resources or set priorities, it is difficult to know which treatment is best, since there is no common metric between competing alternatives. The only basis upon which a treatment could be seen as better is if gain is demonstrated on all dimensions of a health profile.

Utility measures represent an attempt to produce a single summary index of health related quality of life, by collapsing these dimensions into a single score according to how they are valued (Torrance & Feeny, 1989) and are derived form economic and decision theory (von Neumann & Morgernstern, 1944). They reflect the preferences of patients for health states that result from intervention. The key elements of utility measures are that they incorporate preference measurements and relate health states to death (Torgerson & Rafferty, 1999). Thus they can be used in one specific type of economic evaluation that combines duration and quality of life - the cost utility analysis (Drummond, *et al.*, 1997; Torrance, 1987). Through the process of weighting HRQoL states and the duration of those states - quality adjusted life years (QALYs) are calculated (Williams, 1985).

Disease and domain specific instruments

The desire to produce a common metric which is applicable to all patients, conditions and populations is understandable, but in doing so, difficulties arise. The inclusion of only those aspects of health that are 'universal' will mean that important features of, say, a specific disease will not be measured, or their effects will be diluted. A clinically important change in health status might therefore not be recorded by the application of only a generic instrument. One solution might be to produce a generic instrument which is sufficiently comprehensive to measure all aspects of health for all diseases and populations – which would as a consequence be of enormous length and unacceptable (Golligher, 1987). Another approach has been to produce instruments that measure a facet of a specific disease – disease specific measures, or specific domains of functioning which are common to many diseases – domain specific measures (Bowling, 1997). These approaches will be considered below, with examples.

Diseases specific measures

Disease specific measures focus upon the symptoms or facets of health status, which are important for that specific disease. Their major advantage is their brevity (since they measure only what is necessary and avoid the use of redundant questions included in generic instruments) and their ability to detect small but clinically significant changes in health status and levels of disease severity.

Domain specific measures

Important outcomes or domains are often excluded form generic instruments in the search for brevity. Whilst these outcomes are globally applicable to all patients, they are not of equal importance. Thus areas, such as cognitive functioning, or detailed assessment of mood, might not be included in generic instruments, but might be important in certain populations (such as the elderly) when assessing outcome or response to a certain intervention.

Examples of commonly used domain specific measures include mood questionnaires (such as the Hospital Anxiety and Depression scale (Zigmond & Snaith, 1983); cognitive function and memory tests (such as the Mini Mental State Examination - Fostein, *et al.*, 1975); pain questionnaires (such as the McGill Pain Questionnaire - Melzack, 1975). Whilst such measures are very sensitive to changes in the underlying domain which they measure, they are often long and time consuming to complete. This becomes problematic when large 'batteries' of domain and disease specific questionnaires are employed to measure outcome. For this reason, generic questionnaires, with the only essential additional disease/domain specific measures is recommended (McDowell & Jenkinson, 1996; McDowell & Newell, 1996).

The major disadvantage of disease and domain specific instruments is encompassed within their specificity. They do not allow comparisons in terms of health gain between conditions or populations to be made. Also in common with health profiles, they are constructed without reference to 'preferences'. Thus they might show 'improvement' whilst the patient themselves might feel worse off. As a consequence, they are of little use in resource allocation or priority setting (Cairns, 1996).

Strengths and weaknesses of generic and specific measures of health

As described in the preceding section, there are inherent strengths and weaknesses to the various approaches to outcome measurement – these are summarised in table 4. General recommendations on the use of measurement instruments (particularly within clinical trials) involves the use of the minimum number of scales; which should include a generic profile and measurement of patient utility, together

with any necessary domain specific instrument (Bowling, 1995; McDowell & Jenkinson, 1996; McDowell & Newell, 1996; Wilkin, *et al.*, 1992).

This taxonomy will be used throughout this thesis and measures will be described according to the following categories: *Generic, disease specific, and domain specific*. The latter two categories will to a certain extent overlap, and where this occurs this will be noted.

Having introduced a vocabulary within which outcome measures and measure of patient based outcome can be discussed with some degree of clarity, this will be applied in subsequent sections of this thesis.

Table 4: Strengths and weaknesses of generic and specific measures of health

Measure	Strengths	Weaknesses
Generic instruments		
Health profile	Is a single instrument Has established reliability and validity Detects differential effects on different aspects of health status Allows comparisons between interventions or conditions	May not focus adequately on area of interest May not be responsive Difficult to know whether health state is improved (is more desirable) when gain in one dimension is accompanied with loss in another
Utility measurement	Provides single number representing net impact on quality of life Allows cost utility analysis	May involve difficulty in determining utility values Does not allow examination of effect on different aspects of quality of life May not be responsive Sometimes difficult for respondents to comprehend or agree to express preferences.
Specific instruments	Are clinically sensible May be more responsive than generic instrument	Do not allow comparison between conditions May be limited in terms of populations and interventions

Chapter 8 Uses of patient based outcome measures.

Patient based health outcome measures have been defined and their content examined in chapters 1-7. They have been put to several distinct uses in the realm of healthcare practice, research and policymaking. Several writers have described the various *uses* of measurement instruments in healthcare (Fitzpatrick, 1994; Fitzpatrick, *et al.*, 1992a; Kane & Kane, 1981; Nelson & Berwick, 1989; Patrick & Bergner, 1990; Steinwachs, 1989; Ware, 1995). Similarly, some of these uses are more relevant to a discussion of patient based outcome measurement than others. These uses often reflect a US based focus of the research literature and also serve to illustrate the lack of clarity that is apparent when the term 'outcome' is used. The purpose of this section is to introduce a terminology that will be used throughout the rest of thesis in describing the uses of *patient based outcome measures* and to provide illustrative examples of these uses.

Health Care Policy Evaluation

The restructuring and reorganisation of healthcare systems - principally in the US, in response to escalating healthcare costs, has generated an impetus to measure the *health consequences* of these changes. As described previously in Chapter 2, healthcare systems have in the past been evaluated by crude measures of activity and utilisation, rather than patient based measures of health status (Kindig, 1977). The recognition that health care organisation and evaluation requires more complex and patient based measures has been one of the central tenets of the US 'outcomes movement' (Ellwood, 1988). Reorganisation strategies such as cost containment, managed care, co-payment and the reimbursement of episodes of care according 'Diagnostic Related Group' have raised fears that the care (and health) of certain groups of patients will suffer. For example, there is the concern that patients will be discharged from hospitals 'quicker and sicker'.

Two important landmarks in the evaluation of health policy are the Health Insurance Experiment and the Medical Outcomes Study, which are described in some detail below.

The Health Insurance Experiment (HIE) is the largest evaluation of health care policy to date, and has been discussed previously. Briefly, the healthcare effects of two cost containment strategies - cost sharing in a fee for service (FFS) system and

a prepayment method of insurance - were evaluated using standardised surveys of health and social function over a five year period (Patrick, *et al.*, 1973). A total of 4000 people were enrolled and followed up for three to five years, having been randomly allocated to differing health insurance programmes. Co-payment schemes resulted in one third less healthcare utilisation when compared to 'free at the point of entry' care. An expressed aim of those conducting the study was to determine the actual effect of this reduced healthcare utilisation on 'broader health' (Brook, *et al.*, 1983). Subjective health was explicitly measured in addition to harder outcomes, and a health questionnaire was developed for this purpose. The self-completed HIE health questionnaire consisted of 108 items, measuring five dimensions: *physical functioning, mental health, social contacts, and health perception*. According to the authors, the HIE *'clearly demonstrated the potential for scales constructed from self administered surveys as reliable, valid tools for assessing changes in health status for adults and children in the general population'* (Ware & Sherbourne, 1992). Aside form the impact of this study on healthcare policy in increasing the use of cost sharing strategies, the enduring impact of the HIE has been to raise the profile of health status measurement.

The subsequent Medical Outcomes Study (MOS) sought to further develop patient based measures, refining and making more practicable the instruments developed in the HIE, in order to investigate the effect of variations in system of care, clinician speciality, and clinicians' technical and interpersonal style on actual patient outcome. A total of 3000 patients with a number of medical conditions, including diabetes, hypertension, heart disease and depression were recruited and were followed up for two years. Aspects of service use and treatments were monitored and outcomes (both self reported and clinical'/laboratory measures) were examined. The study was able to correlate structures (e.g. method of payment), processes (e.g. aspects of practice style) with outcomes. The relevance of the Medical Outcomes study to mental healthcare evaluation in particular will be examined in detail in section 2 of this thesis.

The self completed health status questionnaires developed in the MOS eventually evolved into the Short Form - 36, which has become one of the most widely used and heavily promoted patient based outcomes measures in the 1990s (Brazier, 1993; Hays, *et al.*, 1993; McHorney & Ware, 1995; McHorney, *et al.*, 1993; Tarlov, *et al.*, 1989; Ware & Sherbourne, 1992; Wright, 1994b).

Health care evaluation

Clinical trials (particularly when randomised and double blind) provide the most valid form of evaluation of one treatment, intervention or technology against another (Guyatt, *et al.*, 1993b). Alternative treatment regimens and technologies can and should be compared in terms of their impact on patient functioning and well being, in addition to traditionally defined biologic endpoints (Guyatt, *et al.*, 1991). In the UK, the Department of Health (1992) suggests that the following should be incorporated into outcome measurement: survival rates, symptoms and complications, health status and quality of life, the experiences of carers and the costs and use of resources. Their report continued: *'many health technologies are intended to improve general health and the quality of life, so it is important to measure patients subjective experiences of illness and the care they receive'*.

In general, and with notable exceptions, patient based measures have not been used in healthcare evaluation (Aaranson, 1989; Sanders, *et al.*, 1998). Broader measures of health status clearly have the potential to complement traditionally defined clinical endpoints in all conditions – but have generally not been measured, although this is not always the case. The two spheres of healthcare that seem to be particularly well advanced in this respect are rheumatology and oncology. The example of outcome measurement in rheumatology and oncology serves to illustrate the need for and contribution of patient based measures.

Rheumatological conditions are chronic in their nature and important patient centred outcomes might include the ability to perform activities of daily living and resolution of pain. In the early 1980s, outcome was traditionally defined in terms of biophysical endpoints, such as blood titres of rheumatoid factor and range of joint movement expressed in degrees (Bombardier & Tugwell, 1982). Innovative treatments were introduced in the 1980s (such as aurofurantoin), which possessed both toxic side effects and a potential to modify disease activity. The evaluation of the relative contribution of adverse and beneficial effects at the patient level required the incorporation of broader measures of health status (Meenan, *et al.*, 1984). The introduction of patient based measures such as the Arthritis Impact Scale (AIMS) (Kazis, *et al.*, 1988) has broadened the way in which outcome is measured in many trials in rheumatology trials. Similarly in Cancer therapies, where traditionally outcome is measured by five year survival rates, the dimension of health status has

been introduced to measure both quality and quantity of life, in order to evaluate therapy. A recent survey of RCTs in cancer showed that 29% of trials reported using some measure of health status or quality of life, compared to less than 2% of trials in other areas of healthcare (Sanders, *et al.*, 1998). It is worthy of note that one of the earliest standardised measures of health status and, until recently, the most commonly used is the Karnofsky index. This scale was introduced to supplement mortality data in the evaluation of cancer treatment in patients in 1948.

Making individual clinical decisions in routine medical practice

In contrast to some of the more research-oriented uses outlined so far, health status instruments might also be used in routine clinical practice with aim of improving the quality of individual care. It has been argued that patient based outcome measures offer an important adjunct to clinicians in the care of their patients (Tarlov, *et al.*, 1989). Here the purpose of patient based instruments might be (1) to aid the recognition of problems which might be otherwise unrecognised or (2) to monitor the progress of the individual patient and hence to monitor and guide treatment (Fitzpatrick, 1994). In the first of these uses, the identification of unrecognised problems, patient based instruments are in effect being used as *screening* or *case recognition* instruments. Traditional forms of screening, such as radiological investigations and biochemical tests, are generally evaluated using the parameters of sensitivity, specificity and predictive value. These parameters should be employed when investigating the performance of health status measures, although this is rarely the case (Fitzpatrick, 1994).

Health professionals are often unaware of many of their patient's health and social problems (Sprangers & Aaranson, 1992) and there might be an expected benefit of health status measures in improving patient care. Generally, this expected benefit has not been seen in experimental evaluations of the health status measures when applied to routine practice. One study is that by Kazis, *et al.* (1990), who conducted a trial to examine the benefits gained by informing clinicians of their rheumatological patients' health status scores. Patients who completed health status questionnaires and had the results of these scales fed back to their clinicians had no better outcome than patients who did not have results fed back to their clinicians. The lack of evidence supporting the application of health status questionnaires in effecting improved care and clinical outcome is reviewed by Fitzpatrick (1994).

It has been noted that patients welcome the opportunity to impart information which is generally outside of the scope of the traditional clinical interview - but which they feel to be important (Nelson & Berwick, 1989). In this use, patient based instruments might be seen as an adjunct to traditional clinical encounters, which facilitate better communication. This is an explicit (but not sole) purpose of a number of health status and quality of life instruments, such as the Nottingham Health Profile (Hunt, *et al.*, 1985a) and the Lancashire Quality of Life Profile (Oliver, *et al.*, 1996). For some instruments, such as the Dartmouth COOP, this is the primary function and most important influence on design (Nelson, *et al.*, 1990) – i.e. it is brief, easy to understand and complete and asks questions which patients will find relevant. The 'ease of use' of an instrument will ultimately influence the degree to which an instrument is acceptable to patients and clinicians. For this reason, short forms have been developed which are easy to incorporate into routine practice (Nelson, *et al.*, 1990).

Economic evaluation and resource allocation

The measurement of both monetary cost and outcome (positive and negative) is the defining feature of an economic evaluation. Cost and outcome can be combined to produce measures of hypothetical benefit, which can be obtained for a given expenditure, such as incremental cost effectiveness ratios, and quality adjusted life years (QALYs) (Williams & Kind, 1992).

One of the most controversial applications of health status and quality of life measures has been their use in allocating limited resources among competing healthcare programmes (Spiegelhalter, *et al.*, 1992). The instrument most used in this context is a specific type of measure - the Quality Adjusted Life Year (QALY). The nature, underlying assumptions and properties of utility measures, such as the QALY, have been introduced in a previous section. Briefly, QALY measures combine quantity and quality of life into a single measure (Williams, 1985), in order to assess benefit brought about by a funded programme. For each programme, this benefit (in terms of QALYs) can be divided by its economic cost and the resulting ratio (cost/QALY) used to allocate resources. QALYs can be used chose between alternative programmes for treating the same patient's or more controversially, to choose among programmes targeted at different groups.

The underlying philosophy behind the use of QALYs and cost/QALY estimates is that rationing of resources is inevitable and that it is best to be explicit and accountable (Smith, 1991).

Clinical audit

Audit consists of reviewing and monitoring current practice and evaluation (comparison of performance) against predefined standards and the use of this information to improve standards (Higginson, 1994; Standing Committee on Postgraduate Medical Education, 1989). Audit has tended to use measures of process in preference to measures of outcome as the 'standards' that are measured (Crombie & Davies, 1997). However, the systematic measurement of outcome has been proposed as a 'standard' in audit (Frater & Costain, 1992). Long & Dixon (1996) identify two scenarios whereby outcome can be usefully measured in the audit process. Firstly, by using adverse events or outcomes as sentinel events that prompt an investigation into the process of care to judge what (if anything) went wrong? An example of this might be confidential inquiries into perioperative deaths and critical incident monitoring in anaesthesia. Secondly, by setting a standard in terms of outcome and monitoring whether this outcome is achieved in routine practice.

Monitoring the health and assessing the needs of population ('healthcare needs assessment').

Those responsible for purchasing and providing health care are increasingly expected to base their decisions about the allocation of health care resources on evidence (Kelly, *et al.*, 1996). The 'needs' of a population is one component of rational allocation of resources. It has been argued that patient based measures provide a feasible and valid measure of health status, which supplement traditions epidemiological indices of mortality and morbidity (Hunt, *et al.*, 1985b). Some authors discuss this use as an example of 'outcomes measurement' (Delamonthe, 1994; Frater, 1992; Geigle & Jones, 1990), although a single snapshot of the health status and needs of a population does not fulfil the definition offered earlier – i.e. a measure of *change* in health.

Chapter 9 Introduction to the rest of the thesis

The preceding discussion has highlighted a number of core themes which emerge in considering the historical development of outcomes measurement in healthcare and the movement towards the use of more 'patient based' measurement instruments. The broad purpose of the current thesis is to produce a critical overview of patient based outcome measurement in psychiatric research and practice. Some of the major themes that have emerged from the preceding chapter are outlined below, together with a discussion of how and why these themes will be explored within the realm of psychiatry.

Section 2 Outcomes measurement in psychiatric research

Given the variety of perspectives and tools that can be adopted in the measurement of outcome, then it will be of interest to know how outcome has come to be measured in psychiatric research. A series of surveys, with illustrative examples will empirically demonstrate the tools and methods that are used in psychiatric research – with particular emphasis on patient based measures.

The output of this section of the thesis will be an overview of the current use of outcomes measures in psychiatric research.

Section 3 Outcomes measurement in clinical practice

The urge to incorporate outcome measurement into the routine day to day care of patients is built upon the supposition that the information collected within such measures reflects the patients' perspective and that such information is useful in actual decision-making. However, we know little about what measures are actually used in routine practice and in healthcare decision making in the UK. Empirical research will therefore demonstrate what (if any) outcome measures are used in actual psychiatric patient care and decision making in the current NHS. This empirical research is intended to complement the review of the use of patient based outcome measures in actual psychiatric research. It will, for example, be of interest to know to what extent measures used in psychiatric research are actually used in routine care.

The collection of patient based outcome data in the context of actual routine care is in itself a fruitless exercise unless it improves the quality or actual outcome of care. The empirical research base from wider healthcare is at best contradictory in providing any support for the routine collection of outcomes in this respect (Fitzpatrick, 1994; Fitzpatrick, *et al.*, 1992a). It will therefore be of interest to know what research evidence exists to support the routine collection and use of outcomes in psychiatric care. The research evidence to support routine outcomes measurement (particularly patient based outcomes measurement) will be examined using a systematic review methodology.

Section 2 - Outcomes measurement in psychiatric research

Section 2.1 Introduction

Section 2.2 Outcomes measurement in clinical trials in psychiatry

Section 2.3 Outcomes research in psychiatry

Section 2.1 Overall Introduction to outcomes measurement in psychiatric research

Chapter 10 Measurement in psychiatry

Measurement in psychiatry has had to incorporate the operationalisation and recording of subjective experience – i.e. the measurement of patient's reports of internal psychic phenomena in the form of psychiatric symptoms, aspects of mood, anxiety, delusions and hallucinations. These are phenomena that cannot be externally observed or verified. There is no (as yet identified) diagnostic pathophysiological basis for 'functional' psychiatric disorders (such as schizophrenia and depression), and most classificatory systems (such as DSM and ICD) diagnose illness according to the presence or absence of mental symptoms that are 'subjective' in their nature in that they are perceived by the patient (Cooper, *et al.*, 1972). These diagnostic systems, for the greatest part, involve the use of trained observers asking standardised questions of patients to record (in a reproducible manner) the presence or absence of internal mental symptoms. Similarly, there has been significant work in the production of 'standardised' measurement instruments with which to diagnose psychiatric disorder in a reliable manner and/or to quantify the degree of severity of a 'disorder'. These standardised instruments have made possible subsequent epidemiological studies of population incidence and prevalence of major mental disorders (e.g. Reiger & Kaelber, 1995; Sartorius, *et al.*, 1986) and investigations of the course of illness (e.g. Shepherd, *et al.*, 1989). Thus standardised instruments, which have been shown to be both valid and reliable in diagnosing and measuring the severity of psychiatric disorders are available to researchers and clinicians – and are seen as valid tools in the conduct, presentation and communication of psychiatric research.

Max Hamilton, the author of one of the most influential standardised instruments in psychiatry, the Hamilton Depression Rating Scale (Hamilton, 1967), writing in 1972 reflected the optimism and embrace of standardised measures in psychiatric research, when he stated:

'A rating scale is, in a sense, an end product of the development of psychiatry. When the phenomena to be studied have been completely defined in nature and range, then it is possible to construct a scale to evaluate them'. (Hamilton, 1972)

Standardised symptom based measures therefore form the backbone of psychiatric research, and there seems to have been an industry in their construction. However,

psychiatry has not restricted its self to the measurement of psychopathology. Thornley & Adams (1998) in a survey of over 2000 randomised trials conducted in schizophrenia found 640 scales to be in use, of which only one third were explicit measures of psychopathological symptoms. The main reason for this proliferation and dominance of standardised outcomes instruments are likely to be the fact psychiatry generally involves the care of persons with chronic and often socially disabling disorders such as schizophrenia, for which standard and easily recordable endpoints such as mortality have limited meaning.

A commonly used classification system for outcomes measures in general (and patient based outcomes measures in particular) divides instruments into generic, disease specific and domain specific measures (Bowling, 1997) –see section 1. Difficulties arise when applying this taxonomy directly in the sphere of psychiatry. Firstly, many authors consider instruments that measure the frequency and intensity of psychiatric symptoms (especially those encountered in mood disorders) to be patient based measures of outcome e.g. (Bowling, 1997; Sanders, *et al.*, 1998), since this is a core component of the dimensions and domains considered to be integral to health related quality of life (Ware, 1987; Ware, 1995). However, this analysis is difficult to support in psychiatry. Other specialities (such as rheumatology and oncology) rightly contrast biophysical measures of outcome with patient based measures of outcome. For example, in rheumatology, the erythrocyte sedimentation rate or the number of joints that are affected may have little bearing on the way in which the individual with arthritis lives their day-to-day life. In order to assess this, patient based measures are adopted. However, in functional psychiatric disorders, there are no biophysical correlates of disease. Instead, instruments are used which measure the frequency and intensity of subjective psychiatric symptoms, with little examination of how these relate to the impact of the disorder on the individual. The nature and basis of common psychopathological ratings scales are considered in more depth below, but for the purposes of the present thesis, these will not be considered as patient based measures of outcome (either generic or domain specific). Secondly, some commonly used measures in psychiatry fall somewhere between measures of psychopathology and measures of functioning – these include some important global measures of outcome. The nature and basis of these measures is also considered below.

In summary, throughout this and subsequent sections, a distinction will be drawn between standardised instruments which count the frequency and intensity of symptoms associated with the diagnosis and severity of a disorder (*symptom based psychopathology measures*), and instruments which judge the impact of psychiatric disorders on the individual and how they live their day to day life (*patient based measures*).

The following section outlines some of the major methods and instruments that are available for use in evaluative psychiatric research, and which will be explored in more detail in subsequent sections.

Standardised measures of psychiatric symptoms

Examples of such symptom-based instruments are the Brief Psychiatric Rating Scale (used in schizophrenia) and the Hamilton Depression Rating Scale (used in depression). The content of these two measures is outlined in Table 5. These are usually (but not always) clinician or interviewer administered and rated instruments.

Table 5: Content of two common symptom-based measures

Hamilton Depression rating Scale (HDRS) (Hamilton, 1967)
The HDRS is a clinician-completed scale, with 17 items that cover the following symptoms associated with depression:

- Depressed mood
- Self depreciation and guilt feelings
- Suicidal impulses
- Insomnia
- Somatic symptoms
- Retardation/agitation
- Anxiety
- Sexual interest
- Ability to work and engage in interests

Brief Psychiatric Rating Scale (Overall & Gorham, 1962)
The BPRS measures the following symptoms associated with schizophrenia, together with depressive symptoms

- Somatic concerns (including delusions)
- Anxiety
- Emotional withdrawal
- Conceptual disorganisation
- Self depreciation and guilt
- Movement disorders
- Depressed mood
- Hostility/suspiciousness
- Hallucinations
- Motor retardation
- Unusual thought content
- Blunted or inappropriate affect
- Disorientation or confusion

Global measures of outcome

Global measures of outcome have a long history in psychiatry, which begins with the Health Sickness Rating Scale by (Luborsky, 1962) in 1962, which represented an attempt to rate health/sickness on a 100 point scale. Subsequent modifications include the Global Assessment Scale in 1976 (Enndicot, *et al.*, 1976), and the Global Assessment of Functioning scale, which forms axis V of the fourth edition of the Diagnostic and Statistical Manual - DSM-IV (American Psychiatric Association, 1994). Most measures have attempted to include some overall assessment of both functioning and psychiatric symptom intensity, usually made by clinicians.

Such scales therefore lie somewhere between symptom based measures, and those measures which tap domains included in instruments which have hitherto been referred to as patient based measures (see below). Spitzer, *et al.*, (1996) in a review of the content and psychometric properties of the GAF, refers to it as an overall measure of 'psychosocial health/sickness'. Global measures, such as the GAF are intended to be applied to all patients with psychiatric disorders, irrespective of diagnosis. The structure of the GAF is outlined in Table 6.

Table 6: An example of a global outcome measure

The Global Assessment of Functioning Scale (Spitzer, *et al.*, 1996)

Clinicians are urged to rate global function between 0 (worst) and 90 (best), considering *'psychological, social and occupational functioning on a hypothetical continuum of mental health-illness.'*

Raters are provided with a series of anchor points to guide their rating:

Code 81-90
'absent or minimal symptoms (e.g. mild anxiety before an exam), good functioning in all areas, interested and involved in a wide range of activities, socially effective, generally satisfied with life'.

Code 41-50
'Serious symptoms (e.g. suicidal ideation, severe obsessional rituals) OR any serious impairment in social, occupational or school functioning'

Code 1-10
Persistent danger of severely hurting self or others (e.g. recurrent violence) OR persistent inability to maintain minimal personal hygiene OR serious suicidal act with clear expectation of death.

Social and role functioning

Mental disorders are generally strongly associated with social dysfunction, particularly schizophrenia and the major affective disorders (Wiersma, 1996). Since the 1960s, there has been a proliferation instruments to measure social and role functioning (Katching, 1983; Weissman, 1975). Wiersma (1996) identifies the major domains that are included in popular measures of social and role function:

- Occupational role (work, eductaion, household, regualr activities)
- Household role (participating and contributing to the household and its economic independence)
- Marital role (emotional/sexual relationship with partner)
- Parental role (relationship with children, caring)
- Family or kinship role (relationship with parents and siblings)
- Social role (relationships with community, with friends and acquaintances)
- Leisure activities and or general interests
- Self care (grooming and appearance)

Commonly used standardised instruments include the Social Adjustment Scale (Weissman & Bothwell, 1976); Katz Adjustment Scale (Katz & Lyerly, 1963); Social Functioning Scale SFS (Remmington & Tyrer, 1979); Index of activities of Daily Living (Katz, *et al.*, 1963).

Quality of life and health related quality of life

There are a number of quality of life, health related quality of life instruments that have been developed specifically for the use amongst persons with mental disorders. The common feature of these instruments is that they measure more than just psychopathological symptoms or single domains of health related quality of life (Ware, 1995), such as social functioning. According to Lehman (2001), common features of quality of life measures designed for use in people with mental disorders is the fact they *'cover patients' perspectives on what they have, how they are doing and how they feel about their life circumstances'*. Specifically, they include sense of wellbeing; functional status; access to resources and opportunities. An example includes Lehman's own Quality of life Interview – QOLI (Lehman, 1983b), which is described in table 7.

Table 7: Lehman's Quality of life Index

The QOLI is a self-report, interviewer-administered measure, which consists of 153 items, and takes 40 minutes to complete. The QOLI measures global life satisfaction as well as objective QOL (what they do) and subjective QOL (how they feel about these experiences) in eight life domains:

Living situation;

Daily activities and functioning;

Family relations;

Social relations;

Finances;

Work and school;

Legal and safety issues and health.

It was designed for persons with severe and persistent mental illness, particularly in community settings, but it has been adapted for those in long term institutional care.. An example of a typical question is given below:

Q. In the past year, how often did you get together with a member of your family?

Answer: Once a day, once a week, once a month, at least once during the year, not at all.

How do you feel about:

A. Your family in general?

B. How often you have contact with your family?

C. The way you and your family act toward each other?

Answer: Terrible; unhappy; mostly dissatisfied; mixed; mostly satisfied; pleased; delighted

(Lehman, 1983b)

Having briefly outlined some of the instruments that are available to researchers in psychiatry, the following section will now examine how these instruments have been used to measure outcome in two major forms of evaluative research: Clinical trials (section 2.2) and outcomes research (section 2.3).

Section 2.2 Outcomes measurement in clinical trials in psychiatry

Chapter 11 Introduction to the survey of clinical trials

The previous section outlined some of the methods and standardised measures that are available with which to measure outcome in psychiatric research. The purpose of this section is to seek to examine, using the taxonomy introduced in the previous section, the way in which outcome is measured in clinical trials conducted in psychiatry.

Clinical trials are considered to be the most robust form of evidence in deciding what works in healthcare in general (Sackett, *et al.*, 1991), and also in mental health (WHO, 1991). In particular, randomised controlled clinical trials (RCTs) have been judged to be the best method available, largely due to their ability to eliminate confounding by ensuring that treatment is allocated according to the play of chance through randomisation (Guyatt, *et al.*, 1993b; Pocock, 1983). The prominence of clinical trials has been recognised within the recent *evidence based* movement, where they form the highest level of clinical evidence, and where the application of this evidence in clinical decision making and policy formulation is encouraged (Sackett, *et al.*, 1991). Similarly, efforts to produce systematic reviews of clinical trials have been seen as a priority, with initiatives such as the establishment of the international Cochrane Collaboration (Chalmers & Altman, 1995).

A central component of the design of any trial is the choice of outcome measure that is used in deciding the success or otherwise of a healthcare intervention. Therefore in applying the results of a trial in clinical practice or in formulating healthcare policy, a core consideration is not just the choice of experimental method used by researchers, but also the choice of outcome measure. For example, Sackett, *et al.* (1991) suggest that in judging the applicability of a clinical trial, a fundamental judgement must be made about whether all clinically relevant outcomes were recorded, including quality of life.

The previous section outlined the diversity of methods that are available to researchers when measuring outcome. There is a danger that outcome may be solely assessed by a limited method, such as by counting the frequency or attempting to measure the severity of psychopathological symptoms associated with common psychiatric disorders, without reference to how these symptoms impact on the individual and how they live their lives. A survey was therefore undertaken in

order to establish the methods that are used in measuring outcome in high quality epidemiological research – randomised clinical trials.

Aims of the survey:

1. To examine the methods that are used in measuring outcome in randomised controlled trials in psychiatry.
2. To examine which, if any, patient based measures are used to measure outcome in randomised controlled trials in psychiatry.
3. To examine how the measurement of outcome has changed over time in randomised controlled trials in psychiatry.

Chapter 12 Methods of the survey of outcomes measurement in psychiatric trials

An empirical survey of controlled trials was conducted, using high quality systematic reviews of randomised trials as a sampling frame for this survey – the Cochrane Database of Systematic Reviews (The Cochrane Database of Systematic Reviews, 2000). A number of topic areas were examined in more detail, in order to provide illustrative examples of patterns that were apparent in the measurement of outcome in clinical trials. Throughout the following section, a contrast will be drawn between two divergent methods of measuring outcome: (1) the use of *symptom based* clinical measures that count or measure the frequency or severity of symptoms of psychiatric disorders, and (2) *patient based* measures which examine the impact of psychiatric disorders on the individual and their quality of life.

Survey method

Target population

The target population for the purposes of this survey was defined as:
Randomised trials of interventions for common functional psychiatric disorders.

Trials relating to the following were therefore excluded:

- Drugs and alcohol problems;
- Child and adolescent populations;
- Cognitive impairment.

Sampling frame

The sample frame for the purposes of the survey was randomised controlled trials included in systematic reviews conducted within the Cochrane Collaboration. Two specific Cochrane groups conduct systematic reviews of interventions in mental health: the Cochrane Schizophrenia Group (CSG) and the Cochrane Depression, Anxiety, and Neurosis Group (CCDAN). Together, these two groups conduct reviews that cover the major diagnostic groups suffering from functional psychiatric disorders.

Cochrane reviews were chosen as a sample frame for the following reasons of practicality, ease of data collection and convenience:

- Cochrane reviews have each judged the methodological quality of their component trials, particularly with respect to randomisation, therefore ensuring only randomised trials be included in the survey.

- In the course of completing a review, researchers are required to record the outcomes measures that are reported in the individual component trials.

- Hard copies of each of the component trials are held in the relevant editorial bases of the respective review groups, allowing further information to be sought, and ambiguous outcomes to be checked.

All reviews published in the Cochrane Library, up to and including issue 2 2001, were sampled. In total, twenty complete reviews conducted under the auspices of the CCDAN, and 59 complete reviews conducted under the auspices of CSG were available for the survey. All potentially relevant CCDAN reviews were included, and a random sample of half of the CSG reviews was taken.

Data collection

For each component trial, the following were sought:

1. Year of publication

2. Mental health problem: the specific disorder of population under examination was recorded, and these were classified into (i) depression, anxiety and related disorders, and/ or (ii) schizophrenia or other severe mental illness.

3. Intervention: the specific intervention under examination was recorded, and these were classified into (i) drug treatments or physical interventions (ii) psychosocial interventions, or (iii) health policy interventions.

4. Standardised outcomes measures used

All standardised instruments used to measure outcome within each trial were recorded. Standardised outcomes instruments were defined as those using an interview schedule or questionnaire format, which was administered in a defined and reproducible manner. Unpublished rating scales, particularly those produced for the purposes of the study, without reference to published literature on the psychometric properties of that instrument were considered as non-standardised measures of outcome, and were not included in this survey.

Each standardised outcome was then subsequently classified into one of the following categories:

A. **Psychopathological rating scale:** defined as a scale or instrument that predominantly measured symptoms association with a common functional psychiatric disorder.

B. **Global outcome measure:** a measure which gave an overall appraisal of disease severity, with reference to the global severity of the disorder or its impact on overall functioning, rather than by counting the number or frequency of individual symptoms associated with as disorder. Examples of this form of outcome measure include the Global Assessment of functioning (GAF) (American Psychiatric Association, 1994) and Global Assessment Scale (GAS) (Enndicot, *et al.*, 1976).

C. **Generic patient based outcome measure:** a measure which examines several domains of health status or health related quality of life, and which is designed to be applied across different population, irrespective of illness or diagnosis. Examples include the Short Form 36 – SF36 (Ware, *et al.*, 1993) or Sickness Impact Profile - SIP (Bergner, *et al.*, 1976).

D. **Disease specific patient based outcome measure:** a measure which examines several domains of health status or health related quality of life, and which is designed to be applied to specific patient groups or a specific disease category.

E. **Domain specific patient based outcome measure:** a measure that examines a specific domain associated with health status or health related quality of life. For the purposes of this survey, the domains identified by Ware (1995) considered to be

the core components of health related quality of life, and include: physical health; social functioning; role functioning; general perceptions of well-being; cognitive capacity. In addition, satisfaction with treatment or healthcare services was included as a domain that is sometimes considered to be a facet of patient based outcome, particularly in mental health (Ruggeri, 2001) – see table 4 for operational definitions of these domains.

F. Other outcomes: in addition to the above, the presence of the following, as outcomes in individual component trials was recorded: relapse; mortality; service use.

Data were extracted from the summary reports of outcomes used in individual trials, as reported in Cochrane Systematic reviews. The content of individual outcomes measure was judged from one of several reference textbooks (Bech, *et al.*, 1993; Bowling, 1995; Bowling, 1997; McDowell & Newell, 1996; Sederer & Dickey, 1996; Thompson, 1989), prior to categorisation, as outlined above. Where this could not be established, clarification regarding content was sought by reference to the original paper.

Data were entered into a custom designed Microsoft Access relational database (Microsoft Corporation, 1998).

Quality assurance

Since the survey relies on the extraction of data from reviews conducted by others, a random sample of 5% of the original trials were obtained and cross checked in order to establish the reliability with which the presence of standardised outcomes measures had been established within Cochrane systematic reviews. Systematic reviews found to have poor reporting of standardised outcomes instruments were then subject to verification by reference to original component studies. Poor reporting was operationally defined as missing more than one standardised outcomes measure.

Data analyses

Descriptive statistics regarding the frequency and type of outcomes were calculated using Microsoft Excel spreadsheets (Microsoft Corporation, 1997a). Specific comparisons were made in order to examine whether outcome was measured in

different ways according to different diagnostic categories or according to different treatments being evaluated. Trends over time with respect to method of outcome measurement were undertaken by weighted regression techniques, using the StatsDirect commercial statistical package (Buchan, 2000).

Chapter 13 Results of the survey of outcomes measurement in psychiatric trials

Included studies

In total 490 individual trials were identified. The topic area of individual reviews and their component trials is given in Table 8. A total of 233 studies of interventions for schizophrenia and related disorders and 257 studies of interventions for depression, anxiety and related disorders were included. Year of publication ranged from 1956 to 2000. The annual publication of studies rose over time (see figure 1).

Figure 1: Publication of trials over time

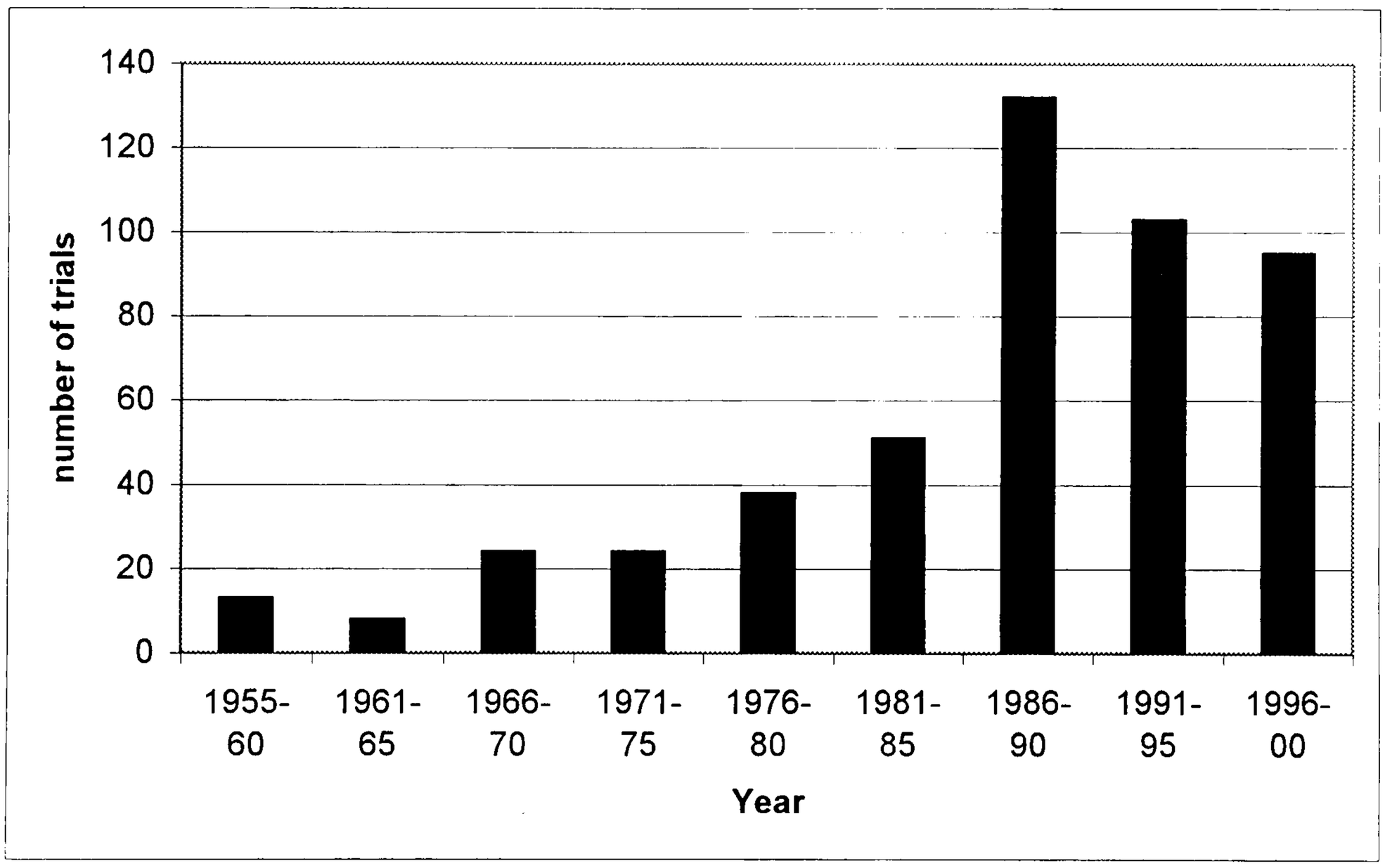


Table 8 Interventions examined in the survey of outcomes measures in clinical trials

Schizophrenia and related severe mental disorders:

Anticholinergic medication for neuroleptic-induced tardive dyskinesia

Assertive community treatment for people with severe mental disorders

Benzodiazepines for neuroleptic-induced tardive dyskinesia

Beta-blocker supplementation of standard drug treatment for schizophrenia

Carbamazepine for schizophrenia and schizoaffective psychoses

Calcium channel blockers for neuroleptic-induced tardive dyskinesia

Case management for people with severe mental disorders

Chlorpromazine versus placebo for schizophrenia

Clotiapine for acute psychotic illnesses

Cognitive behaviour therapy for schizophrenia

Cognitive rehabilitation for people with schizophrenia and related conditions

Crisis intervention for people with severe mental illnesses

Depot bromperidol decanoate for schizophrenia

Depot fluphenazine for schizophrenia

Depot pipothiazine palmitate and undeclynate for schizophrenia

Droperidol for acute psychosis

Family intervention for schizophrenia

Length of hospitalisation for people with severe mental illness

Life skills programmes for chronic mental illnesses

Molindone for schizophrenia and severe mental illness

Olanzapine for schizophrenia

Psychoeducation for schizophrenia

Risperidone versus other atypical antipsychotic medication for schizophrenia

Risperidone versus typical antipsychotic medication for schizophrenia

Sertindole for schizophrenia

Zotepine for schizophrenia

Depression, anxiety and related disorders

Antidepressant drug treatment for postnatal depression

Antidepressant plus benzodiazepine for major depression

Antidepressant versus placebo for depressed elderly

Antidepressants for depression in people with physical illness

Antidepressants using active placebos

Brief psychological interventions ("debriefing") for trauma-related symptoms and prevention of post-traumatic stress disorder

Cognitive behaviour therapy for adults with chronic fatigue syndrome

Counselling for Depression in primary care

Drugs versus placebo for dysthymia

Lithium for maintenance treatment of mood disorders

Pharmacotherapy for Posttraumatic Stress Disorder

Psychosocial and pharmacological treatments for deliberate self harm

SSRIs versus other antidepressants for depressive disorder

St John's wort for depression

General overview of method of outcome measurement

The majority of studies examined outcome using a standardised symptom based outcome measure. Global measures were also used commonly to measure outcome. Patient based outcome (generic, disease specific or domain specific) was not measured in the vast majority of studies. Table 9 summarises the methods adopted in measuring outcome amongst trials, with a breakdown according to patient or diagnostic group, and by type of intervention. The specific patterns of outcome measurement are explored in more detail below.

Psychopathological rating scales

Symptom based psychopathological rating scales were the most commonly used standardised outcome measure for all disorders and interventions. They were used more commonly for drug-based interventions, compared to psychosocial interventions in both schizophrenia (79.9% vs 49.2%, difference = 30.5%, 95% CI 17.2 – 43.5%), and depression, anxiety and related disorders (90% vs 70.3%, difference = 19.7%, 95% CI 0.06% – 36.2%). For schizophrenia and related disorders, the most commonly used measures were: the Positive and Negative Syndrome Scale (Kay, 1991), and the Brief Psychiatric Rating Scale (Overall & Gorham, 1962). For depression and related disorders, the most commonly used measures were: the Hamilton Depression rating scale (Hamilton, 1967)

Global measures

Global measures were used in less than half of all trials. They were more commonly used in drug trials, than in psychosocial interventions, for both schizophrenia (54% versus 21.7%, difference = 32.3%, 95% CI 18.9% - 43.2%), and depression, anxiety and related disorders (35.0% versus 10.0%, difference = 25.0%, 95% CI 9.2% - 33.9%). The most commonly used measures were the Global Assessment of Function (American Psychiatric Association, 1994), and the Global Assessment Scale (Enndicot, *et al.*, 1976).

Generic patient based outcomes measures

In contrast to symptom based and global measures, there was little evidence of the use of generic patient based outcome measures, with approximately 1% of trials using these measures. Those that were used were the Short form 36 (Ware, *et al.*,

1993) - n=5, the Dartmouth COOP (Nelson, *et al.*, 1990) – n=1. These were used in both drug based and psychosocial interventions conducted in the mid to late 1990's.

Disease specific measures.

In contrast to generic measures, there was evidence that a substantial minority of trials of interventions for schizophrenia and related disorders used a disease specific measure. Psychosocial interventions were evaluated more commonly than drug-based interventions using disease specific measures (15.9% versus 2.5%, difference = 13.4%, 95% CI 5.8 – 24.0%). The survey found no examples of disease specific patient based measures being used to evaluate interventions for depression, anxiety or related disorders. The measures used were the Heinrichs Quality of Life Scale – QLS (Heinrichs, *et al.*, 1984); the Lehman Quality of Life Interview (Lehman, 1983a); the Oregon Quality of Life Questionnaire – OQLQ (Bigelow, *et al.*, 1982).

Domain specific patient based outcomes measures

A substantial minority of trials in schizophrenia used a domain specific measure of patient based outcome, with 40% of psychosocial interventions using such a measure. The most commonly evaluated domain was that of social functioning (n=30), followed by cognitive functioning (n=10); role functioning (n=8) and perceptions of wellbeing.

Social functioning was largely measured using four major scales: the Social Adjustment Scale (Weissman & Bothwell, 1976); Katz Adjustment Scale (Katz & Lyerly, 1963); Social Functioning Scale SFS (Remmington & Tyrer, 1979); REHAB scale (Baker & Hall, 1988).

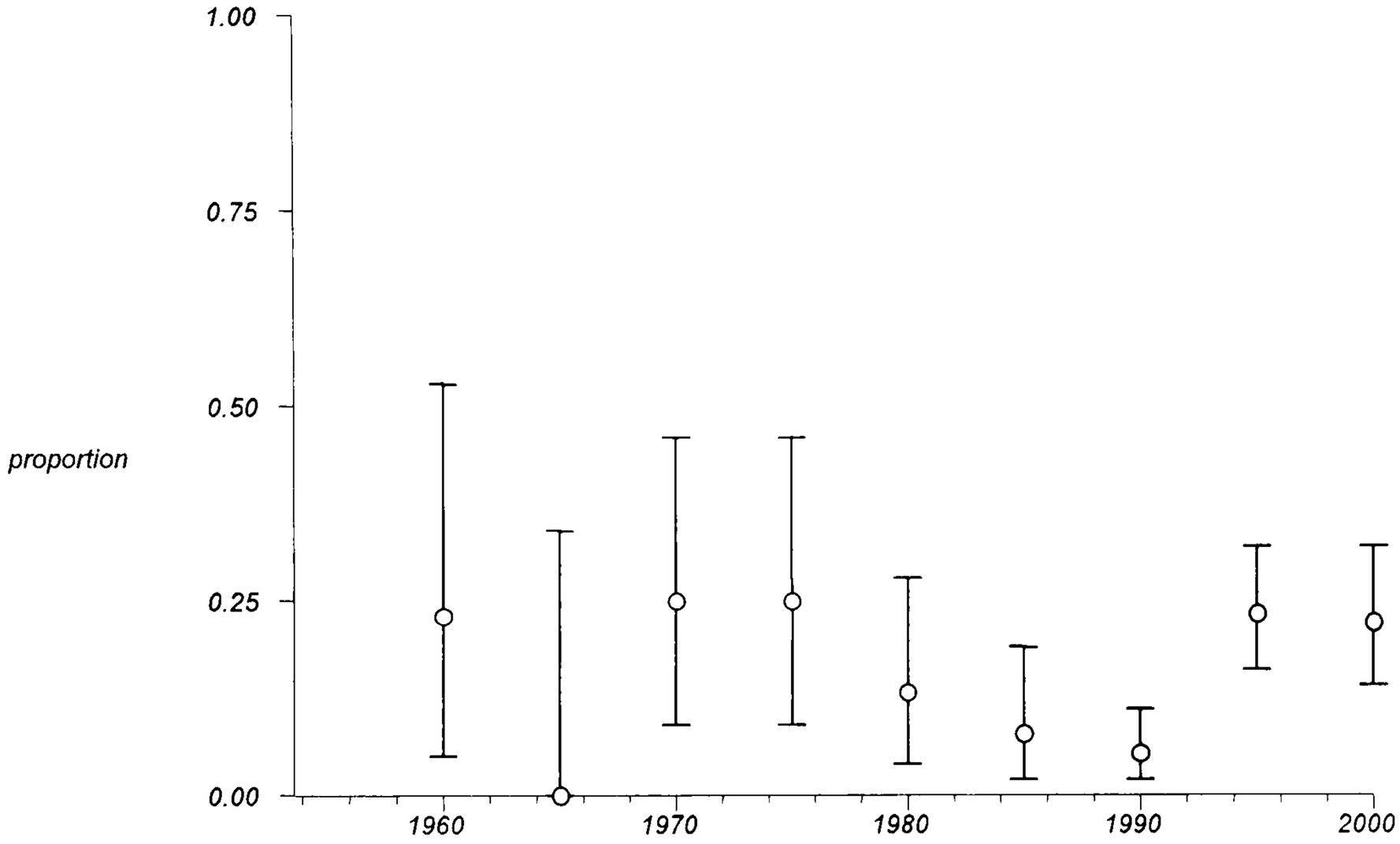
Domain specific measures were much less commonly used in trials for depression, anxiety and related disorders. Details of domains measured and instruments used are given in table 10.

Trend over time in the measurement of outcome

In order to examine changes over time in the measurement of outcome, all patient based measures (generic, disease specific, and domain specific) were conflated,

and the presence or absence of such a measure was recorded for each trial (figure 2).

Figure 2: Proportion of trials using a patient based outcome measure, measured over time



Five-year periods indicated, together with 95% confidence intervals

Trends over time in terms of the measurement of patient based outcome were examined by pooling five year periods from 1955 onwards, and conducting a regression of year (Yr) against proportion of studies using a patient based measure - weighted by the absolute number of trials in any five year period. The regression analysis showed no increase in the use of these measures over time ($r^2 = 0.002$, two sided $p = 0.9$).

A feature of the plot of use of patient based measures over time is the observation that studies prior to 1970 used patient based measures, whereas those conducted between 1980 and 1990 did so less frequently. Coincident with this observation is the finding that randomised trials conducted during the 1980s were dominated by

drug trials (Figure 3) comparing new anti-depressants with older tricyclic drugs. In these trials, patient based outcome was very rarely measured, and the sole criterion for success was a statistically significant change on a symptom-based measure such as the Hamilton Depression Rating Scale (Hamilton, 1967). The use of symptom-based measures in drug trials is discussed below.

Figure 3: Type of intervention in trials measured over time

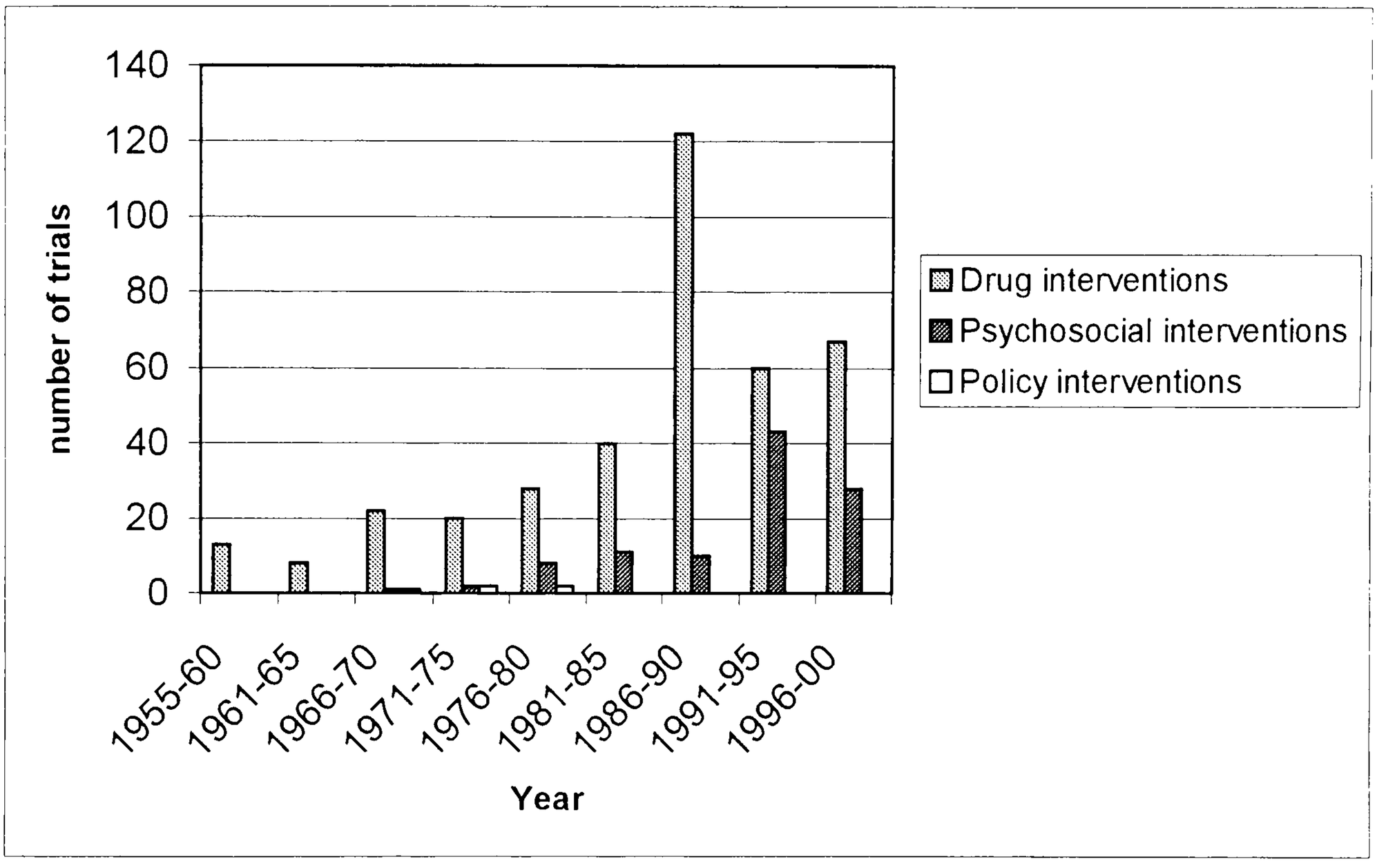


Table 9: Overview of outcomes measures used in a survey of 490 randomised trials

		Symptom based measure	Global measure	Generic	Disease specific	Domain specific	mortality	Service use
Schizophrenia and related disorders	**All interventions (n=233)**	164/233 **70.4%**	102/233 **43.8%**	2/233 **0.8%**	15/233 **6.4%**	48/233 **20.6%**	21/233 **9.0%**	50/233 **21.5%**
	Drug interventions (n=159)	127/159 **79.9%**	86/159 **54%**	1/159 **0.6%**	4/159 **2.5%**	18/159 **11.3%**	5/159 **3.1%**	4/159 **2.6%**
	Psychosocial Interventions (n=69)	34/69 **49.2%**	15/69 **21.7%**	1/69 **1.4%**	11/69 **15.9%**	28/69 **40.6%**	16/69 **23.2%**	41/69 **59.4%**
	Policy interventions (n=5)	3/5 **80%**	1/5 **20%**	0	0	2/5 **40%**	0	5/5 **100%**
Depression, anxiety and related disorders	**All interventions (n=257)**	224/257 **87.2%**	81/257 **31.5%**	5/257 **1.9%**	0	10/257 **3.8%**	4/257 **1.5%**	10/257 **3.9%**
	Drug interventions (n=220)	198/220 **90.0%**	77/220 **35.0%**	3/220 **1.4%**	0	6/220 **2.7%**	3/220 **1.4%**	3/220 **1.4%**
	Psychosocial Interventions (n=37)	26/37 **70.3%**	4/37 **10.0%**	2/37 **5.4%**	0	4/37 **10.8%**	1/37 **2.7%**	7/37 **18.9%**
	Policy interventions (n=0)	0	0	0	0	0	0	0

Table 10: Domain specific patient based measures used in 490 randomised trials

	Domain	Frequency	Instruments used
Schizophrenia and related disorders (n=233)	**Physical Health**	nil	
	Social Functioning	30/233	Social Adjustment Scale SAS (Weissman & Bothwell, 1976) – n=18; Katz Adjustment Scale (Katz & Lyerly, 1963) – n=5; Social Functioning Scale SFS (Remmington & Tyrer, 1979) – n=3; REHAB scale (Baker & Hall, 1988) – n=2.
	Role functioning	8/233	Index of activities of Daily Living (Katz, *et al.*, 1963) – n=6
	Perceptions of well-being	8/233	Rosenberg's Self Esteem Scale (Rosenberg, 1979) – n=6
	Cognitive functioning	10/233	IQ and intelligence tests - various
	Satisfaction	6/233	Client Satisfaction Questionnaire (Larsen, *et al.*, 1979) – n=4
Depression, anxiety and related disorders (n=257)	**Physical Health**	1/257	Pain and Disability Index (Tait, *et al.*, 1987)
	Social Functioning	9/257	SAS (Weissman & Bothwell, 1976) – n=6; Katz Adjustment Scale (Katz & Lyerly, 1963) – n=3
	Role functioning	4/257	Karnofsky Perfomance index (Karnofsky, *et al.*, 1948) – n=2
	Perceptions of well-being	nil	none
	Cognitive functioning	2/257	IQ and intelligence tests - various
	Satisfaction	1/257	Client Satisfaction Questionnaire (Larsen, *et al.*, 1979) – n=1

Three case examples of methods used in measuring outcome in a specific area of psychiatric treatment and healthcare delivery:

In order to illustrate some overall patterns in the measurement of outcome in psychiatric trials, three case examples are chosen: the evaluation of new drugs for schizophrenia and depression, and the use of specific models of community care for those with severe mental disorders

New drugs for schizophrenia and depression

The 1980s and 1990s have seen an intense period of research activity, largely by the pharmaceutical industry, with the emergence of new classes of first, anti-depressants (Serotonin Specific Re-uptake Inhibitors), and then anti-schizophrenia drugs (atypical anti-psychotics).

Almost one hundred trials have been located and included in a recent systematic review of trials comparing new and older anti-depressant drugs (Geddes, *et al.*, 2001). Amongst these trials, the primary endpoint of interest to researchers is consistently the suppression of depressive symptoms, measured using a handful of rating scales, applied serially over a six-week period. No trials included in this review measure broader health related quality of life, although ten trials do measure global outcome, using the CGI.

A similar pattern of outcomes measurement is seen amongst trials conducted to examine the comparative effectiveness of new atypical anti-psychiatric drugs in the care of those with schizophrenia. Several Cochrane reviews have been conducted into individual drug entities (Bagnall, *et al.*, 2001a; Duggan, *et al.*, 2001; Gilbody, *et al.*, 2001a; Kennedy, *et al.*, 2001; Lewis, *et al.*, 2001a; Lewis, *et al.*, 2001b; Srisurapanont, *et al.*, 2001; Tuunainen & Gilbody, 2001), and these have recently been collated in a review of the value of atypical drugs to the UK National Health Service, commissioned by the Health technology Assessment Programme (Bagnall, *et al.*, 2001b). All trials use symptom-rating scales as their primary outcome of interest, with successful treatment being operationally defined as a 50% shift from baseline score on one of two rating scales (the BPRS and the PANSS) over a six-week period. Approximately half also measure global functioning. However, 27% of trials also measure social functioning on one of the available clinician rated scales. Only two published

trials of new anti-psychotics, which have thus far been included in systematic reviews, measure broader quality of life. One published trial incorporates the Short Form 36, whilst three use a mental health specific quality of life measure.

The dominance of short-term measurement of psychiatric symptoms as a primary outcomes measure is explored in more detail in the discussion.

Case management and assertive community treatment for severe mental disorders

Two important reviews included in the present survey examine the use of different models of community care for those with severe mental illness, including schizophrenia and related disorders (Marshall, *et al.*, 2001; Marshall & Lockwood, 2001). Trials included in these two reviews measured psychiatric symptoms much less frequently than in drug trials, but instead measured a much more broad range of patent based outcomes. Symptom based scales were used in only one third of trials, whereas quality of life measures (e.g. Lehman's Quality of Life Scale (Lehman, 1983b) and or domain specific measures (especially social function, role function, self esteem, and satisfaction) were measured in over half of included studies.

A large majority of studies also measured simple aspects of service use, such as hospital admission and length of stay, in addition to using standardised measures of outcome. The clear focus in a number of these trials was not just psychiatric symptoms amongst persons with often-chronic disorders, but rather the impact of their illness and attendant symptoms on how they lived their day-to-day life, and their need for services.

Chapter 14 Discussion of the survey of outcomes measurement in psychiatric trials

Survey methods

The survey used a ready available source of systematic reviews of randomised trials as its sampling frame. The generalisabilty and representativeness of these data deserves further consideration. The survey was not a random survey of randomised trials in mental health, but rather a survey of trials that have been included in systematic reviews.

The advantages of this approach include the fact that data had been already extracted in a standardised way by those preparing systematic reviews under the auspices of Cochrane guidelines (Mulrow & Oxman, 1999). This allowed a much larger number of trials to be sampled in a much shorter period of time, and at less cost than would have been possible if primary studies had been used in the first instance. The quality of the data was monitored within the survey and was found to be acceptable.

The danger from using data from systematic reviews is the fact that studies included in reviews may not be representative of studies that are conducted in general. However, it might be argued that the studies included in systematic reviews represent the trials which individuals find the most important in the context of their day to day work and decision making processes, since topics are likely to have been selected for the process of systematic review for these reasons rather than for intellectual curiosity alone. One particular limitation that needs to be borne in mind with, for example, Cochrane reviews in the sphere of schizophrenia, is that particular effort has been made to review each of the new drug entities and that particular funding has helped this process received via the NHS Health Technology Assessment Programme (www.hta.nhsweb.nhs.uk/). The result of this reviewing activity might be that the trials included in the survey might include a disproportionate number of trials conducted into new drugs and that these might not be representative of schizophrenia trials as a whole. The summary statistics presented showing the proportion of studies measuring outcome in one specific way need to be viewed with this potential bias in mind. However, the survey has revealed important trends that have been examined in more detail in the case examples. Of note is also the fact that some of the

results found in this survey are in line with trends revealed in an earlier survey by (Thornley & Adams, 1998), which limited its scope to schizophrenia trials.

Main findings

The main findings of the study are that the dominant method of outcomes measurement in randomised trials in psychiatry remains symptom based psychopathology scales. The increasing popularity of generic patient based measures, such as the Short Form 36 – available since the early 1990s, is not reflected in psychiatry. Similarly, the existing quality of life measures developed specifically for those with mental illness have largely not been included as measures of outcome in clinical trials. The findings of this survey therefore mirror the findings of other surveys of patient based outcomes measurement in other specialities. Sanders, *et al.* (1998) in a survey of trials included in the Cochrane Controlled Trails Register (The Cochrane Controlled Trials Register, 2000) found that less than 5% of trials overall (in any speciality) use patient based measures. The main exception to this being cancer and cardiovascular disease trials, where 29% and 26% of trials respectively used a patient based measure. Psychiatry seems therefore no worse then the majority of specialities in its use of patient based measures in evaluative research.

However, there remain a substantial minority of trials where symptom based measurement is supplemented by the measurement of domains that can be considered facets of patient based outcome. Specifically these include social and role functioning. The survey shows that these instruments are commonly used in trials of psychosocial interventions for those with mental illnesses, such as schizophrenia. These instruments have been in existence for many years, and have therefore been available to research as outcomes instruments in evaluative research. The present survey shows that patient based outcomes measurement is therefore not a new phenomenon in psychiatry, and that domains of patient based outcome, such as social functioning, have been incorporated into trial designs since the 1960s.

Wiersma (1996) outlines two reasons why social functioning has traditionally been of interest and has come to be measured as an outcome in its own right. Firstly, the trend towards community oriented care models required careful evaluation, with respect to its consequences. In order to judge the consequences of community versus hospital treatment, a separate series of

measures is justified for those with chronic and enduring mental disorders whose social functioning has traditionally been poor. Secondly, there is evidence that disease progression, symptomatology and social dysfunction may vary relatively independently. Social disablement of a patient may be characterised much more by using measures of social disabilities than by measures of psychiatric symptoms. Further, interventions targeted at social disability may be successful in helping gain or maintain independence, whilst having little impact on psychotic symptoms.

The dominance of symptom-based measures in the majority of the trials surveyed in this review deserves further discussion. Commentators on the measurement of outcome in other specialities such as rheumatology and oncology (e.g. Albrecht 1994) have drawn attention to the fact that the development and use of 'patient based measures' represent a reaction to the over dependence on biophysical measures of outcome, which say little about the impact of disease or illness on the individual. This analysis may not directly relevant to psychiatry, since there have not been biophysical correlates of disease, which can be appropriated as (limited) measures of outcome. However, the over reliance upon instruments which record the number or severity of the individual 'symptoms' associated with a specific psychiatric disorders might be seen as analogous. The use of 'symptom based' measurement scales as an outcome, without reference to how these symptoms affect the individual in their day to day life provides a similar scenario to that seen in wider healthcare. The present survey therefore lends empirical support to observations that have been made previously, for example in the sphere of schizophrenia research, Collins, *et al.* (1991) has stated that:

'*A recurrent criticism of measurement in schizophrenia research is that symptom suppression is overemphasised as the sole criterion measure of treatment effectiveness, to the neglect of other endpoints, such as the quality of life and subjective experience of the patient*'

This observation is especially true in the case of drug trials in psychiatry. The success or otherwise of new drugs is almost entirely measured using symptom based measures, without reference to the value of these new and relatively expensive new technologies in terms of wider quality of life. One example of this comes from a widely disseminated and cited trial of the value of one new anti-schizophrenia drugs – olanzapine, manufactured by Eli Lilley. This

industry-sponsored trial is one of the largest drug trials ever conducted in psychiatry, with almost 2000 participants (Tollefson, *et al.*, 1997). The outcomes used in this trial included four symptom based measures and a series of standardised assessments of side effects, each of which were applied every two weeks. In total, two million questions were asked of its nearly 2000 participants, but failed to ask whether patients felt they were substantially better {Professor Clive Adams, personal communication}. The main cause of this over-dominance of symptom based measures is likely to be the fact that these trials are essentially designed to meet the demands of drug licensing authorities, such as the US Food and Drug Administration, and the UK Medicines Control Agency. These bodes require evidence of the value of a new drug entity (efficacy), and are happy that this is demonstrated by the use of symptom based measures. They make no demands that effectiveness or the ability to make substantial changes to patients' wider health related quality of life should be demonstrated before granting product licence (FDA, 1997). There is therefore no economic incentive to conduct trials which measure patient based outcome.

Suggestions for further research

The present survey has demonstrated that there is a dominance of symptom based instruments in the measurement of outcome in clinical trials. This is despite the existence of disease specific and generic patient based measures. This prompts two main topics for further research.

First, fundamental research is needed into the suitability of patient based measures for inclusion in clinical trials in mental health. Fitzpatrick, *et al.* (1998) have produced a general series of recommendations based upon a systematic review of the methodological literature surrounding patient based outcomes, which can be applied in all areas of health. They recommend that before inclusion in a trial, judgements should be made according to eight criteria: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability and feasibility. There is little point in including an instrument in a trial if it is valid and reliable, but shows no response to change in underlying dimensions of quality of life that are important to the patient. Similarly, many patient based measures are over-long or unacceptable to patients, and their addition to an already lengthy battery of questionnaires might prove too onerous to trial participants. For example, Lehman's QOLI, designed for persons with

severe mental disorders takes 45 minutes to complete (Lehman, 1983b). Generic patient based outcomes measures, such as the Nottingham health Profile or the Short Form 36, may be difficult to apply to patients with mental health problems if they concentrate on physical functioning by asking about an individual's ability to climb stairs, whilst ignoring those aspects of social and role functioning that are important in chronic and severe mental illness. They may therefore be insensitive to underlying change in health status, and may include large numbers of questions that are irrelevant to the individual, also making them unacceptable to respondents.

Clearly, the desirable attributes outlined by Fitzpatrick, *et al.* (1998) may be present for many measures and the fact that they are not included in trials represents an omission. A systematic summary of these attributes for available instruments, when used in populations with mental health problems is needed as a matter of urgency. Such a summary would be an invaluable resource for researchers and those who must interpret the meaning of research which uses patient based outcomes.

Second, despite the theoretical appeal of patient based instruments, in that they extend the measurement of outcome beyond symptom suppression, it remains to be demonstrated if the results of trials are substantially different when they are used. If the results of trials are in fact substantially different according to how outcome is measured, then there needs to be an examination of what should be the primary endpoint of trials, and which results should be used in decision making processes, which incorporate trial based evidence.

It had been anticipated that the current survey would provide sufficient examples of trials that measure both patient based and symptom based outcomes, in order that this question be examined empirically. Unfortunately this proved not to be the case. An ongoing piece of research by the present author seeks to examine this question further (see appendix 1 for a protocol of this study)

Section 2.3 Outcomes Research in Psychiatry

Chapter 15 Background to the survey of outcomes research in psychiatry

Randomised controlled trials have generally been accepted as the 'gold standard' design when deciding what interventions work in psychiatry (WHO, 1991). Most randomised studies in psychiatry have investigated the effect of drug or psychotherapy interventions in tightly controlled and largely artificial experimental conditions (Hotopf, *et al.*, 1997; Thornley & Adams, 1998), while patients, clinicians and other decision-makers need to know how treatments work in the real world and whether they are cost effective under routine conditions (Wells, 1999). Important questions relating to the organisation and delivery mental health services are also rarely addressed in randomised trials (Gilbody & Whitty, 2001).

The need for research relating to effectiveness (rather than efficacy) has prompted a number of responses: One has been the call to conduct randomised trials in real-world settings, using *pragmatic designs* (Hotopf, *et al.*, 1999). Another has been to synthesise various data sources using *decision analysis* (Lilford & Royston, 1998). A response which has been highly influential in the United States in the past decade involves the analysis of large databases of patient data collected in routine care settings – known as *outcomes research* (Anonymous, 1989; Ellwood, 1988; Wennberg, 1991).

The origins of outcomes research.

Outcomes research forms a cornerstone of the outcomes movement discussed in section1, and outlined by Paul Elwood in his 1988 Shattuck lecture (Ellwood, 1988). In this lecture, he called for the routine collection of outcomes measures by clinicians, in order to create a *'technology of patient experience'*. He proposed that these data should be assimilated in large databases that would form a resource for clinical and health services research. Such data could eventually be used *inter-alia* to compare existing treatments and to evaluate new technologies, thereby avoiding both the expense of clinical trials and the loss of generalisability that resulted from the selective recruitment to conventional efficacy trials.

A core component of outcomes research, according to Elwood, was the type of outcomes that would be collected and analysed. Accorrding to Elwood:

'The centre piece and unifying ingredient of outcomes research is the tracking and measurement of functioning and well being or quality of life'. i.e. the collection of patient based outcomes.

The Agency for Health Care Policy and Research – AHCPR (now the Agency for Healthcare Research and Quality – AHRQ) was established under public law in 1989 in order to conduct 'outcomes research' into common medical conditions, with the establishment of Patient Outcome Research Teams - PORTs (Wennberg, *et al.*, 1993). The research programme was allocated $6 million in its first year, rising to $63 in 1991, with the purpose of using routine outcomes data to determine *'outcomes, effectiveness and appropriateness of treatments'* (Anderson, 1994). It was decreed by Congress, via the General Accounting Office, that new primary research conducted by the PORTs was not to be the traditional randomised controlled trial, rather it was to be observational in design, utilising the vast amounts of data routinely collected on US patients (General Accounting Office, 1992). This health research policy produced a new breed of health researchers; known as database analysts (Anonymous, 1989; Anonymous, 1992), with the motto *'Happiness is a humongous database'* (Smith, 1997).

Outcomes research differs from traditional observational or quasi-experimental research in a number of ways, particularly with respect to the outcomes that are used, and the setting in which these outcomes are collected. In outcomes research, competing interventions that are already used in routine care settings are compared by analysis of routine data collected by clinicians or by other agencies (such as insurance companies), whereas quasi-experimental studies implement interventions in one setting or amongst one group of patients, and compare outcomes with patients who have not been subject to the intervention (Gilbody & Whitty, 2001). Quasi-experimental studies are therefore more like randomised trials, and are considered to be clearly different in their approach and ethos to outcomes research (Aday, *et al.*, 1998). The outcomes that are studied in outcomes research are generally those that are already collected as part of routine care, although there is no reason why these cannot be in added to in the light of the specific question being asked.

Outcomes research in psychiatry

The previous survey of clinical trials has demonstrated the infrequency with which patient based outcomes are used. A clear aspiration of Elwood's was that outcomes research would address the limited methods by which outcomes are measured in traditional evaluative research. It would be expected that outcomes research in psychiatry might use a more patient based approach than has been demonstrated within this thesis.

Enthusiasm for outcomes research has, in the US, led to the establishment by the American Psychiatric Association of Practice Research Networks - PRNs (Zarin, *et al.*, 1997; Zarin, *et al.*, 1996). This initiative involves the recruitment of 1000s of practising psychiatrists, who will routinely measure a broad range of outcomes for their patients, in order to: provide benchmarking for practice, judge the extent and consequences of variations in practice; and to examine the effectiveness in real world settings of all manner of healthcare interventions – as an alternative to the randomised trial. There are advocates of outcomes research in non-US mental health services research, particularly in psychotherapy (Barkham, *et al.*, 1998; Guthrie, 2000; Marginson, *et al.*, 2000; Mellor-Clarke, *et al.*, 1999). Similarly, the pharmaceutical industry is keen to extend the method in the evaluation of new and relatively expensive drug therapies; for example the Schizophrenia Health Outcomes Study – SOHO, funded by Eli Lilly, aims to recruit European collaborators to collect outcomes from patients with schizophrenia in receipt of typical and atypical drugs. Others have urged caution (Sheldon, 1994), and the principle concerns that have been expressed about outcomes research include: (1) their observational (rather than experimental) design; (2) the poor quality of the data which are used; (3) the inability to adjust sufficiently for case mix and confounding; (4) the absence of clinically meaningful outcomes in routinely collected data (Iezzoni, 1997).

As in the survey of randomised trials reported in preceding sections, a key component in interpreting and using the results of research is the type of outcomes that are collected and presented. The purpose of the present research is to produce the first systematic survey of the use of outcomes research in psychiatry, since this has not hitherto been described.

Aims

1. To examine the specific types of outcomes that have been collected and used within outcomes research, in examining the effectiveness of interventions in psychiatry.

2. To examine the specific uses to which routinely collected data have been put in examining the effectiveness of interventions in psychiatry.

Chapter 16 Methods of the survey of outcomes research in psychiatry

Sources of outcomes research

No specific database of outcomes research was available for the conduct of this research, and the source of potentially relevant studies was therefore the large amounts of literature that were identified in the searches detailed in appendix 1.

Survey method

Target population

All examples of outcomes research which fulfilled the following inclusion and exclusion criteria.

Inclusion criteria

Reports were included if they fulfilled each of the following criteria:

1. The research was conducted in a care setting that was part of usual care in a healthcare system.

2. The outcomes data used were those collected routinely for all patients – either for administrative purposes, or as a means of monitoring outcomes within the service being evaluated.

Exclusion criteria

Studies were excluded if they fulfilled any of the following criteria:

1. Research that only examined the costs and processes of illness and healthcare from routinely collected data, with no linkage to the outcomes of care. For example, primary care prescription databases have been used to conduct research into newer psychotropic drugs (eg Donoghue, *et al.*, 1996), but since they are not linked to patient level data and outcomes, they cannot be considered outcomes research.

2. Quasi-experimental or non-randomised evaluations of new technologies, where an intervention is implemented and outcomes measurement systems established only in the course of its evaluation (Cook &

Campell, 1979). For example, the PRiSM Psychosis study (Thornicroft, *et al.*, 1998) is an example of a quasi-experimental evaluation of a model of community care for those with severe mental illness, where districts were non-randomly allocated to implement an experimental service, and outcomes were measured under experimental and control conditions *as part of the study*.

3. Studies that only examined the relation between patient characteristics and outcome, with no direct comparison between competing treatments or health policy strategies (e.g. Rosenheck, *et al.*, 1997).

4. Reports of routine outcomes measurement in practice, with no direct report of comparative service or treatment evaluations based on the data.

Sampling frame

Studies were drawn from all published studies that were included in the following databases, up to and including the following dates:
MEDLINE; EMBASE; Cinahl; British Nursing Index; Cochrane Controlled trials register – to June 2000. See appendix 2 for details of the search strategy used.

Sampling method

Studies formed a complete sample of all those identified using the search terms outlined in appendix 2.

Data extraction

Data were extracted on the following:

- Population.
- Clinical or organisational question being asked
- Setting
- Sample size and length of follow up
- Outcomes studied, and source of outcomes studied.
- Adjustment for case mix and confounding
- Results

Data synthesis

It was anticipated that relatively few examples of outcomes research would be identified. The principle form of data synthesis was a descriptive overview of major trends in terms of the following:

- Outcomes studied, and source of outcomes studied.

With due consideration of:

- Clinical or organisational question being asked
- Setting
- Sample size and length of follow up
- Adjustment for case mix and confounding
- Results

Salient examples were used to illustrate trends, particularly in terms of outcomes measurement.

Chapter 17 Results of the survey of outcomes research in psychiatry

Despite the widespread advocacy of outcomes research in healthcare, relatively few published examples relating to mental health were found. Several of these studies were published in the past three years, highlighting an increase in the use of the design. The scope, design and analysis of the studies we identified is summarised in table 11. In the following section important characteristics these nine studies are reviewed.

Research questions addressed

Outcomes research has been used in broadly two areas of mental health research:

(1) The evaluation of mental health policy, including aspects of service delivery, organisation and finance

The earliest and perhaps most important example of outcomes research in mental health is the Medical Outcomes Study (MOS) conducted by the RAND corporation in the United States in the late 1980s (Tarlov, *et al.*, 1989; Wells, *et al.*, 1989; Wells, *et al.*, 1996). The design and objectives of this study were shaped by US health care policy debates: on the role of financing and reimbursement strategies in private care (fee for service versus pre-payment), and on the place of speciality (secondary) care.

The authors justified the use of observational methods in two ways. First, the authors claimed that the cheaper design and reduced burden on participants could maximise the number and range of collaborators and patients, particularly from non-research settings. Second, the authors claimed that the specific research questions precluded the use of randomisation, since the very act of randomisation would alter the functioning of existing health care delivery systems (Wells, *et al.*, 1996).

Three other studies researched health policy and organisation questions, such as the consequences of the withdrawal of mental health benefits from insurance plans (Rosenheck, *et al.*, 1999a); the effectiveness services directed at homeless persons (Lam & Rosenheck, 1999); the difference in outcome

between private and publicly funded health providers (Leslie & Rosenheck, 2000).

(2) The evaluation of new technologies.

Four studies (Croghan, *et al.*, 1999; Hong, *et al.*, 1998; Hylan, *et al.*, 1999; Melfi, *et al.*, 1998), utilised an outcomes research design to demonstrate the worth of new antidepressants and anti-psychotics in routine care settings. One further study (Rosenheck, *et al.*, 2000) examined the value of an innovative psychosocial intervention for those with war-related Post Traumatic Stress Disorder.

Source and choice of cases and outcomes

Outcomes studies can be broadly be divided into:

(1) Those which collect data prospectively on a service-wide level, where the choice of outcomes is decided a priori and is influenced by the research question or population under examination, and (2) those which utilise existing outcomes data, collected for other purposes.

The MOS is the best-known example of prospective outcomes research. The authors set out to measure patient-centred outcomes, in addition to clinician-rated depressive symptoms within existing healthcare services. The enduring legacy of the MOS is the fact that patient-centred measures of health status were developed for the study, and eventually evolved into the Short Form 36 (Stewert & Ware, 1992) – now the most commonly used generic measure of health related quality of life.

A further study (Rosenheck, *et al.*, 2000), measured multiple outcomes, including disease-specific measures relating to the underlying condition (Post Traumatic Stress Disorder), measures of social function, health-related quality of life, and service use. This study used a large and already existing dataset describing all of the 600,000 patients in receipt of mental healthcare under the US Veterans Administration (National Committee on Quality Assurance, 1995), supplemented with routinely collected disease-specific patient outcomes measures collected for all patients in receipt of care for PTSD (Rosenheck, 1996).

All the other studies that were identified utilised existing outcomes already entered on large administrative databases, studying a much more limited range of outcomes. For example, studies examining the value of new anti-depressants in routine care settings use a commercially available medical insurance database (eg MarketScan™) of linked pharmacy and medical claims data on 750,000 individuals (Croghan, *et al.*, 1999; Hylan, *et al.*, 1999; Melfi, *et al.*, 1998). Cases of depression were identified retrospectively, either from a reimbursement claim for anti-depressant medication or by the presence of one of six ICD codes indicative of depression. This approach is problematic, since antidepressants are commonly prescribed for a number of conditions other than depression (Streator & Moss, 1997). Similarly, depression is consistently under-identified by clinicians (Jencks, 1985), and mislabelled or underreported, in part as a consequence of the stigma of mental illness (Rost, *et al.*, 1994).

Administrative databases such as MarketScan™ also hold no direct information relating to disease severity, such as scores on symptom rating scales. Disease progression, relapse or remission cannot be directly measured and database studies are forced to use alternatives. For example, (Hylan, *et al.*, 1999) used continuous six-month claims for refills of prescriptions as a proxy measure of acceptable pharmacotherapy and therefore good outcome, ignoring the fact that patients discontinue medications for a whole host of reasons other than treatment failure.

Sample size and length of follow up

Sample size was generally much greater than that achieved in the traditional randomised trial, with a median sample size of n= 2678 (range 1034 to 20,814). Those studies that recruited subjects prospectively in the context of a study, such as the MOS (Wells, *et al.*, 1989), achieved smaller sample sizes (n=1772) than those which selected subjects retrospectively from large existing datasets (Croghan, *et al.*, 1999; Rosenheck, *et al.*, 1999a) - median n=4052. Periods of follow up were of median six months (range 4 to 48 months).

Adjustment for confounding and case mix

All studies made some attempt to describe and adjust for confounding factors, typically using some form of regression analysis, or propensity scoring (Rubin,

1997). Authors rarely reported each of the potentially confounding factors that were entered into their analysis – often restricting reports to those that were positive and related to outcome. However, it was clear that the ability of studies to adjust for confounding was determined by the collection or availability of suitable measures. Two studies serve to illustrate the contrast between limited and more complete adjustment for confounding.

The authors of the MOS prospectively measured a broad range of case-mix variables, including disease severity and co-morbidity, in addition to traditional demographic characteristics, such as age, sex and socio-economic status. This is especially important in the MOS since the type of healthcare provider is inexorably linked to disease severity, making unadjusted comparisons of outcome un-interpretable.

One of the more unexpected results of the MOS demonstrates the limitation of an observational approach and the need to measure and adjust for case-mix and confounding. In unadjusted samples, the receipt of any treatment (anti-depressant medication or counselling) was associated with a much worse 2-year outcome than the receipt of no treatment. In analyses that adjusted for baseline health differences, treated and untreated patients had a comparable 2-year outcome. In a subgroup analysis, designed to minimise unmeasured biases by restricting the analysis to those with the most severe depression, treatment was in fact associated with a significantly better 2-year outcome (Wells, 1999; Wells, *et al.*, 1996).

In contrast, outcomes studies based on administrative data are much more limited in their ability to measure and adjust for confounding. For example, in retrospective database studies of new anti-depressants (eg Hylan, *et al.*, 1999; Melfi, *et al.*, 1998) disease severity could not be measured since these data were not directly included in administrative data, and could only be crudely inferred from the setting in which care was given (primary versus secondary care).

Table 11: Examples of outcomes research in psychiatry

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Medical Outcomes Study (MOS) (Wells, *et al.*, 1989)	Depression (major depression, dysthymic disorder & sub-threshold depression) being managed in Family practices & specialist healthcare providers	1. How does treatment for depression differ by speciality and payment system? 2. How does outcome for depression differ by speciality and payment system? 3. How can care for depression become more cost effective?	Data routinely collected by clinicians and research workers during the course of the study on 1772 patients	Detection of depression by physicians Adequacy of treatment Depressive symptoms (incl Hamilton Depression scale scores) Health status (incl. Short Form 36)	Baseline demographic data and case-mix measured and adjusted for (including medical co-morbidity, psychiatric co-morbidity, past history of depressive episodes)	Depression is generally under recognised, inadequately treated and is associated with a poor level of functioning. Depression is associated with poorer quality treatment and outcome when a pre-payment plan is in place, rather than a Fee for Service.
Lam & Rosenheck (1999)	Severe mental illness amongst the homeless contacted through 'street outreach'	Is case management as effective for those homeless contacted on the streets, as for those contacted through shelters and other service agencies?	Routinely collected data from a five year, 18 site demonstration project which established and sought to evaluate outreach services for the homeless mentally ill (n= 5431) (Randolph, *et al.*, 1997)	Depressive and psychotic symptoms; alcohol and drug abuse; housing; paid employment; social support; quality of life and service use	Those in receipt of street outreach (n=434) were compared to those receiving conventional outreach after adjusting for baseline socio-demographic differences, and baseline differences in psychosis and substance abuse	Assertive outreach resulted in client improvement in 14 of 20 outcome indicators. These benefits persisted and were similar to conventional outreach, following adjustment for case-mix and confounding.

Table 11: Examples of outcomes research in psychiatry (continued)

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Rosenheck, *et al.* (1999a)	Mental health service use amongst enrolees in a health insurance plan following mental health spending cut backs	Do cutbacks of mental health coverage by an insurer result in increased non-mental health service utilisation and reduced productivity?	Employee work records and health care claims data relating to 20,814 employees in a single US corporation	Mental health and non-mental health service use (number of days of inpatient and outpatient healthcare). Healthcare costs Days absent form work.	Baseline differences between years in terms of socio-demographic factors, employment, income and state of employment	Reduction in mental health care utilisation was accompanied by a marked increase in non-mental healthcare service use and costs, and sick time.
Leslie & Rosenheck (2000)	Individuals in receipt of US public sector (VA) and privately insured inpatient mental healthcare. Followed up for six months following discharge	Is publicly insured healthcare of lower quality and associated with poorer outcome compared to privately insured healthcare	Routinely collected VA outcomes data were available on 180,000 inpatient episodes Routinely collected data from seven million privately insured lives were available on a commercially available databases (MEDSTAT MarketScan) – 6000 inpatient episodes were selected.	Length of stay Readmission rates (14, 30 and 180 days post discharge) Proportion receiving outpatient care	Adjustment made for known and measured confounders (age, sex, gender, diagnostic category, and psychiatric co-morbidity). No data available on important confounders, including socio economic status, employment, homelessness, health status and level of disability.	VA patients were older, and more prone to psychiatric illness. Quality indicators and outcome were poorer for VA care than privately insured care. The results are largely un-interpretable, given the observed difference may be real, or an artefact of casemix.

Table 11: Examples of outcomes research in psychiatry (continued)

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Rosenheck, *et al.* (2000)	US patients with chronic war related post-traumatic stress disorder being treated in veterans (VA) inpatient programmes. Followed up for 4 months	Is an innovative psychosocial treatment (Compensated work Programme – CWP) effective in routine care settings	Routine data for all patients in receipt of VA inpatient mental healthcare. Supplemented by disease specific measures collected for all patients in receipt of care for PTSD. Compete data on 542 patients in receipt of CWT, with 542 matched controls, in receipt of routine or standard care for PTSD.	PTSD symptoms; Substance abuse; Violent behaviour; employment and medical status	Matching patients to controls by selecting those characteristics that predict participation in the intervention condition (Propensity scoring Rubin, 1997). Logistic regression of baseline differences on PTSD symptom scores between CWP patients and controls.	CWP has no impact on any of the outcomes measured, compared to controls, when adjusted analyses were conducted. The treatment is likely to be clinically and cost ineffective. A formal randomised trial is not justified on the basis of this observational study.

Table 11: Examples of outcomes research in psychiatry (continued)

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Melfi, *et al.* (1998)	US patients in receipt of anti-depressant medication for depressive disorders	Does adherence to anti-depressant treatment guidelines prevent the relapse and recurrence of depression?	Compliance with treatment guidelines operationally defined as having made a claim for four or more antidepressant prescriptions over a six month period following initiation of medication. 4052 patients classified into one of three groups, according to whether they met this criterion from Medicaid claims records.	Relapse or recurrence during an 18 moth follow up period was defined as the initiation of a new anti-depressant prescription; or by evidence of a suicide attempt, hospitalisation, mental health related emergency room visit, or receipt of electro-convulsive therapy.	A series of general co-morbidity adjustments were made using hospitalisation for any other physical disorder, together with demographic variables. Severity of depression was controlled for using proxy measures, including whether an individual was seen by a mental health specialist.	Patients with 4 or more prescriptions of anti-depressants were less likely to relapse

Table 11: Examples of outcomes research in psychiatry (continued)

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Croghan, *et al.* (1999)	Depression being managed in primary care	Does specialist referral for psychotherapy improve compliance with anti-depressants, compared to those managed exclusively in a primary care setting?	A commercially available medical insurance database (MarketScan ™) of linked pharmacy and medical claims data on 750,000 individuals. Those with complete claims data, and a new prescription of antidepressants were followed up over 12 months from initiation of prescription (n=2678)	Use of anti-depressants ascertained from claims. Continuous medication use over 6 moths is taken to be a proxy measure of effective antidepressant therapy and good outcome (Agency for Health Care Policy Research, 1993). Total healthcare costs were also measured from cost claims data.	There were substantial differences between those in receipt of care in primary and specialist settings in terms of age, sex, and previous history of depression. Previous claims. Hospitalisations and diagnoses of depression; used to adjust, using logistic regression	Referral to a specialist increases the chance of receiving continuous anti-depressant therapy by 11% in adjusted analyses. The authors calculate cost effectiveness ratios to achieve this benefit, and conclude that continuous medication is likely to be a good proxy measure of improved outcome.

Table 11: Examples of outcomes research in psychiatry (continued)

Author/study name	Clinical problem/population and setting	Clinical or organisational question or hypothesis being examined	Source of outcomes data & Sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Hylan, *et al.* (1999)	Patients in receipt of pharmacotherapy for depression in primary care settings	Is there a difference between different serotonin specific re-uptake inhibitor (SSRI) anti-depressants in terms of patient compliance?	A commercially available medical insurance database (MarketScan™) of linked pharmacy and medical claims data on 750,000 individuals. Complete episodes available on 1034 patients in receipt of a new SSRI prescriptio.	Continuous prescription of the same anti-depressant, without dosage change, or switch between different drugs or drug classes over six months was taken a proxy measure of a successful initial choice of anti-depressant.	Logistic regression of available confounders included: demographic details; severity of depression from ICD codes; co-morbid drug and alcohol problems; co-morbid physical disorder (counts of other ICD codes); provider characteristics (primary care or specialist).	Patients in receipt of fluoxetine were more likely to receive continuous prescriptions over a six-month period, when compared to sertraline or paroxetine. The authors conclude that fluoxetine is better tolerated than either sertraline or paroxetine.
Hong, *et al.* (1998)	US patients with relapsing schizophrenia and high levels of healthcare resource use.	Is a newer anti-psychotic (quetiapine) associated with better compliance, and therefore lower rates of re-hospitalisation, when compared to conventional treatment?	Those with schizophrenia (n= 1400) selected from the MarketScan ™ claims database, coupled with a Medicaid claims file, providing detailed healthcare costs and resource use on 5% of the 5 million Californian Medicaid population.	Hospital readmission rates and the prevalence of high service utilisation were calculated. The were imputed into a power calculation, which was used to design a prospective randomised trial	NA	The annual hospital readmission rate was 50%. A prospective randomised trial would need 182 patients per arm, in order to detect a 15% reduction in readmission with 80% power.

Chapter 18 Discussion of the survey of outcomes research in psychiatry

Survey methods

The survey uses only published examples of outcomes research, and only those examples that are included in widely available databases. It is likely that a proportion of outcomes research is either unpublished or if published, is not included in the databases sampled. Outcomes research is a method that may be used to produce relatively quick and cheap evaluations, often for the purposes of healthcare providers or pharmaceutical companies (Anonymous, 1989; Anonymous, 1992), and this research may not be published. Similarly, if published, it may be in the form of reports and monographs. These forms of publication are not included in databases such as MEDLINE, and are often termed 'grey literature' (NHS Centre for Reviews and Dissemination, 2000).

Failure to include unpublished or difficult to find literature in surveys and systematic reviews introduces a potential bias, in that published literature or that which is in common databases may be unrepresentative of the research as a whole. The problem of publication bias is discussed in more detail in chapter 24. The results of this survey should be considered in the context of this potential bias.

Main findings

Despite the enthusiasm with which outcomes research was adopted and funded in the US, by the 1990s, its value was being called into question. The US Office of Technology Assessment offered a stinging appraisal:

'Contrary to the expectations expressed in the legislation establishing the AHCPR.... administrative databases have generally not proved useful in answering questions about the comparative effectiveness of alternative medical treatments' (Office of Health Technology Assessment - US Congress, 1994)

Clearly, the superficially appealing opportunity to generate large-scale studies from readily available and existing data sources should be approached with caution. The

present survey highlights both the strengths and the limitations of outcomes research as a method for evaluating mental health services.

Strengths of outcomes research

The criticism is often made that randomised trials are undermined by the fact that the participants form a highly selected and homogenous group, and their healthcare and follow up is different from that received by the majority of patients (Anonymous, 1994). The consequence is that it is not always possible to apply the results in clinical practice – that is, trials lack external validity (Naylor, 1995).

One potential advantage of outcomes research is that observational data are routinely collected for all patients and the results can therefore be applied more generally. Further, data are generated in routine healthcare services, rather than in artificially constructed trials. Lastly, outcomes research might be able to deliver answers to some questions relatively quickly and cheaply and with greater statistical power and without the need to seek ethical approval and individual patient consent, compared to the time consuming, and costly, randomised trial.

The present review suggests that outcomes research in psychiatry has indeed realised these advantages - incorporating large numbers of subjects from real life clinical populations and following them up for clinically meaningful periods of time.

Weaknesses of outcomes research

Elwood's original vision of outcomes research required that a rich and clinically meaningful set of outcomes would be collected for all patients during their routine care (Ellwood, 1988). However the feasibility and cost of such data collection has meant that the building blocks of much outcomes research (with notable exceptions) have been data that are collected as part of the administrative process (Iezzoni, 1997). These administrative data (produced by federal health providers, state governments and private insurers) contain the minimum amount of information required to fulfil an administrative function, particularly billing. They generally include little more than routine demographic data, ICD-9 diagnostic codes, details of interventions received during a hospital episode, length of stay and mortality during a hospital episode. The fundamental problem with research using these data is that the outcomes that are available are generally not those that we would like to study.

Research becomes driven by the availability of data rather than by the need to answer specific questions, as acknowledged by one outcomes researcher:

"I utilise data that are available. I do not start with 'what is the problem and what is the outcome?' I say 'given these data, what can I do with them?'" (Blumberg, 1991).

The other major problem with outcomes research, as with all observational research, is the problem of confounding and selection bias (Cook & Campell, 1979; Iezzoni, 1997). The treatment that a patient receives will often be determined by a number of factors that are related to outcome, such as disease severity. Thus patients will differ in many ways other than the treatment they receive, and it is therefore difficult to attribute any differences in outcome to the treatment itself (Green & Byar, 1984).

The present survey suggests that, in psychiatry, large-scale studies using 'humongous databases' are largely achieved at the expense of clinically meaningful outcomes and limited opportunities to adjust for confounding. Only two studies stand out as having collected a broad range of clinically important outcomes and case mix variables, reflecting not just disease severity, but the facets of service use and health-related quality of life – the MOS (Wells, *et al.*, 1989), and Rosenheck's study of PTSD (Rosenheck, *et al.*, 1999b).

Can outcomes research ever be useful in the UK?

Professor Nick Black has recently called for the establishment of large-scale high quality clinical databases across all disciplines in the UK (Black, 1999). The most ambitious example of this work in the UK has been in intensive care (Rowan, 1994). According to Black, such databases need not be seen as an alternative to the randomised trial, but rather a complement. The attractions for researchers include the possibility of generating large samples from multiple participating centres, and including clinically important subgroups of patients, who might be traditionally excluded from trials. Outcomes research can also be used to promote rather than replace randomised trials in a number of ways: First, by raising the level of uncertainty among clinicians as to the effectiveness of established interventions, they might increase clinicians' likelihood of participating in a randomised trial. Second, by providing a permanent infrastructure for mounting multi-centre trials. Finally, the adoption of such databases means that research is no longer the

preserve of a minority of clinicians working in specialist centres; thus enhancing the generalisability of the results.

Suggestions for further research

In the UK, there are research initiatives underway. For example, The Centre for Outcomes Research and Effectiveness (CORE) has been established under the auspices of the British Psychological Society (Clifford, 1998) in order to generate 'practice based evidence' of effectiveness framed within routine services (Marginson, *et al.*, 2000). At this juncture, it would be timely to learn from the examples of outcomes research in the US, and to recognise the limitations and potential of the approach.

Rosenheck, *et al.* (1999b), who provide one of the more rigorous examples of outcomes research, outlines several ingredients of a successful clinical outomces database, capable of producing rigorous and informative research. Outcomes databases should: (1) include large numbers of subjects; (2) use standardised instruments that are appropriate for the clinical condition being treated; (3) measure outcomes in multiple relevant domains; (4) include extensive data in addition to outcomes measures, in order to support matching; (5) collect data at standardised intervals after a sentinel event such as entry to hospital, or discharge from the hospital; (6) take aggressive steps to achieve the highest possible follow up rates. Data should also be collected prospectively if they are to meet these aims

Such databases are going to require substantial time, effort and expense to establish, making outcomes research far from the quick and cheap research option that is envisaged. For example, the whole MOS cost US$12 million, and the depression component cost about US$4 million (Wells, *et al.*, 1996). They are also going to require resolution of the practical and ethical problems of using clinical data for research purposes – as highlighted in recent debates about the data protection act; the European Human rights act and Health and Social Care Bill (Al-Shahi & Warlow, 2000; Anderson, 2001; Kmietowicz, 2001; Medical Research Council, 2000)

The pharmaceutical industry is especially keen to use outcomes research to examine the effectiveness of its products. The current survey highlights that, so far, outcomes studies conducted by the pharmaceutical industry have been of generally

poor quality and do not adhere to the sensible recommendations outlined by Rosenheck, *et al.* (1999b). The use of this method has clear advantages for the pharmaceutical industry – particularly in terms of cost. In conducting such research, the industry can claim that expensive (pragmatic) randomised trials are no longer needed in order to examine clinical and cost effectiveness in routine care settings, nor will they have to provide and dispense the drugs for the many thousands of patients who are included in these studies. Informed consent and ethical approval may no longer be required, since treatment is as received as part of usual care and outcomes are those that are collected anyway. Large-scale outcomes studies that are currently underway – such as the SOHO study – will need to demonstrate that they are methodologically robust and that their results are believable. The current survey provides a framework within which the quality of such studies can be judged.

Mental health researchers must give clear thought as to how outcomes databases should be constructed; how resources might be put in place and to what extent informed consent is required for research conducted using these data. A necessary, but not sufficient condition in the implementation of outcomes research as a distinct method is the collection of a wide variety of outcomes, including patient based outcomes, by psychiatrists in the context of their routine care. The following section considers in detail the practicalities, advantages and potential barriers to this approach.

Section 3 Outcomes measurement in clinical practice

Section 3.1 How do psychiatrists measure outcome?

Section 3.2 Does outcome measurement make a difference?

Section 3 Outcomes measurement in clinical practice

Section 3.1 How do psychiatrists measure outcome?

Section 3.2 Does outcome measurement make a difference?

Section 3.1 Measuring outcome in psychiatric practice – a survey of UK consultant psychiatrists

Chapter 19 Background to the survey

Outcome measurement forms a central component of recent mental health policy formulations. For example, in the UK, there have been a number of initiatives in recent years aimed at the introduction of outcomes measurement tools into routine mental health practice, as part of a government health strategy to 'improve significantly the health and social function of mentally ill people' (Department of Health, 1991).

It was shown in section 1 that outcomes measures broadly serve four purposes: (1) the evaluation of the clinical and cost effectiveness of interventions in experimental situations, such as trials; (2) the monitoring of population health; (3) clinical audit, and; (4) as an aid to clinical decision making in routine practice and patient care (Faden & Leplege, 1992; Fitzpatrick, 1994; Fitzpatrick, *et al.*, 1992a; Ware, 1995).

Despite the availability of various standardised tools with which to measure symptom severity of common psychiatric disorders, and wider quality of life and health status, little is known about the actual use of standardised outcomes measures by clinicians (Slade, *et al.*, 1999). One previous survey of 73 consultant psychiatrists from 1989 established which of a pre-specified range of symptom based clinical measures were in use at that time. This survey suffered from a number of methodological problems, including: small sample size; being restricted to one health region, and failing to examine in detail the actual specific uses of these measure in clinical practice. This survey is also now out of date.

Little is therefore known about the extent to which instruments developed in response to The Health of the Nation Document (Department of Health, 1991), and the National Health and Community Care Act (House of Commons, 1990) have been adopted in practice. This is especially important for measures such as the Health of the Nation Outcome Scale (HoNOS), which were intended to measure outcome, need and inform the provision of healthcare at a population level. For these data to be useful in this respect they must be collected by clinicians routinely, for each and every patient, and for clinicians to do this, such measures must be useful in the care of individual patients.

In order to establish the use of outcomes measures by UK psychiatrists, a survey of the current use of outcomes measures in psychiatric practice in the UK was undertaken.

Aims of the survey

1. To examine the use of outcomes by practising psychiatrists in the day-to-day care of their individual patients.
2. To examine the use of outcomes measures by practising psychiatrists for the purposes of clinical audit.
3. To examine the collection of outcomes measures by hospitals and Trusts, and their use in planning and organising the care of patients.
4. To establish barriers and advantages to the use of outcomes measures by practising psychiatrists.

Chapter 20 Methods of the survey

A questionnaire survey of consultant psychiatrists practising in the UK was conducted. Since there are approximately 4000 general adult psychiatrists practising in the UK (Department of Health, 2000), then a survey of all clinicians was neither practical within the time and resources available, nor an efficient use of resources. A sampling procedure was therefore employed to extract the required information in a rigorous and methodologically efficient manner. The methods employed in the conduct of the survey are outlined below, and follow best practice guidelines outlined in key texts by Moser & Kalton (1971) and Fowler (1993).

The four key stages in the conduct of the survey are as follows and are discussed in detail below:

1. Respondent identification and sampling procedure
2. Questionnaire design and administration.
3. Survey methods
4. Survey analysis

Respondent identification and sampling procedure

Target population.

The target population for the purposes of the survey was defined as practising consultant psychiatrists responsible for the care of working age adult patients in the National Health Service of England Wales, Scotland and Northern Ireland.

Sample frame

The sample frame was drawn from consultant psychiatrists listed in the Medical Directory CD-ROM (Financial Times Healthcare, 2000). This is a commercially available resource, updated and published annually, containing the details of all medical practioners listed in the medical register held by the General Medical Council. In compiling and updating the Medical Directory, all practioners with an entry in the medical register are contacted by post on an annual basis, and invited to provide up to date information, including: their background details (medical school and year of qualification); postgraduate qualifications and membership of Royal

Colleges and societies; area(s) of clinical speciality; current appointments and places of work. In addition, this includes an up to date correspondence address provided by the individual, most usually the place of work.

An initial plan to use a database of addresses supplied by the Royal College of Psychiatrists was abandoned, since this database was unable to provide details of clinical speciality, clinical grade and a hospital based correspondence address. Additionally, it was prohibitively expensive to purchase this resource. The medical directory CD ROM was therefore used in preference, since this contained personal contact addresses, which are likely to be up to date. Previous surveys of consultant psychiatrists have also shown that the Medical Directory is an efficient and reasonably up to date resource for the purposes of conducting postal surveys. Previous surveys of consultant psychiatrists identified from this resource have been shown to achieve response rates of 72% (Adams, *et al.*, 1999) and 80% (Bristow, 1999).

Sampling procedure

Practising psychiatrists are classified within the Medical Directory under the following specialities:

- *'Psychiatry – general adult'*
- *'Psychiatry – old age'*
- *'Psychiatry – child and adolescent'*
- *'Psychiatry – learning disabilities'*
- *'Psychiatry – forensic'*
- *'Psychotherapy'*

When compiling the Medical Directory, these specialities are presented to respondents when completing their individual entry as 'forced choice', non-exclusive categories to indicate their own field of interest. In addition, a free text box is provided which allows stated areas of clinical interest and areas of subspecialty to be added. Of the available categories above, the category 'Psychiatry – general adult' was felt to be the most relevant to the stated target population of the survey.

A computerised search for entries under 'Psychiatry – general adult', excluding all those 'retired', resulted in 3992 individuals. A random sample of 500 adult

psychiatrists was drawn from this pool, using a computer generated random number table (STATA corporation, 1999).

Confirmation that subjects fulfilled the specified inclusion criteria was also sought; by examining whether their stated main speciality corresponded with their 'free text' description of their areas of interest and sub-speciality, and that they included a NHS hospital as their place of work. Those that did not fulfil these criteria were replaced by further random sampling of the Medical Directory database.

All 500 contact addresses were manually entered into a computerised database (Microsoft Corporation, 1998).

Questionnaire construction.

A self-completed/self-report questionnaire was produced. The content of the questionnaire was informed by a comprehensive and systematic literature survey, which had (1) identified the main clinical uses of routine outcome measures and (2) had identified the outcomes measures which are most commonly reported in published psychiatric research (See previous chapters).

Information sought in the questionnaire

The questionnaire sought to identify the following:

1. For commonly encountered psychiatric disorders, which standardised outcomes measures were used by adult psychiatrists for the purpose of:

A. Identifying and assessing the severity of clinical disorders

B. Identifying patients' needs and deficits in social functioning, and quality of life.

C. Monitoring patient progress.

D. Clinical audit.

Common clinical psychiatric disorders were subdivided into the following four broad categories:

- Depression/anxiety and related disorders
- Schizophrenia & other psychoses;
- Drugs and alcohol problems;
- Dementia and related organic disorders.

2. Outcomes measures routinely collected by hospitals/Trusts. (Including administrative outcomes such as length of stay, re-admission rates and standardised measures such as HoNOS scores).

3. Clinicians' reports of outcomes measures being used in the allocation of resources and the planning of psychiatric services.

4. Clinicians personal views on the use of outcomes measures in psychiatric practice

Questionnaire design and administration

The design and response format followed best practice guidelines outlined in key texts by Fowler (1993) and Dilman (1991), and summarised in a recent systematic review by McColl, *et al.* (2001). Brown, *et al.* (1989) provide a useful framework within which to consider the conduct of a survey in order to maximise response rate. A task analysis model of respondent decision making in completing and returning a questionnaire breaks the task down to:

(1) Interest in the task; (2) Evaluation of the task; (3) Initiation and monitoring of the task, and; (4) Completion of the task.

Each of these stages can be facilitated within the design of the survey, by attention to a number of factors. For example, interest in the task is maximised when a personal letter is used, which addresses the questionnaire to the specific respondent, and uses a visually well-designed questionnaire. The use of a personalised covering letter is also useful in this respect, and the content of this letter can maximise interest in a task by highlighting the timeliness and relevance of the question being asked, and identifying the sponsor of the research and outlining the credentials of the investigator. Tangible rewards can also be offered for participation, in order to maximise interest. Table 12 summarises a task analysis of questionnaire design and administration.

Table 12: Task analysis of factors influencing questionnaire design and completion

Stage 1 Interest in the task	Stage 2 Evaluation of the task	Stage 3 Initiation and monitoring of the task	Stage 4 Completion of the task
Personal contact	Time and effort required	Actual difficulty encountered	Provision of SAE
Personalisation of letter	Length of questionnaire	Clarity of question wording	
Personalisation of envelope	Size of pages	Clarity of instructions	
Class of mail	Supply of stamped addressed return envelope	Complexity of questions	
Questionnaire appearance	Cursory evaluation of difficulty	Sensitivity of requests	Reminders to return
Cover illustration	Number of questions	Number and nature of sensitive questions	
Colour of cover	Complexity of questions		
Layout and format			
Quality/clarity of type			
Topic		Actual time required	
Questionnaire title			
Cover illustration			
Content of cover letter			
Timeliness			
Relevance/salience			
Source credibility/trust			
Image of sponsor			
Credentials of individual investigator			
Message in cover letter			
Reward for participation			
Tangible and intangible			
Persistence of source			
Follow up procedures			

The questionnaire design and administration adhered to these principles, and the specific methods used are elaborated below. The questionnaire was designed using a commercially available desktop publishing computer package (Microsoft Corporation, 1997b) – see appendix 3 for final version.

Pre pilot survey

A two stage pre-pilot survey was conducted in order to identify any problems with individual question items, and to test the return rate of the questionnaire. A specified return rate of 60% of questionnaires and a completion rate of 90% for all items was set in advance as an acceptable result, suggesting a representative sample (McColl, *et al.*, 2001).

In order to establish any initial design flaws and to identify any ambiguous terms or items, a *cognitive interviewing technique* was used, as described by Fienberg, *et al.* (1985). Briefly, this technique involves the administration of a postal questionnaire in the presence of a researcher, with the purpose of identifying design flaws and ambiguous items, which might not have been apparent in the initial design. Using this technique, five respondents were individually asked to complete the questionnaire following the instructions given on the forms. They were also asked to raise any problems that they encountered in completing the questionnaire with the interviewer. These were noted and modifications to the wording and layout were effected in response to these comments.

The key modifications made at this stage were:

- The exclusion of a 'does not apply' response category,
- Change in the sequence of questions and the allocation of numbers for each item.
- The provision of a definition and example of an 'outcome measure'.
- The provision of a free text box to give examples of some uses of outcomes measures.

The modified survey was then mailed to 50 individual consultants, together with a covering letter (appendix 4). The accompanying letter followed established guidelines in maximising response rate (Dilman, 1991; Harvey, 1987), in that it:

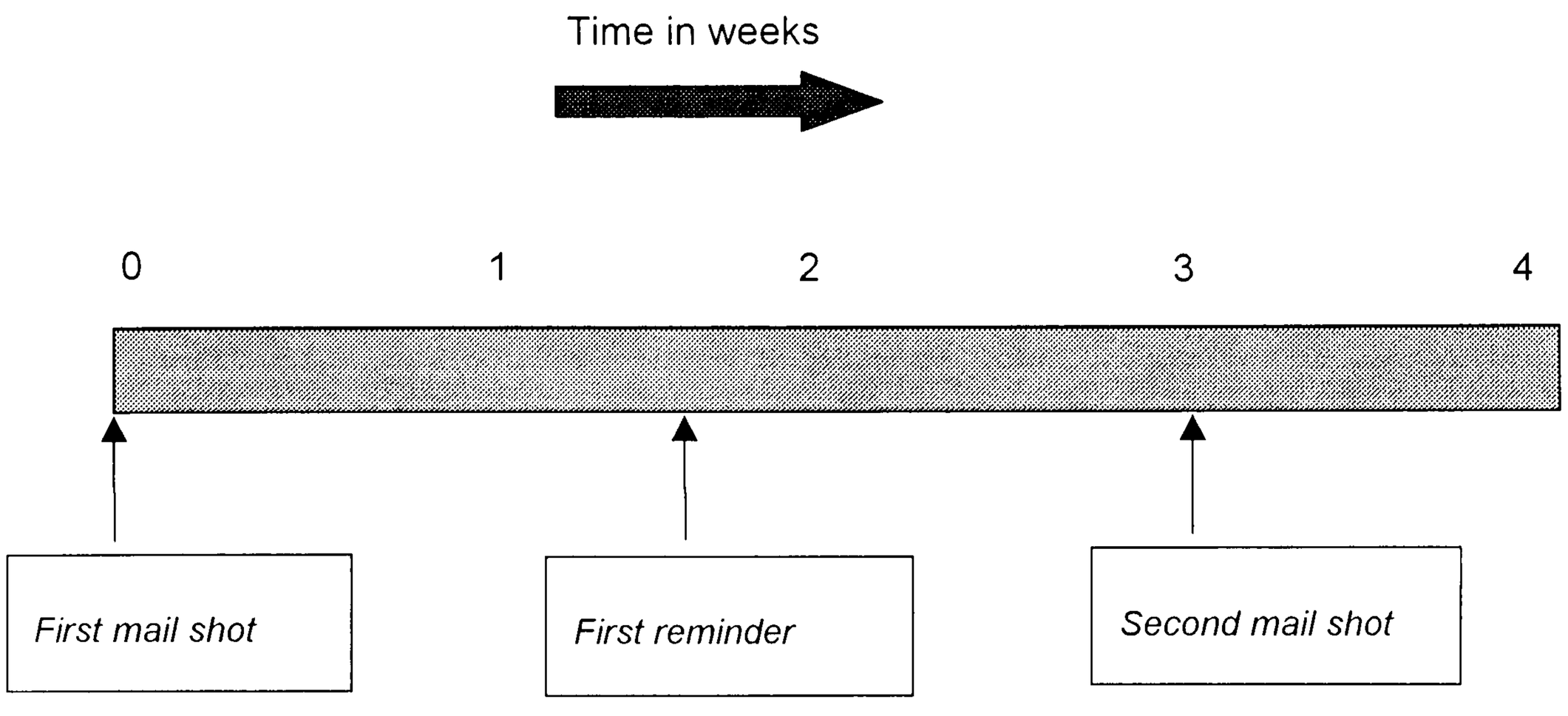
- Briefly set out the purpose of the survey,
- Highlighted the funding source of the research,
- Appealed to respondent's altruism in asking for a reply,
- Identified the key researcher (SG) as someone who was likely to be aware of the issues being surveyed,
- Provided a nominal inducement to participate (entry to a draw to win a copy of the Oxford Textbook of Psychiatry – an up to date textbook, recently published and likely to be coveted by respondents)
- Provided contact details of the researcher (SG) to answer any questions
- Provided a pre-paid envelope for return of the questionnaire.

This approach, together with a single reminder letter at ten days (appendix 4), resulted in a response rate of 56% within three weeks of mailing. A second reminder and further copy of the questionnaire increased this response rate to 66% (appendix 5). All items provided sufficient useable data in over 98% of cases. However, 9% of respondents were no longer engaged in routine clinical NHS practice and had either retired or were only involved in private practice. Retrospective analysis of the Medical Directory indicated that they had been incorrectly identified by the Medical Directory as not being retired, and in active NHS clinical practice. However, by studying year of qualification, the majority of these clinicians were in fact likely to be over 55 years of age, a common retirement age of psychiatrists in the UK (assuming medical qualification at the age of 23 years). The original database of 500 psychiatrists was therefore modified, by eliminating all clinicians likely to be over the age of 55 by virtue of having qualified before 1968. Further psychiatrists fulfilling the inclusion criteria were selected at random to bring the total up to 500 adult psychiatrists. No further modifications were therefore made at this stage, and a further pilot survey was not deemed necessary

Survey

The survey proper was conducted by mailing the questionnaire, covering letter and a pre-paid reply envelope to each of the remaining 500 respondents on the contact database. Reminders and second copies of the questionnaire were sent in accordance with the following time-scale:

Figure 5: Time scale of the survey



All data were entered into a specifically designed Microsoft Access relational database (Microsoft Corporation, 1998), and respondents were sequentially eliminated from the mail address database. Reminders and second mail shots were sent only to non-respondents.

Analysis

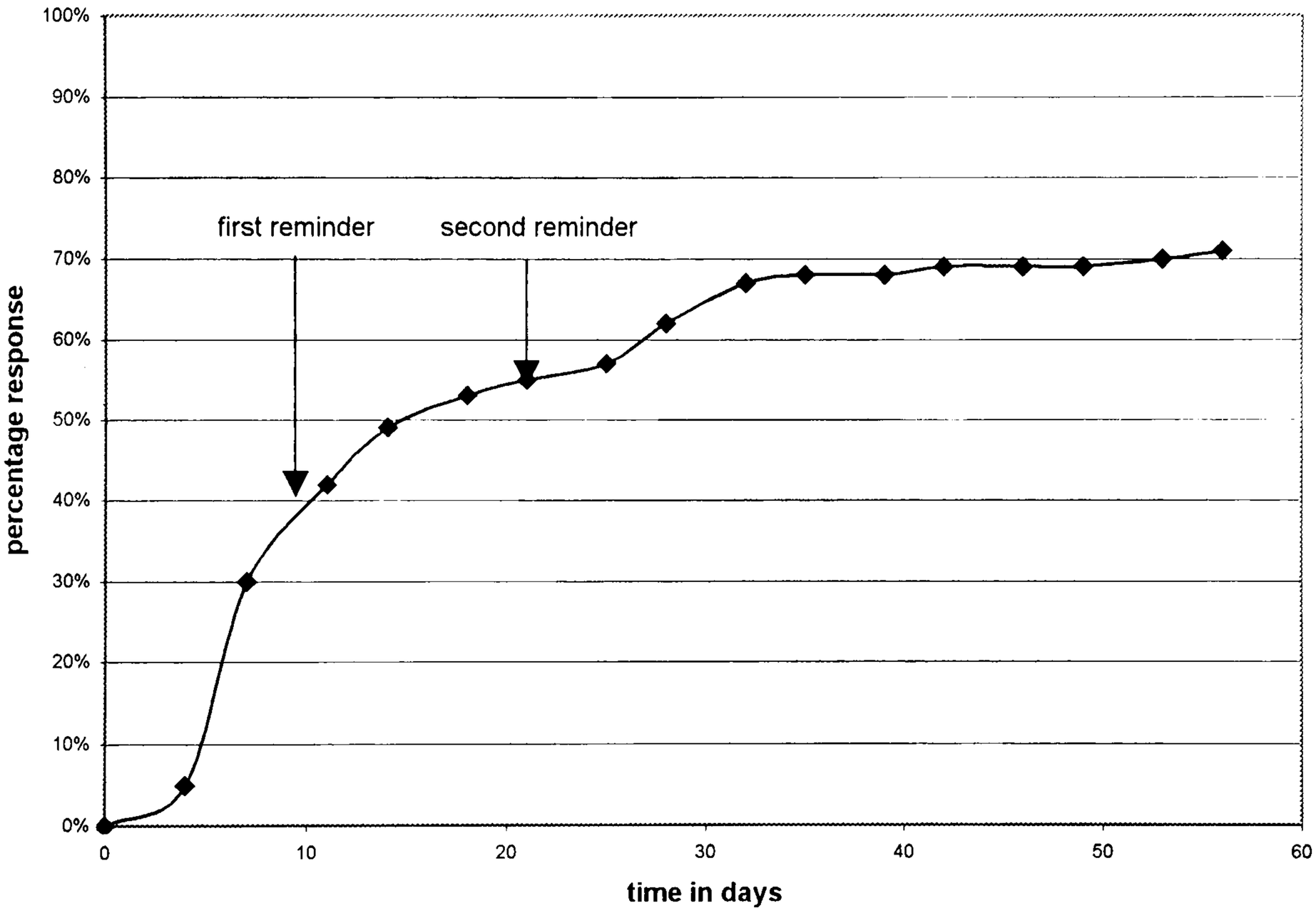
Data were sorted and analysed using Microsoft Access relational database (Microsoft Corporation, 1998). Responses to each individual item and confidence intervals for proportions were calculated using StatsDirect statistical package (Buchan, 2000).

Chapter 21 Results of the survey

Questionnaire responses

In total, 369 (74%) of the 500 questionnaires were returned in a six-week period. Figure 6 demonstrates the time scale over which responses were received. The response rate in the first 10 days was only 42%, and the effect of the first reminder and the mailing of the second questionnaire were key in increasing the response rate to 71%.

Figure 6: Cumulative responses to postal questionnaire over time



Of the 369 returned questionnaires, 29 were either not been completed (n=8), or were competed by consultants who fell outside of our inclusion criteria (n=21).

Reasons for being returned without completion were (total n=8):

- No longer in post/not known at this address (n=4);
- Deceased (n=2),
- Reason not stated (n= 2).

Reasons for non-fulfilment of exclusion criteria were (n=21):

- No longer in NHS practice (n=7);
- Retired (n= 9);
- Child/learning disability/old age psychiatrists (n=5)

Twenty nine of the 369 retuned questionnaires were therefore excluded from the final analysis (final eligible response rate = 340/500 – 68%). When ineligible responses were excluded from the denominator, the final response rate was 72% (340/471).

Details of respondents.

The vast majority of respondents gave their main stated speciality as 'general adult psychiatry' (82%). The breakdown of respondents by speciality is given in table 13. Most respondents reported working in a non-teaching hospital/non teaching community mental health trust (225/340 – 65%), whilst others reported working in a teaching hospital/community trust (117/340 – 35%). Survey respondents reported having been a consultant psychiatrist for a mean 12.4 years (range 2 to 25), and were each responsible for an average of 14 in-patients (range 0 to 42); 17 day-hospital patients (range 0 to 36), and 29 outpatients (range 0 to 44) in any one week.

Table 13: Specialities of respondents

Speciality	Main speciality	Sub specially/special interest	Total
General adult psychiatry	264	14	278/340 82%
Community psychiatry	40	36	76/340 22%
Rehabilitation psychiatry	16	24	40/340 12%
Liaison psychiatry	8	20	28/340 7%
Drugs and alcohol	10	14	24/340 7%
Academic psychiatry	8	10	28/340 7%
Forensic psychiatry	12	10	22/340 6%
Psychotherapy	6	6	12/340 4%

Some respondents indicated more than one speciality, so figures add up to >100%

1. Reported uses of standardised outcomes measures by clinicians in the day-to-day care of patients.

A. Case identification and assessing the severity of specific psychiatric problems

Respondents were asked about the use of outcomes measures in identifying cases and assessing the severity of the following problems: depression/anxiety; schizophrenia/psychosis; cognitive impairment; drugs/alcohol problems. Depression/anxiety and cognitive impairment were the disorders where outcomes measures were most commonly used for this purpose, with 44.6% (95%CI 39.3-50.2%) and 55.3% (95%CI 49.8-60.7%) respectively reporting using these measures, either routinely or occasionally. For disorders such as schizophrenia, and drug and alcohol problems, outcomes measures were reportedly never used for this purpose amongst the majority of consultants (for schizophrenia 72.9%, 95%CI 67.9-77.6%, and drugs/alcohol 83.3%, 95%CI 79.1-87.3%, report never using a standardised measure for this purpose). The most commonly used measures for the detection of depressive and anxiety disorders were the Beck Depression Inventory – BDI (61/340); the Hospital Anxiety and Depression scale – HAD (53/340); and the Hamilton Depression Rating Scale – HDRS (46/340). The most commonly used measure in detecting cognitive impairment was the Mini Mental State Examination – MMSE (Fostein, *et al.*, 1975) (134/340). Although infrequently used, the most commonly reported measures used in the detection of psychotic illnesses were the Positive and Negative Symptom Scale - PANSS (Kay, 1991) (25/340), the Health of the Nation Outcome Scale – HoNOS (Wing, 1994) (25/340), and the Brief Psychiatric Rating Scale - BPRS (Overall & Gorham, 1962) (17/340). For drugs and alcohol problems, the most commonly reported measure was the CAGE questionnaire (Mayfield, *et al.*, 1974) (10/340).

Exact response rates are given in table 14 and figure 7. The most commonly used instruments are given in tables 20-23.

Table 14: Case identification and assessing the severity of specific psychiatric problems

	Never (95% CI)	Occasionally (95% CI)	Routinely (95% CI)
Depression/anxiety	188/340 55.3% (49.8-60.1%)	116/340 34.1% (29.0-39.4%)	36/340 10.5% (7.5%-14.4%)
Schizophrenia/psychosis	248/340 72.9% (67.9-77.6%)	70/340 20.6% (16.4-25.3%)	22/340 6.5% (4.1-9.6%)
Cognitive impairment	152/340 44.7% (39.3-50.2%)	138/340 40.6% (35.3-46.0%)	50/340 14.7% (11.1-18.9%)
Drugs/alcohol	284/340 83.3% (79.1-87.3%)	36/340 10.6% (7.5-14.3%)	20/340 5.9% (3.6%-8.9%)

B. Identifying deficits in social functioning, quality of life or the assessment of patients needs.

Respondents were asked about the use of outcomes measures in detecting deficits in social functioning, quality of life or the assessment of patients needs. Very few clinicians reported using standardised instruments at all for this purpose, amongst any patient groups. The following percentages of clinicians reported never using a questionnaire amongst the following clinical groups: Depression/anxiety 80.6% 95%CI 75.9-84.7%); schizophrenia/psychosis 75.6% (95%CI 70.4-79.8%); cognitive impairment 83.5% (95%CI 79.2-87.3%); drugs/alcohol 88.8% (95%CI 84.9-91.9%). For the small minority who did report using a standardised questionnaire, only a small percentage specified which measure they chose to use. For depression, the most commonly reported measures were the HoNOS (Wing, 1994) (20/340) the Social Adjustment Scale (Weissman & Bothwell, 1976) (9/340); and the Social Functioning Schedule (Remmington & Tyrer, 1979) (5/340). For schizophrenia/psychosis, the most

commonly reported measures were the PANSS (20/340); the BPRS (13/340), and the HoNOS (Wing, 1994) (16/340). For cognitive impairment and drugs and alcohol problems, the most commonly reported measure was the HoNOS (13/340 and 12/340 respectively).

Exact response rates are given in table 15 and figure 8. The most commonly used instruments are given in tables 20-23.

Table 15: Identifying deficits in social functioning, quality of life or the assessment of patients needs.

Question: Do you use standardised outcomes measures to check for deficits in social functioning, quality of life, or to assess patient needs?

	Never (95% CI)	Occasionally (95% CI)	Routinely (95% CI)
Depression/anxiety	274/340 80.6% (75.9-84.7%)	44/340 12.9% (9.6-17.0%)	22/340 6.5% (4.1-9.6%)
Schizophrenia/psychosis	256/340 75.6% (70.4-79.8%)	46/340 13.5% (10.1-17.6%)	38/340 11.2% (8.0-15.0%)
Cognitive impairment	284/340 83.5% (79.2-87.3%)	36/340 10.6% (7.5-14.4%)	20/340 5.9% (3.6-8.9%)
Drugs/alcohol	302/340 88.8% (84.9-91.9%)	20/340 5.9% (3.6-8.9%)	18/340 5.3% (3.2-8.2%)

C. Measuring clinical change over time and therapeutic response

Standardised measures were most commonly used in order to measure change over time amongst those with depression and anxiety problems, with 41.7% (95%CI 36.5-47.2%) of clinicians reporting using a measure at all, although only 11% (95%CI 8.0-15.0%) reported using a measure on a routine basis. A larger proportion of clinicians reported never using a standardised questionnaire for cognitive impairment (66.5%, 95%CI 61.2-71.5%), schizophrenia (73.5%, 95%CI

68.5-78.1%) and drugs and alcohol (91.2%, 95%CI 87.6-94.0%). The most commonly reported questionnaires, in the case of depression/anxiety were the BDI (Beck & Ward, 1961) (49/340); HAD (Zigmond & Snaith, 1983) (41/340); HDRS (Hamilton, 1967) (23/340); and the HoNOS (18/340). The most commonly used measure in the case of schizophrenia/psychosis were the PANSS (Kay, 1991) (20/340); the BPRS (Overall & Gorham, 1962) (13/340), and the HoNOS (Wing, 1994) (16/340). The most commonly used questionnaires in the case of cognitive impairment was the MMSE (Fostein, *et al.*, 1975) 60/340, and the HoNOS (Wing, 1994) (13/340). Of the few clinicians who reported using a standardised questionnaire to measure change over time amongst those with alcohol problems, the most commonly stated measure was the HoNOS (Wing, 1994) (10/340).

Exact response rates are given in table 16 and figure 9. The most commonly used instruments are given in tables 20-23.

Table 16: Measuring clinical change over time and therapeutic response

Question: Do you use outcomes measures to measure *change over time* or therapeutic response?

	Never (95% CI)	Occasionally (95% CI)	Routinely (95% CI)
Depression/anxiety	198/340 58.2% (52.8-63.5%)	104/340 30.5% (25.7-35.8%)	38/340 11.2% (8.0-15.0%)
Schizophrenia/psychosis	250/340 73.5% (68.5-78.1%)	68/340 20.0% (15.9-24.7%)	22/340 6.5% (4.1-9.6%)
Cognitive impairment	226/340 66.5% (61.2-71.5%)	84/340 24.7% (20.2-29.6%)	30/340 8.8% (6.0-12.4%)
Drugs/alcohol	310/340 91.2% (87.6-94.0%)	14/340 4.1% (2.3-6.8%)	16/340 4.7% (2.7-7.5%)

D. *Standardised questionnaires used for audit.*

Overall, standardised questionnaires were used much less for clinical audit, than for the other purposes outlined above. The most commonly reported condition for which they were used was depression/anxiety, where 19.4% (95%CI 15.3-24.0) of clinicians reported their use either occasionally or routinely in the course of clinical audit. The most commonly reported measures for this condition were the BDI (Beck & Ward, 1961) (18/340), the HoNOS (Wing, 1994) (18/340); the HDRS (Hamilton, 1967) (13/340), and the HAD (Zigmond & Snaith, 1983) (12/340). For those with schizophrenia/psychosis, 21.2% (95%CI 16.9%-25.9%) of clinicians reported using a standardised measure occasionally or routinely, and the most commonly reported measures were the HoNOS (Wing, 1994) (24/340); the PANSS (Kay, 1991) (6/340), and the BPRS (Overall & Gorham, 1962) (8/340). Standardised measures were very rarely used for those with cognitive impairment or drugs or alcohol problems.

Table 17: Standardised questionnaires used for audit.

Question: Do you use standardised outcomes measures as tools for clinical audit?

	Never (95% CI)	Occasionally (95% CI)	Routinely (95% CI)
Depression/anxiety	260/340 76.5% (71.6-80.9%)	52/340 15.3% (11.6-19.6%)	14/340 4.1% (2.2-6.8%)
Schizophrenia/psychosis	268/340 78.8% (74.1-83.0%)	40/340 11.8% (8.5-15.7%)	32/340 9.4% (6.5-1.3%)
Cognitive impairment	294/340 86.5% (82.4-89.9%)	36/340 10.6% (7.5-14.4%)	10/340 2.9% (1.4-5.3%)
Drugs/alcohol	310/340 91.2% (87.6-94.0%)	12/340 3.5% (1.8-6.1%)	18/340 5.3% 95%CI 3.2-8.3%

In addition to standardised questionnaires, an enquiry was made into the use of the following routinely collected data in the process of audit: Length of stay; Use of the Mental Health Act; Mortality; Suicide; Readmission rates.

The most commonly used measure was length of stay, with 60.6% (95%CI 95% CI 55.2-65.8%) of clinicians reporting experience of the use of this measure for audit purposes. Other routine data were reported to be used by over half of the clinicians.

Table 18: administrative data used for clinical audit

	Proportion of respondents reporting routine collection
Mortality	108/340 31.7% 95% CI 26.8-37.0%
Suicide	200/340 58.8% 95% CI 53.3-64.1%
Length of Stay	206/340 60.6% 95% CI 55.2-65.8%
Readmission	194/340 57.1% 95% CI 51.6-62.3%
Use of the Mental Health Act	186/340 54.7% 95% CI 49.2-60.1%

2. Outcomes measures collected under the instruction of the Trust/hospital

Very few clinicians reported being required to collect outcomes measures by their trust. Only 13.5% reported being required to collect outcomes data themselves for all of their patients (irrespective of diagnosis). Of those that specified which measure they were required to collect, the HoNOS was the most common (25/340), with some also reporting the requirement to collect the Global Assessment of Functioning (5/340). Exact response rates are given in table 19.

Table 19: Outcomes required by the Trust

Question: Are you asked to routinely collect standardised outcomes measures by your hospital/Trust for the following patients/problems?

	Proportion of respondents reporting routine collection
All patients	46/340 13.5% 95% CI 10.0-17.6%
Depression/anxiety	14/340 4.1% 95% CI 2.3-6.8%
Schizophrenia	18/340 5.3% 95% CI 3.2-8.2%
Drugs and alcohol	10/340 2.9% 95% CI 1.4-5.3%
Cognitive impairment	6/340 1.7% 95% CI 0.6-3.8%

Clinicians were specifically questioned about being asked or required to collect the Health of the Nation Outcome Scale (Wing, 1994), or specific Needs Assessment Tools by their hospital trust. With respect to the HoNOS, 26% (95% CI 21.3-30.1%) reported being asked to collect these data on their patients, whilst only 8.2% (95% CI 5.5-11.7%) reported being asked to use

specific needs assessment tools (such as the Camberwell Assessment of Need and the MRC Needs for Care).

Data collected routinely by hospitals/trusts

Clinicians were asked about data that they knew to be routinely collected by hospitals and trusts. In contrast to standardised questionnaires such as the HoNOS, trusts commonly collected the following data:

- Use of the Mental Health Act (88.2%, 95% CI 84.3-91.5%);
- Length of stay (86.5%, 95% CI 82.7-89.9%);
- Suicides (82.4%, 95% CI 77.9-86.3%);
- Deaths (75.3%, 95% CI 70.3-79.8%);
- Readmission rates (70.6%, 95% CI 65.4-75.4%).

These data were also commonly fed back to individual clinicians, with 72.6% reporting that this happened in their individual trust or hospital.

Table 20: The use of questionnaires for depression & anxiety

	Proportion of consultants using instruments occasionally or routinely	Measures/instruments used
Screening for depression/anxiety	152/340 – 44.6%	BDI – 61/152; HAD – 53/152; HDRS – 46/152; HoNOS – 11/152; MADRS – 10/152; Other (GAF; GHQ; Zung; GDS; SCAN) – 1/152
Screening for deficits in social functioning/QoL/needs	66/340 – 19.3%	HoNOS – 20/66; SASS – 9/66; SFQ – 5/66 GAF – 4/66; CAN – 3/66; QL checklist – 2/66; MRC needs for Care – 1/66
Measuring therapeutic response	143/340 – 42.1%	BDI – 49/143; HAD – 41/143; HDRS – 23/143; HoNOS – 18/143; MADRS – 10/143; GAF/CGI – 9/143; Speilberger – 4/143; BAI – 3/143; Zung – 2/143 GDS – 1/143
Clinical audit	80/340 – 24%	BDI - 18/80; HoNOS - 18/80; HDRS - 13/80; HAD - 12/80; MADRS - 3/80; Speilberger - 2/80; Zung – 1/80

Table 21: The use of questionnaires for schizophrenia/psychosis

	Proportion of consultants using instruments occasionally or routinely	Measures/instruments used
Screening for and diagnosing schizophrenia/psychosis	93/340 – 27.4%	PANSS – 25/93; HoNOS – 20/90; BPRS –17/90; KGV – 9/90; PSE/SCAN 6/90; GAF – 5/90; CAN – 2/90
Screening for: deficits in social functioning/QoL/needs	84/340 – 21.5%	PANSS – 20/84; BPRS – 13/84; HoNOS – 16/84; KGV – 6/84; SFS – 3/84; Lancashire – 2/84; CGI – 1/84
Measuring therapeutic response	91/340 – 26.7%	HoNOS – 33/91; BPRS – 13/91; PANSS – 12/91; GAF/CGI – 9/91; Lancashire – 3/91; SFS – 1/91; CAN – 1/91; SAD/SANS – 2/91
Clinical audit	73/340 – 21.5%	HoNOS – 24/73; PANSS – 6/73; BPRS – 8/73; Lancashire – 4/73; CAN – 2/73

Table 22: The use of questionnaires for cognitive impairment

	Proportion of consultants using instruments occasionally or routinely	Measures/instruments used
Screening for and diagnosing cognitive impairment	188/340 – 55.3%	MMSE - 134/188; WAIS – 9/188; CAMCOG – 3/188
Screening for deficits in social functioning/QoL/needs	56/340 – 16.5%	HoNOS – 13/56; QL checklist – 3/56; CAN – 2/56
Measuring therapeutic response in cognitive impairment	91/340 – 26.7%	MMSE – 60/188; HoNOS – 13/188; WAIS – 6/188; ADASCOG – 1/188; CAMDEX – 1?188
Clinical audit of cognitive impairment	47/340 – 13.8%	MMSE – 13/47; HoNOS – 9/47; WAIS – 2/47

Table 23: The use of questionnaires for Drugs and Alcohol problems

	Proportion of consultants using instruments occasionally or routinely	Measures/instruments used
Screening for and diagnosing drugs and alcohol problems	56/340 – 16.5%	CAGE – 10/56; SADQ – 3/56; HoNOS – 4/56; SCID – 2/56; Maudsley Addictive Profile – 2/56
Screening for: deficits in social functioning/QoL/needs	37/340 – 10.9%	HoNOS – 12/37; SAS – 2/37; MRC – 1/37; GAF – 1/37
Measuring therapeutic response in drugs and alcohol problems	29/340 – 8.5%	HoNOS – 10/29; Maudsley Addictive Profile – 2/29
Clinical audit of drugs and alcohol problems	30/340 – 8.8%	HoNOS – 8/30; ARPQ – 2/30; Maudsley Addictive Profile – 2/30; QLI – 1/30

Figure 7: Use of standardised measures in screening for specific psychiatric problems

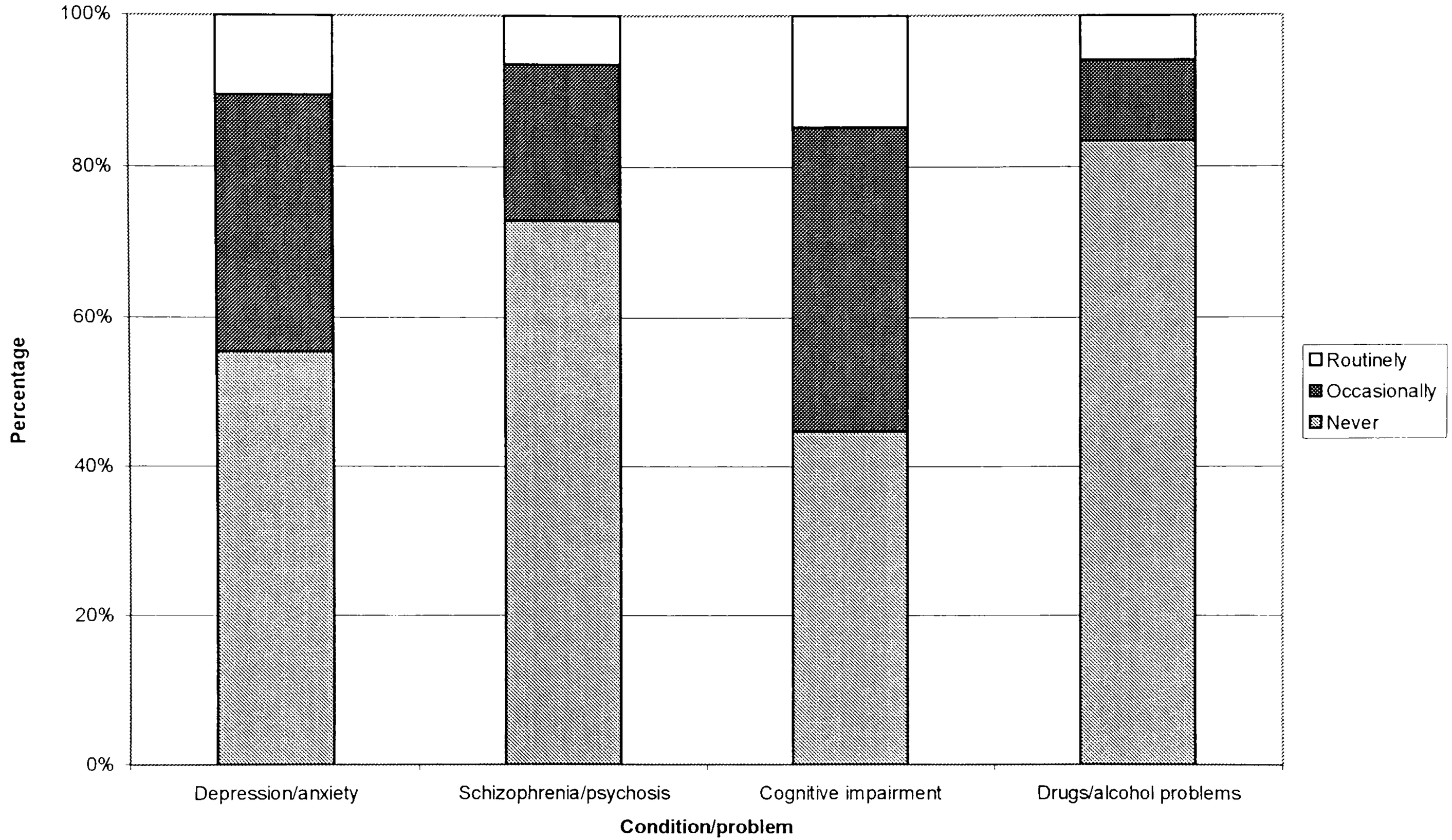


Figure 8: Use of standardised measures to check for deficits in social functioning, quality of life, or to assess patient needs

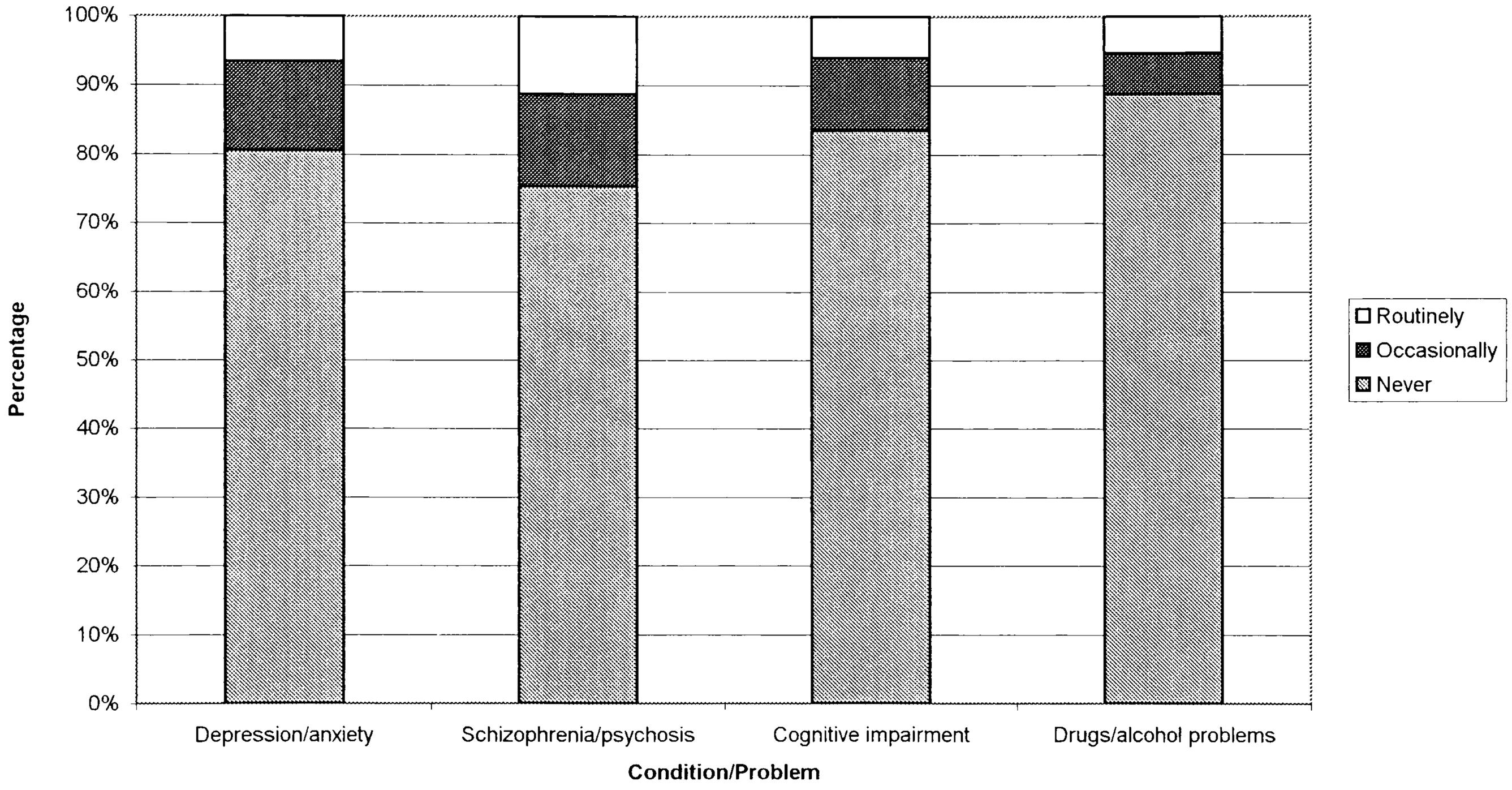


Figure 9: Use of standardised measures to investigate change over time or therapeutic response

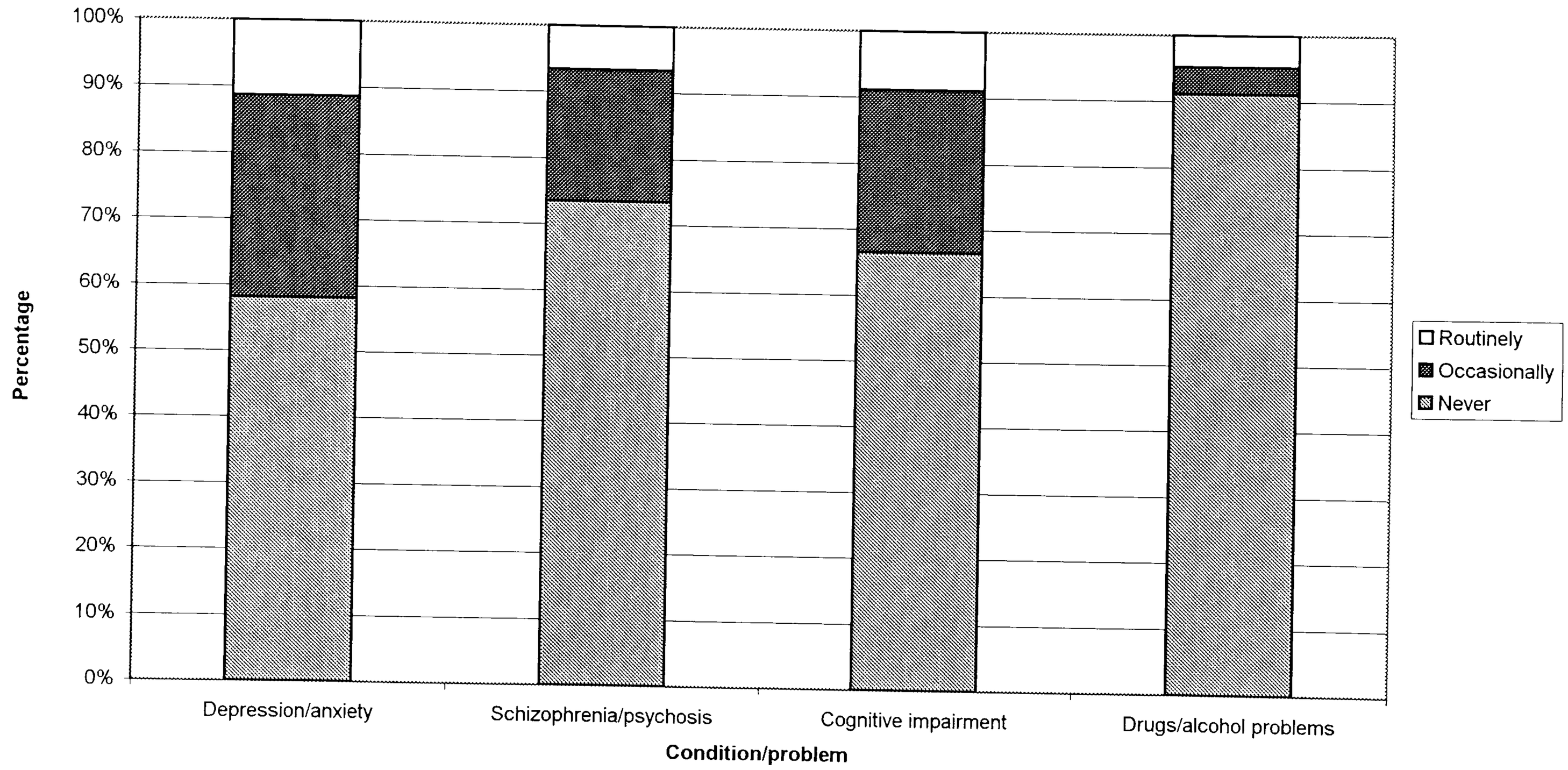
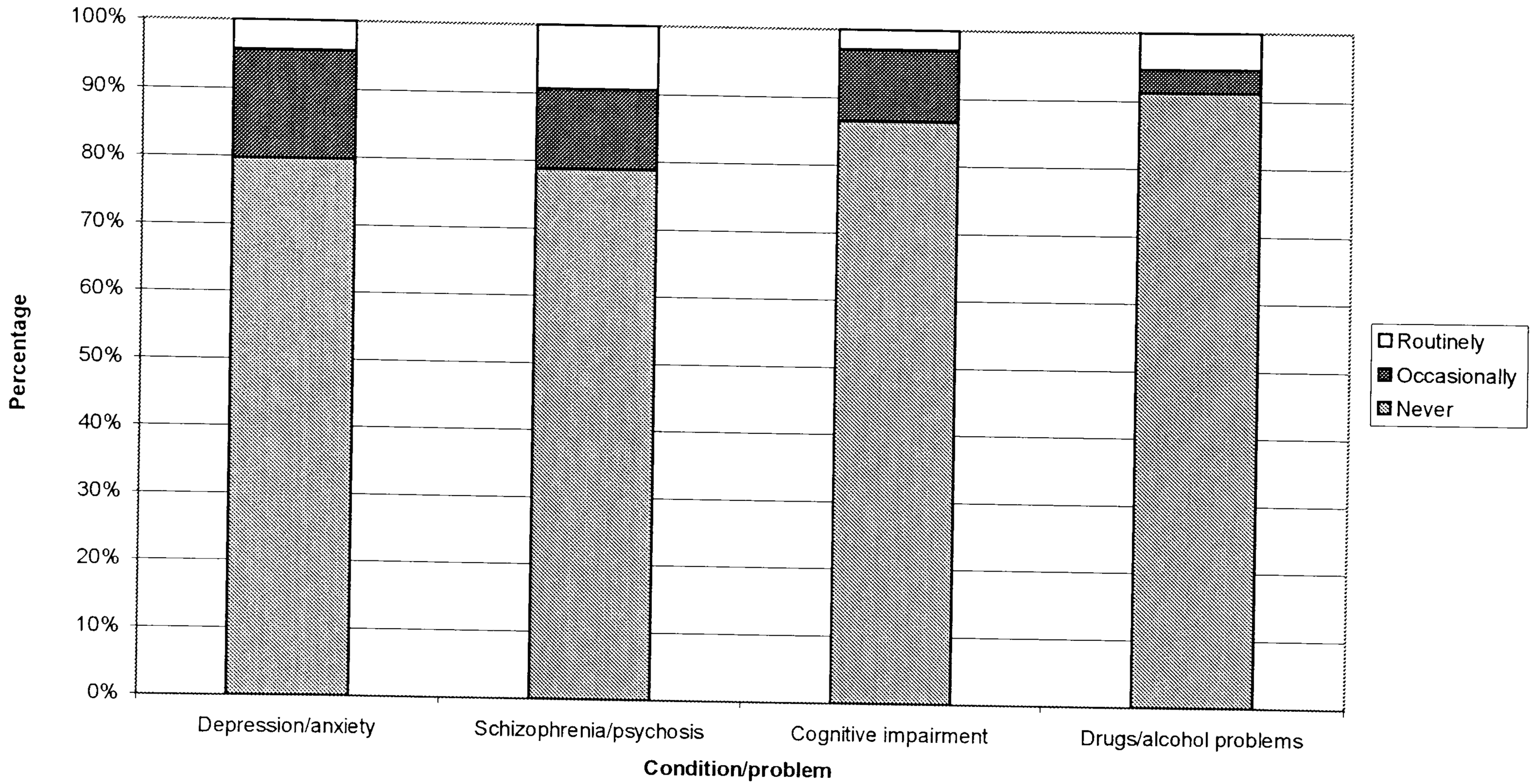


Figure 10: Use of standardised measures in clinical audit



3. The use of outcomes data in planning services and allocating specific funds.

Only a minority of clinicians (107/340 - 31.5%) reported knowledge that outcomes measures had ever been used in planning services or in allocating specific resources within their hospital or trust. Of the 107 respondents reporting the use of outcomes measures in planning services, 53 gave specific examples. Analysis of the content of the comments given shows that broadly four specific uses of outcomes measures are defined, which are outlined below, together with specific examples:

Demonstrating the effectiveness of new treatments and models of service delivery (n=23).

Examples included the monitoring of the use and effectiveness of atypical anti-psychotic medication (n=17), and the use of specific community based treatment models of care – such as assertive outreach (n=7). Specific examples are given below

Specific examples:

'Effectiveness of crisis response team and assertive outreach, admissions and re-admissions before and after.'

'HoNOS scores have provided evidence for the benefit of clozapine, and this has helped get Health Authority funding prescribe it.'

'Demonstration of the effectiveness of CM. Finances appropriately directed.'

One respondent described the use of outcome data to compare performance between different hospital trusts, and to justify funding:

'We use what are largely positive outcomes in comparison to neighbouring trusts to show we are using resources properly and to petition for more (consultants, CPNs, ICU beds, increased drug budgets etc.).'

Defining specific local problems within the clinical catchment area, and responding to these appropriately (n=16).

Specific examples were given of the use of outcomes measures in identifying specific problems in a locality and using this to develop services. Examples included: the use of waiting times for those with alcohol problems to justify a new post; the use of depression and suicide measures amongst hospital populations in the development of liaison services; the use of out of area referral rates to justify home treatment services.

Specific examples:

'Waiting list times led to development of a post to assess people with alcohol problems referred to alcohol treatment unit.'

'Depression and suicide scores amongst general hospital patients used to develop liaison psychiatry.'

'High rate of out of area admissions led to the development of home treatment services. Audit of clinical caseloads allowed for prioritisation of severely mentally ill.'

Rational planning and organisation of services (n=11).

Most frequent examples were the definition of clinical catchment areas (n=4), and the provision of appropriate staffing levels and caseload sizes (n=3). Two respondents described the use of needs assessment tools in order to target resources more specifically at those with severe mental illness. Another described the use of measures in the closure of long stay beds and the re-provision of services.

Specific examples:

'Audit of clinical caseloads allowed for prioritisation of severely mentally ill.'

'Attempts have been made to reform catchment areas using some of the above measures (specified within the questionnaire).'

'HoNOS being used in the development of our 16 years-19 years (adolescence) community and in patient service.'

Negative comments

In addition to the above, three respondents used this space to explicitly state their negative views of the use of outcomes data in planning services. These statements centred on the unreliability of the data.

Specific examples:

'They make them up as they go along from useless data collection, which is unreliable in the first place.'

'The data have been of low quality and unreliable. Collection systems have been poor and it is only now that appropriate systems are being installed.'

4. Clinicians' personal views relating to the use and experience of outcomes measurement

Respondents were asked in a non-directive manner to give their views on the use of outcomes measures in clinical practice. Approximately one third (120/340) used this space to give comments, which centred on the following themes:

- The nature of measurement and outcome in psychiatry (n=40)
- The psychometric properties of the instruments available (n=28)
- The skills, time and resources used in measuring outcome on a routine basis (n=63)
- The utility of measures in clinical practice (n=22)
- The response of organisations to routinely collected outcome measures (n=3)
- Specific comments relating to the Health of the National Outcome Scale - HoNOS (n=26)
- The role of routine outcome measurement within the wider multi-disciplinary team (n=15)

Each of these themes will now be considered in turn

The nature of measurement and outcome in psychiatry (n=40)

40 respondents expressed a negative view of standardised outcomes measures, questioning the ability of outcomes measures to capture the subtlety of multi-faceted outcome and to describe the individual patient.

Specific examples:

'Outcome measure such as those described above are rather simplistic. Most of clientele have severe enduring illness and require a much more sophisticated outcome measure.'

'Often find little advantage over proper clinical assessment. Can become a paper exercise unless specific purpose. We use specific individual care plans with objectives.'

'Deep reservations about the value of any scale which divides a continuous fluctuating process into arbitrary categories which are themselves the subject of entirely personal evaluation.'

'I am appalled at the direction psychiatry has taken, patients are not so much examined and listened to and responded to as human beings. They are categorised by symptoms and evaluated according to their 'scores'. It is a semi-robotic process.'

'A bit time consuming. Not clinically relevant.'

'Have been interested in their use, but never been convinced of their usefulness/reliability. They also seem time consuming, add a pseudo-scientific gloss.'

'Very limited clinical application. Diagnoses are ambiguous'

'Rehabilitation psychiatry is more about maintaining stability and quality of life than on change and getting results. So these measures are less relevant.'

'Never used in routine care. No time, questionable value in the real world.'

'I monitor my patients carefully - using a sort of Gestalt of their well-being or by identifying their needs. Pursuing and trying to address them.'

The psychometric properties of the instruments available (n=28)

Respondents explicitly questioned the basic psychometric properties of validity, reliability and sensitivity to change for available measures (n=28).

Specific examples:

'Outcome scales are time consuming, of questionable validity, very subjective and variable depending on rater.'

'Used exclusively on their own, they are very imprecise and fairly unhelpful when assessing risk and outcomes.'

'Have been interested in their use, but never been convinced of their usefulness/reliability. They also seem time consuming, add a pseudo-scientific gloss.'

'Doubt validity of many outcome scales.'

'Most questionnaire measures were found not sensitive enough to be of use. HoNOS may be an exception.'

'The validity and appropriateness of outcome measures concerns me.'

The skills, time and resources used in measuring outcome on a routine basis (n=63)

Respondents stated that outcome measurement requires training in order that it is done in a valid and reproducible manner (n=25), and that a robust infrastructure, particularly in terms of administration and information technology resources, is needed to support the process (n=20). Respondents generally felt that these were lacking, representing a barrier to their use.

Specific examples:

'Would like to use them, but need more time.'

'Difficult to use in CMHT, no time.'

'Use of measures requires a robust infrastructure and the time required; skilled staff; IT support. Such an infrastructure has not been made available, nor will it ever become available in my working lifetime; sadly!'

'Our service is pressurised so that we have little time at present to use outcome measures'

'A bit time consuming. Not clinically relevant.'

'My concern is: 1) the time involved- haven't 2) I don't know how to use it.'

'If doctors in psychiatry have to use them routinely our workload has to be reduced by 50%.'

'My own strong view is that 'bolt on' forms, risk assessment, CPA will never work, and add to risk because the notes become impossibly bulky and are a) not used b) not read.'

'The use of outcomes measures represents an opportunity cost, and my precious time will be distracted from more useful and productive activities.'

'I had used the HoNOS for inpatients but it took so long that I dropped it.'

'To be meaningful they would have to be part of well thought through collaborative effort-accepted and taken on board sufficiently well resourced and fed back. These conditions do not apply here.'

'would love to use outcome measures, but my adult service has been on the beech at Dunkirk for years.....Unless there are more adult psychiatrists.....!'

The utility of measures in clinical practice (n=22)

Respondents stated that they did not find the results of standardised outcomes measures particularly useful in clinical practice (n=21). One respondent stated that they were more *'research tools'*, rather than instruments that are useful in clinical

practice, and that they *'are more indirect measures than my overall knowledge of the patient'*. Another stated that the *'use of scales detracts from therapeutic relationship.'*

Specific examples:

'Do not find scores and scales useful in treating and monitoring psychiatric patients.'

'Generally unhelpful in clinical practise, of some use in planning service.'

'Rating scales are useful in research to provide objective measure of change but they do not fulfil a useful role in clinical practice. They are more indirect measures than my overall knowledge of the patient.'

'In practice these are rarely used. For formal audit never. For assessment if severity or progress sometimes.'

'My own strong view is that 'bolt on' forms, risk assessment, CPA will never work, and add to risk because the notes become impossibly bulky and are a) not used b) not read.'

Use of scales detracts from therapeutic relationship.

'Seldom makes use of outcome measures, normally relies on clinical judgement.'

'Never used in routine care. No time, questionable value in the real world.'

'Very useful in routine practice. No noticeable impact in service planning or resource allocation.'

The response of organisations to routinely collected outcome measures (n=3)

Three respondents expressed concern that there is no support within trusts for the collection of outcomes measures, or that if there were, then these would not be used in planning services.

Specific examples:

'I'm interested, but never get any feedback or assistance, so enthusiasm has waned.'

'Use of measures requires a robust infrastructure and the time required; skilled staff; IT support. Such an infrastructure has not been made available, nor will it ever become available in my working lifetime; sadly!'

'Managers pay little attention to such unhelpful details, such as clinical data, but blithely follow political dictat.'

Specific comments relating to the Health of the National Outcome Scale (HoNOS) (n=26)

26 responses specifically related to the HoNOS, whereas no other measure was mentioned specifically by name. Comments were largely critical (n=21), and related to: time to complete (n=16); inadequate psychometric properties (n=8); the lack of additional information that it adds to the routine clinical assessment (n=5); the lack of enthusiasm amongst staff (n=7). Positive comments (n=7) included the fact that it could be completed by non-clinicians (n=4), and that it acted as a useful aide memoire in clinical decision making (n=3). One person stated that *'the HoNOS, although scientifically flawed, is useful for bringing together all members of the multi-disciplinary team'.*

Specific examples:

'Attended HoNOS training day at RCPsych, considerable difficulties in implementing into general usage in this Trust. Training offered to all clinical staff but little enthusiasm to use HoNOS in practice.'

'We have used HoNOS with CPA patients in quite a lengthy pilot study but I have not found it particularly helpful. It does not add anything to a clinical assessment, tends to distort the CPA (care programme approach) process.'

'We tried using HoNOS as a routine measure but it wasn't found to be useful for anything. I have often thought that we should use a severity rating scale on each patient admitted (eg. HDRS for depressives, PANSS for schizophrenia) and re-rate prior to discharge. We haven't managed it yet.'

'HoNOS: useful as an aide memoire to patient and their current state-good process measure but very poor as measuring outcomes over time, also aggregate score meaningless doesn't allow comparison of services. Despite this is best available.'

'Insufficient understanding of which scale is best for which condition/situation. The HoNOS is gaining ground, especially as it doesn't require a doctor to complete it, that frees up important time.'

'I had used the HoNOS for inpatients but it took so long that I dropped it. HoNOS is a useless tool, conceived by a government lackey!'

'We were involved in the piloting of HoNOS. It was a disappointing experience. Hours of work were put into collecting data. The Research Unit promised that we would get useful data back and would be able to compare our performance with other trusts. What we got back was very disappointing and of very little clinical relevance. Nothing of great relevance was revealed. All the staff involved felt that the routine collection of data which cannot be readily used is of little benefit.'

'HoNOS although scientifically flawed is useful for bringing together all members of the Multi Disciplinary team.'

The role of routine outcome measurement within the wider multi-disciplinary team (n=15)

15 respondents commented that other members of the multi-disciplinary team, particularly nursing staff, often carry out outcomes measurement. Similarly, others thought that the use of outcome measures fostered greater interdisciplinary communication.

Specific examples:

'For many conditions I rely on rating scales carried out by trained nursing staff, social workers and psychologists.'

'Discuss in the MDT leads to better assessments and analysis of outcomes, as Psychiatry is still a very inexact science. Rating scales only improve things marginally - shared experience in the MDT setting presents better evaluation, though of course rating scales etc may be helpful in giving a fuller picture.'

'Nursing staff do routine assessments, depression/anxiety and risk assessment in IPCU.'

'Nurses collect HoNOS.'

'HoNOS although scientifically flawed is useful for bringing together all members of the MD team.'

'To be meaningful they would have to be part of well thought through collaborative effort-accepted and taken on board sufficiently well resourced and fed-back. These conditions do not apply here.'

'Largely an activity by the nursing staff'

Chapter 22 Discussion of the main results of the survey

Survey methods

The method employed in this study was that of a postal survey, which represents a relatively cheap and efficient way of accessing the thoughts and experiences of a large number of individuals (Fowler, 1993). Postal questionnaires are, for example, cheaper and less time consuming to conduct than face to face or telephone interviews (Oppenheim, 1992). However, postal surveys, as with all forms of research, are prone to a number of potential biases that require further discussion.

Surveys with low response rates are especially prone to biases, since the danger is ever apparent that those who complete surveys differ in systematic ways from those who fail to respond. For example in postal attitude surveys of the general population, response is consistently found to positively correlate with socio-economic status, and to negatively correlate with age, since respondents are shown to younger and better educated than non-respondents (Goyder, 1987). For this reason, recommendations are made that response rates should be in excess of 65-70% (Borg & Gall, 1983; Fowler, 1993; McColl, *et al.*, 2001), in order to dilute this and other potential biases, which can ultimately never be avoided. It is not known to what extent this general observation of respondent bias might be directly applicable to surveys of consultant psychiatrists. However, the observation that those over 55 were found to respond less in the pilot survey may in part been explained by the fact that older clinicians may have been less inclined to respond, in addition to having retired from practice. The consideration must be borne in mind that the results of the present survey may have been biased towards the responses of younger or in some way selected consultants. However the choice to exclude clinicians over the age of 55 can be defended on practical grounds, given the very poor response rate and plausible explanation for their non-response – i.e. that they had retired.

In addition to the systematic biases that exist between those who respond to surveys and those who do not, there are a number of practical reasons why people fail to respond to surveys. These include: moving on from the listed address; being away from the listed address over the duration of the survey; and refusal to participate. Active steps were taken to minimise the effect of the first two biases, by using an up to date data resource which has previously yielded high response rates, and ensuring that reminders spread out over several weeks, together with further

copies of the questionnaire instrument were sent. Further active steps that were planned in the event of a low response rate were to make contact by phone to non-respondents, in order to confirm that a correct correspondence address was in fact obtained, and to encourage completion of the questionnaire. The fact that a response rate was achieved which is comparable to other surveys using this resource (Adams, *et al.*, 1999; Bristow, 1999), at the upper range of respondent rates of clinician surveys in general (Asch, *et al.*, 1997), and in the region of the 65-70% recommended response rate, lends credibility to the findings of the survey.

Factors that are known to affect response rates in general were attended to in the design of the questionnaire and survey methodology. Of particular relevance is the fact that surveys that are perceived as 'boring' or where the topic of research is seen as unimportant to the respondent are known to have lower response rates (Herberlein & Baumgartner, 1978). The fact that a relatively high response rate was obtained may in some way reflect the fact that clinicians perceived this to be an important topic, about which they felt it important to record their actions and their experiences. This, however, also raises the possibility that those responses that were obtained were not typical of practising clinicians. It can be postulated that those with strong positive or negative views regarding the use of outcomes measures chose to respond. Some of the responses to open ended questions were in fact quite strongly negative or positive, suggesting that respondents were more likely to use this space if they had something other than a neutral stance on the topic.

One major limitation of all forms of respondent survey is the fact that what is recorded is a self-report of real events and practices, rather than a direct and independent observation of what actually happens. Additional research would be needed which directly observed clinician behaviour in order to corroborate the findings of this survey.

Of potential concern within the survey is the ability of clinicians to understand what is meant by 'outcome' and an outcome measure. The process of cognitive interviewing identified this at the pre-pilot survey design stage. In particular, the use of the term outcome without an operational definition left some clinicians unclear about what was being asked. The addition of an operational definition was intended to rectify this. However, it was clear from some responses that a small minority of clinicians confused *outcomes* with *process* measures (e.g. waiting list times),

particularly when asked to give free text examples. This ambiguity of terminology is not unique to psychiatrists, and ambiguity exists in the research literature (see chapter 5). The results of the survey must be considered alongside the general difficulty that exists in assuming that the questions and terminology were understood. The use of cognitive interviewing at the design stage follows best practice guidelines in ensuring this bias was kept to a minimum.

The use of forced choice questionnaires is particularly good for eliciting factual information, but is less good for enquiring into the nature of topics about which little is previously known (Mays & Pope, 1996). For this reason, a mixture of forced choice factual questions and open-ended questions were employed, with the latter being used to elicit respondents' experience of outcomes measurement. That many chose not to elaborate on this topic does not devalue the responses that were obtained. However, these data are limited in many respects. Most importantly, hitherto unexplored issues were raised within these 'free text' responses, which would justify further research or investigation – for example some of the barriers that were identified to the routine use of outcomes measures. Respondents generally tend to be less expansive in writing than in speaking. Therefore, there is clearly a role for further interview based research, where respondents might be encouraged to elaborate on some of the topics raised, allowing more spontaneous and richer information to be obtained and recorded verbatim. The research needed to explore some of the issues identified within this quantitative survey is of a qualitative nature. Techniques such as 'in depth interviews' would usefully allow some of the important issues and hypotheses to be explored in greater detail (Denzin & Lincoln, 1994).

Main findings

Use of outcomes measures by adult psychiatrists in the day-to-day care of their patients

The main finding is that the majority of clinicians do not use outcomes measures at all in their day-to-day practice. The only exception to this is in screening for cognitive impairment; although only a minority of clinicians do this routinely and this condition represents only a small component of the case mix in general adult psychiatry. What is particularly surprising is the infrequency with which patient needs and psychosocial problems are measured in any standardised way, despite

political pressures and explicit government policy (Glover, *et al.*, 1997; Secretary of State for Health, 1999) to adopt measures such as the HoNOS and needs assessment tools. This may reflect a failure simply to use standardised measures, or perhaps a wider indifference towards and failure to address psychosocial outcomes and needs.

HoNOS does seem to have found a place in measuring outcome in UK mental health services, albeit a small one. It is only used by a small minority of clinicians, but seems to be the main tool that is used in measuring psychosocial outcome for those schizophrenia and other psychoses.

Outcomes measures routinely collected by hospitals/trusts

When data are collected by trusts, they are administrative outcomes – such as length of stay and readmission rates. These are generally the measures that are the easiest to collect, but which potentially bear little relation to the clinical or psychosocial outcome of the individual patient or clinical population. Interestingly, it is these data that are routinely fed back to clinicians, and are used in clinical audit, rather than standardised patient based measures. This is perhaps not surprising, since it is administrative outcomes that will form the basis upon which success of individual trusts or clinicians is to be judged in the performance management framework of the 'New NHS' (Secretary of State for Health, 1999). The desirability of these 'performance indicators' as the main measurement of success or failure is debatable (Davies & Crombie, 1997; Davies & Lampel, 1998). Of particular concern that these figures are the easiest to manipulate or 'improve', without conferring any overall health gain on the population or service under consideration (Smith, 1996b). Organisations (both medical and non-medical) are known to concentrate on the manipulation and improvement of single outcomes indicators, at the expense of all others, when they are elevated to the status of 'performance indicators'. This distortion of the behaviour of organisations has been termed 'gaming' (Davies & Lampel, 1998; Smith, 1996a). There is a very real danger that the elevation of easy to collect data, rather than clinically meaningful data, to the position of a performance indicator will adversely affect the outcome of patients, or will at best, confer little advantage.

Use of outcomes measures in the allocation of resources and the planning of psychiatric services.

Relatively few examples were found of measures of patient based outcome or need being used in planning services. Several of the examples that were offered by consultants related to the use of outcomes measures to demonstrate the worth of new and expensive technologies in psychiatry, such as new drugs for the treatment for schizophrenia. The use of outcomes measures collected in the context of routine practice, rather than experimental research settings, raises a number of issues.

First, the collection of routine data in order to assess the effectiveness of interventions has several drawbacks. These include the fact that effectiveness needs to be evaluated using robust methodological research, ideally using comparison or control groups, and with due consideration of confounding and extraneous variables that could offer plausible explanations for a demonstrated effect or lack of effect (Sackett, *et al.*, 1991). Examples of published versions of the use of routine outcomes measures to demonstrate the clinical and cost effectiveness of new drugs for schizophrenia in routine care settings suggest that flawed methods are used. For example, the clinical and cost effectiveness of new drugs such as clozapine (an example cited by one of the respondents) has been judged in local settings using underpowered, and uncontrolled before and after studies (eg Aitchison & Kerwin, 1997), with little consideration of basic epidemiological principles when judging the results of such studies (eg Adams, 1997). Better-resourced attempts have also been made to use routinely collected data in order to judge the effectiveness of new technologies in general (Sheldon, 1994), and in psychiatry in particular (see earlier chapters). These have largely been unsuccessful, and have failed to make a convincing case for the use of this method being used appropriately in practice.

Second, the successful application of routinely administered outcomes measures to evaluate the effectiveness of interventions or policy initiatives presupposes that the instruments are fit for this purpose. Instruments must be valid, reliable, and most importantly sensitive to change (Guyatt, *et al.*, 1993a; Streiner & Norman, 1995). Unfortunately most respondents failed to mention the specific instrument that was used for the purposes that were outlined. The suitability of the instruments used cannot be therefore commented on in most cases. However, several respondents mentioned that the HoNOS was used to measure the effectiveness of interventions, including the effectiveness of new atypical drugs. The basic psychometric

properties of the HoNOS have been questioned (Stein, 1999), and this instrument has specifically failed to show sensitivity to change in the underlying condition (Bebbington, *et al.*, 1999), making it of limited use as a measure of outcome and responsiveness in individual patients. It is possible that these limitations are not well appreciated by clinicians when using instruments to infer clinical change and assume that the effectiveness of an intervention has been demonstrated in a local setting. Conversely, it was also apparent from a number of respondents that some clinicians are all too aware of the limited psychometric properties of the available instruments. The HoNOS was specifically named by respondents when voicing negative comments about the potential for outcomes measures to be used in routine practice.

Clinicians personal views on the use of outcomes measures in psychiatric practice

The largely negative views regarding the use of outcomes measures in psychiatric practice are important in several respects. These views give an insight into the reasons behind the general reluctance to use outcomes measures that has been demonstrated in the survey.

The concern regarding the basic psychometric properties of available measures and the time taken to complete them represent real barriers to their use. Slade, *et al.* (1999) have speculated that outcomes measures will never be used on a routine basis unless instruments are available that are psychometrically robust, brief, quick and easy to administer. The research presented here provides empirical evidence to support this assertion.

The greatest number of respondents articulated the widely held view that outcomes measurement is an activity that consumes resources – particularly time. It is clear that clinicians either do not view this as a productive use of resources, or believe that sufficient resources have not been provided in order to make routine outcomes measurement a reality.

Routine outcomes measurement represents a 'technology' (Ellwood, 1988), and as such, its implementation should be justified on the grounds of demonstrated clinical and cost effectiveness. The evidential basis for the clinical and cost effectiveness of routine outcomes assessment for those with mental disorders is discussed in greater depth in the following chapter. However, respondents themselves directly

questioned whether this was in fact a clinically and cost effective approach. Even those writers that have lent support to the idea that routine outcomes measurement (e.g. Marks, 1998; Slade, *et al.*, 1999) have claimed that this will only be achieved if sufficient resources are provided to make this a reality. Importantly, one aspect of these resources might be adequate information technology to record, store and allow easy retrieval and feedback of outcomes to clinicians. Several respondents directly commented that outcomes measurement had been imposed as a top down initiative, with no other resources provided to support this. Specifically, it was clear that in many cases outcomes measurement was expected to be undertaken in addition to clinicians existing workload, and that information technology was not adequately provided or resourced. This generated a certain amount of resistance amongst clinicians to the implementation of this strategy. Clinicians also expressed the concern that outcomes measurement within trusts was a largely bureaucratic exercise, with little feedback of centrally collated outcomes, and little perception that they had been actually used in changing or organising services for the better.

The general reluctance amongst clinicians to measure outcome in a standardised way may also be explained by the reservations that were expressed about the ability of such measures to adequately capture the subtlety and complexity of the individual patients' health and well-being. This is an issue that goes beyond the traditional psychometric concerns of validity and reliability, and extends into the realms of the very nature of measurement and a belief that complex experiences can not be easily operationally defined and condensed to a series scores on a scale. Of significance was the expressed view that outcomes measures add little to the normal processes of patient assessment, such as history taking and multi-disciplinary assessment. The use of terms such as 'dehumanising' represents an extreme expression of this belief. Clearly, if clinicians believe that standardised outcomes measurement adds nothing to their traditional way of working, then they are unlikely to use them, or if they do use them with some reluctance, then they are unlikely to incorporate their results into clinical decision making.

The following chapter will address the question as to whether there is in fact any demonstrable benefit in terms of improving the care of the patient in more depth. However, on the basis of the findings of the present survey, that a significant barrier to the use of standardised outcomes measures in routine practice is the fact that clinicians do not perceive their use to be of any direct benefit to themselves or the care of the individual patient.

Specific comment is justified for the responses offered regarding the HoNOS. This measure has above all come to symbolise the shift towards outcomes measurement within British psychiatric practice, since it was conceived in response to early policy documents that held the measurement of outcome as central to quality improvement (Department of Health, 1991). It was also developed by psychiatrists' own professional organisation – the Royal College of Psychiatrists (Wing, 1994). It forms a central component of the most recent major policy document in mental health (Department of Health, 1998), which stipulates that a minimum data set (Glover, *et al.*, 1997) be collected for all those with severe mental illness in the course of care planning.

The HoNOS does seem to have found a definite place in the measurement of outcome in UK psychiatric practice, since it is the main method by which outcome is measured for mental disorders such as schizophrenia, albeit by only a small minority of clinicians. Aspirations that it would initially be collected on a service wide basis, so that it could be used in both individual patient care, and in assessing the needs and adequacy of service provision at a population level (Curtis & Beevor, 1995), have clearly not been realised. The general barriers to the routine use of outcomes measures, outlined above, apply to this measure. More specifically however, this was the only measure mentioned by name, when respondents were asked to give their personal views regarding the use of outcome measurement in routine practice. Clinicians who offered their views felt it to be psychometrically unsound, cumbersome and over long, thus not fulfilling the criterion of usability set down by Slade, *et al.* (1999). Paradoxically, the instrument was said by a small minority of respondents to be a useful adjunct to history taking, and a useful focus of discussion within multi-disciplinary team meetings. The enduring benefit of this measure might therefore be as an adjunct to improve the process by which care is given – by improving professional communication, rather than as a measure of outcome, where it is widely held to be a flawed instrument. The future role of the HoNOS was recently summarised by Stein (1999), who said:

'Eventually, the HoNOS will find its place within the research armamentarium, but whether it will improve the health of this nation, or any other nation, remains open to question'

Sharma, *et al.* (1999) provide a useful insight into the real value of this instrument when adopted on a service wide basis as a routine outcomes instrument. Sharma, *et al.* (1999) routinely administered the HoNOS questionnaire to 204 consecutive patients in an inner city psychiatric service, and showed that scores changed in the anticipated direction over time. However, the most interesting observation of the authors is the statement that:

'HoNOS ratings were rarely used in the care meetings in our team.....We found that [patient] review meetings were the place for rating HoNOS, rather than for using the HoNOS ratings to formulate a care plan. Even if the HoNOS ratings were made available in review meetings, their value in care planning would have been limited.'

Upon completion of the project, the instrument fell from use. In providing some explanation for this, Sharma et al commented that:

'The use of any standardised schedule in routine clinical practice will require adequate administrative support, as well as the motivation of health professionals. National Health Service trusts should take account of both of these factors, before introducing this or any other instrument into routine work.'

Section 3.2 Does routine outcome measurement improve outcome in mental illness? A systematic review

Chapter 23 Background to the review

Routine outcome measurement

The previous section (3.1) highlighted the fact that outcomes measures are rarely used on a routine basis in the care of those with psychiatric illnesses. Several barriers to the use of these measures were identified, and the question of whether this was a clinically and cost effective intervention was raised. The purpose of this section of the thesis is to examine the evidence base to support the policy of the routine use of outcomes measures in improving the quality of care for those with psychiatric illness. As a preliminary, the theoretical basis of the potential for routinely administered measures to improve the care of those with psychiatric disorders will be examined.

The benefits of routine outcome measurement

When used as aids to decision making in routine care, outcome measures are thought to be useful in improving patient care in a number of ways. First, by identifying problems which might not otherwise be recognised by clinicians or those responsible for care. For example, clinicians are often unaware of patients' social and psychological problems (Sprangers & Aaranson, 1992), and the identification of these problems might trigger an appropriate response and improve the overall quality of patient care. Second, outcome measures might be used to monitor the course of patients' progress over time, to make decisions about treatment and to assess subsequent therapeutic impact. Third, surveys have suggested that clinicians find these data useful in formulating a more comprehensive assessment of the patient (Kazis, *et al.*, 1990; Young & Chamberlain, 1987). Lastly, patients often welcome the opportunity of giving clinicians information regarding their health status, particularly when they perceive this information is not otherwise comprehensively assessed, thus aiding effective patient-doctor communication (Nelson, *et al.*, 1990).

There are broadly two areas in which routine outcome assessment might be applied in the improvement of the quality of care for those with psychiatric disorders, and which will be examined in this review. The first is in the recognition and management of psychiatric disorders, such as anxiety and depression, in non-psychiatric settings (such as primary care and the general hospital). The second is

in the management of already recognised psychiatric disorders in specialist care settings.

Disorders such as anxiety and depression are especially prevalent in both primary care and general hospital settings. Evidence for this comes from a number of sources and the most robust evidence involves the use of research interviews - designed to allow diagnoses in a reproducible and standardised manner against accepted diagnostic criteria. For example, the work of Goldberg and colleagues (Goldberg, *et al.*, 1970; Goldberg & Huxley, 1980) has shown that attenders at general practices show a prevalence of depression and anxiety several times greater than that in the general population, and that this often goes unrecognised. Similarly, Feldman, *et al.* (1987) have studied the prevalence of psychiatric disorders in general hospital inpatients and found it to be 15-20% (2-3 times the general population incidence). Only half of those 'cases' were detected by clinicians. Research by others has shown higher than expected rates of psychiatric disorder in general hospital outpatient attenders (van Hemert, *et al.*, 1993).

Less robust evidence comes from the use of psychiatric screening ('case finding'), questionnaires administered in these settings, which consistently show an elevated prevalence of psychiatric disorder - compared to that observed with standardised interviews (Meakin, 1992). Examples of such questionnaires include the General Health Questionnaire (Goldberg, 1972), and the Beck Depression Inventory (Beck & Ward, 1961) and the Hospital Anxiety and Depression scale (Zigmond & Snaith, 1983). According to the author of one of these instruments, high scores on these screening questionnaires should, therefore, lead to closer investigation to confirm or eliminate the presence of minor psychiatric illness - which might warrant further intervention (Goldberg, 1986). However, the use of such measures in non-psychiatric settings to identify problems and to monitor progress would be consistent with their use as an outcome measure.

Similarly, recently introduced measures of health status and health related quality of life, such as the short form 36 (SF36) contain items and sub-scales which measure 'psychological well-being' (Ware & Sherbourne, 1992). In the case of the SF36, the mental health sub-scale was validated by its correlation with already established measures of depression and in its ability to discriminate between those with and without clinically diagnosed depression (McHorney & Ware, 1995; McHorney, *et al.*, 1993). Psychological well being is in fact a core component of many 'health status

measures' (Bowling, 1997; Ware, 1995) and is clearly related to the domains which are measured by instruments such as the GHQ, HAD and BDI. Measures of health status and health related quality of life have been advocated as being suitable for routine use in clinical care settings (Greenfield & Nelson, 1992). Where such measures are used to explicitly identify minor psychological problems (such as depression and anxiety symptoms) and to monitor changes over time, then this is consistent with their use as a routinely administered outcome measure - and the suitability of their use in this context will be considered within this review.

Routine outcome measurement has also been advocated as an adjunct to patient care within psychiatric services (Marks, 1998), where measures of psychiatric symptoms might be applied in order to measure therapeutic response and to inform management decisions. Similarly, associated health status and health related quality of life amongst those with commonly encountered psychiatric disorders such as depression and schizophrenia has been shown to be poor, and at least as bad as that seen in chronic medical conditions such as rheumatoid arthritis and ischaemic heart disease (Orley, *et al.*, 1998; Wells, *et al.*, 1989). In the case of schizophrenia, impairments in quality of life and health status are often unrelated to the number or severity of symptoms, such as delusions and hallucinatory experiences (Anthony & Rogers, 1995; Becker, *et al.*, 1993). This is especially important, since it is the level of symptoms that forms the major focus of clinical consultations and practice, and is the major criterion by which the success (or otherwise) of treatment is judged in both practice and research (Revicki & Murray, 1994). Consequently, clinicians' perceptions of these problems are often poor; clinicians underestimate the health status or health related quality of life of patients when patient and clinician ratings are compared (Becker, *et al.*, 1993; Lehman, 1983a; Lehman, 1983b; Sainfort, *et al.*, 1996). The use of more comprehensive outcome measures, which capture both symptoms and wider health related quality of life, might therefore, be useful in identifying needs, monitoring clinical response and making clinical decisions in those with severe mental illness. Further, it might be supposed that the use of patient based measures in addition to symptom-based measures might provide a more comprehensive assessment of patient outcome, since they potentially move the clinical consultation beyond the isolated consideration of the severity of clinical symptoms such as delusions and hallucinations.

In consideration of these possible benefits, in the UK, there have been a number of initiatives in recent years aimed at the introduction of outcomes measurement tools

into routine mental health practice, as part of a government health strategy to *'improve significantly the health and social function of mentally ill people'* (Department of Health, 1991). For example, the Health of the Nation Outcome Scale (HoNOS) has been developed with a number of uses in mind, including the assessment of local service requirements and psychiatric morbidity at a population level (Stein, 1999). However, a key aim of the developers of the HoNOS is that it should be useful to clinicians in actual individual care planning, since without this feature it would not be widely used and so the data would not be collected which would ultimately inform decisions at a population level (Wing, 1994). In a related vein, there has also been substantial research activity into the development of instruments aimed at assessing the needs of those with severe mental illness. Such *needs assessment* tools are intended to define health and social needs at both a population level and, ideally, at an individual level (Thornicroft, *et al.*, 1992), such that healthcare provision might be more rational, responsive and 'appropriate' (Stevens & Gillam, 1998; Wright, *et al.*, 1998). Examples of individual patient needs assessment tools for use in severe mental illness include the Camberwell Assessment of Need (CAN) (Phelan, *et al.*, 1995); MRC Needs for Care Assessment (Brewin & Wing, 1993).

Possible disadvantages of routine outcome measurement

The routine measurement of outcome has not been without its critics (Crombie & Davies, 1997; Davies & Crombie, 1997), and concerns have been raised that 'outcomes measures' are un-interpretable, unwieldy and a bureaucratic hindrance to successful patient care.

One way in which the success or usefulness of these measures in everyday routine care might be judged is by evaluation of the degree to which their adoption improves the outcome and quality of care. Research in other specialities has generally not been positive in this respect. For example one important study by Kazis, *et al.* (1990) examines the benefits of informing clinicians of their patients' health status scores. Patients included in this study all had a diagnosis of rheumatoid arthritis and were attending routine outpatient follow up. The health related quality of life instrument examined was the patient completed disease specific Arthritis Impact Measurement Scale (AIMS) (Meenan, 1982) or modified Health Assessment Questionnaire (MHAQ) (Pincus, *et al.*, 1983). Patients in the experimental group completed health status instruments that were then sent to clinicians on a quarterly

basis over a year. An 'attention placebo group' completed instruments quarterly, but these data were not fed back to their physician. A 'control group' only completed instruments at the beginning and end of the study. There were no detectable differences between groups at the end of the year in terms of outcomes such as patient satisfaction or changes in health related quality of life (as measured by the AIMS and MHAQ). Nor were there any differences in terms process variables, such as changes in medication or referrals to other agencies.

There are various reasons to be cautious about the likelihood that the routine use of outcome measures would improve outcome and quality of care, which might explain the inability to establish any benefit in the above experimental example. Firstly, many clinicians find the information conveyed by outcome measures and health status measures irrelevant to clinical decisions, time consuming, difficult to interpret and too cumbersome to be integrated into routine practice (Deyo & Carter, 1992; Deyo & Patrick, 1989; Greenfield & Nelson, 1992). Additionally, the measures may not be sufficiently psychometrically robust to inform individual patient care. The most important facet of validity is *sensitivity to change*, if they are to be informative as outcome measures (Guyatt, *et al.*, 1993a). If they are not sensitive to change, then their results will not be interpretable and important changes will not be detected or acted upon (Fitzpatrick, *et al.*, 1992b). Reliability is often demonstrated at a 'group' level (using correlational statistical analysis), but high indices of 'group level' reliability can obscure large 'between-individual' and 'within individual' variation scores which make instruments uninformative at an individual patient level (Dunn, 1996; Streiner & Norman, 1995).

The measurement of outcome in the context of individual patient care is not without cost. Instruments must be developed, administered (often by clinicians), coded, stored and retrieved - all of which have resource implications. Similarly, there is a danger that outcome measurement triggers resource intensive interventions which are of no proven benefit to patients, and which might actually harm them. Perhaps, more subtly, there is also a danger that the uptake of outcome measurement in this context represents a marketing ploy, in which measurement is used to demonstrate an institution's 'customer orientation', but which does not inform the provision of care (Fitzpatrick, 1994).

In summary, the case for the benefit of routine outcomes measurement is far from clear.

Aims of the review

To review systematically the best available evidence of the value of routine outcome and needs assessment in the day to day care of those with common mental disorders such as anxiety, depression and schizophrenia and related disorders.

Chapter 24 Systematic reviews and their application in mental health

The method chosen to answer the specific question of whether there is any evidence to support the clinical and cost effectiveness of routine outcomes measurement for common mental disorders is that of *systematic review*. The following section briefly outlines the rationale behind systematic reviews, and the stages involved in the conduct of a review. Particular reference is given to the application of a systematic review methodology to questions pertaining to the delivery, organisation and quality of health services, and the application of systematic reviews in the sphere of mental health services research.

The origins of systematic reviews.

In the field of mental health, as in all healthcare, evidence of the effectiveness of interventions is often contradictory, partly because of differences between the studies in terms of methodological rigour, patient populations and interventions (Gilbody & Petticrew, 1999). In order to make sense of this disparate and often contradictory literature, practitioners, policy makers and consumers of healthcare have relied on traditional 'review' articles, which are generally prepared by 'content experts' in a field. Unfortunately, such reviews have been shown to be prone to a number of biases and their conclusions can be just as contradictory as the primary research (Mulrow, 1987). For example, content experts may come to a particular field with their own prejudices and there is a risk that the primary research will be plundered selectively in order to confirm the authors opinion (a "confirmatory bias"). The reader is left unclear as to how the primary studies have been selected for inclusion or how a particular conclusion has been reached. In the face of growing dissatisfaction with the lack of transparency of methods and lack of trust in the conclusions of traditional review articles, the *systematic review* article has emerged.

The need for and rationale behind systematic reviews is neatly summarised by Mulrow (1987):

'Through critical exploration, evaluation and synthesis the systematic review separates the insignificant, unsound and redundant, from the salient and critical studies that are worthy of reflection'

Systematic reviews adopt an explicit method in order to limit bias in the search, and selection of studies for review. This takes the form of extensive (including electronic) literature searches, followed by selection of the highest quality studies for review - ideally these should be randomised-controlled trials (RCTs), where these are available or feasible. This evidence is (where appropriate) synthesised in order to produce a clear message or conclusion regarding effectiveness. It may be summarised narratively, or in some cases it is possible to summarise the results of the primary studies quantitatively, in the form of a meta-analysis (NHS Centre for Reviews and Dissemination, 2000).

The use of such methods in mental health has a long history. Smith & Glass (1977) pioneered the use of mathematical techniques (and coined the term 'meta-analysis') in order to synthesise disparate, contradictory and under-powered studies of the effectiveness of psychotherapy. This pioneering meta-analysis found no differences in effectiveness between different psychotherapeutic techniques. From this early start the methods of systematic review and meta-analysis evolved rapidly, and have been employed to evaluate the effectiveness of a range of interventions in mental health. For example, systematic reviews have been used to examine: the comparative efficacy of new drug entities in both depression and schizophrenia (NHS Centre for Reviews and Dissemination, 1999a; Song, *et al.*, 1993); the value of family based psychotherapeutic interventions in schizophrenia (Mari & Streiner, 1994); and the relative merits of different methods of organising and delivering care to those with severe mental illnesses in the community (Marshall, *et al.*, 2001; Marshall & Lockwood, 2001).

There are reasons why systematic reviews are especially important in the sphere of mental health. Firstly, many interventions in psychiatry have been evaluated within randomised controlled trials, but the quality and real world validity of many of these data have been shown to be poor (; Hotopf, *et al.*, 1997; Thornley & Adams, 1998; Gilbody, *et al.*, 2001e). By bringing a systematic approach, then the highest quality and most applicable studies are separated from those that are of limited value within an explicit framework, or the dearth of high quality literature within a specific area is made explicit – thus identifying the need for primary research. Secondly, mental health is bedevilled by the problem of small sample size and under-powered research (Hotopf, *et al.*, 1997; Thornley & Adams, 1998). By applying (where appropriate) the mathematical technique of meta-analysis, then statistical power is brought to an area of enquiry where, in the absence of large-scale trials, this would

not otherwise be possible. A clear example of this is in the application of meta-analytic pooling to comparative trials of new and older anti-depressants, where in excess of 100 randomised trials are known to exist (Hotopf, *et al.*, 1997), but where no individual study has the statistical power to demonstrate the superiority of one class of drugs over the other, or their equivalence.

Stages in a systematic review

There are several stages in the conception, design and conduct of a systematic review. The following section provides a brief overview of the architecture of a systematic review, with particular reference to those stages that are especially important in the conduct of systematic reviews in both the sphere of mental health, and in the evaluation of interventions intended to enhance the quality with which healthcare services are delivered or organised. This section draws upon guidelines outlined by the NHS Centre for Reviews and Dissemination (2000), and the Cochrane Collaboration (Mulrow & Oxman, 1999), and upon personal experience in the conduct of systematic reviews in the sphere of mental health.

Formulation of a question or hypothesis.

A review begins with a research question or hypothesis, in the same way as a rigorous piece of primary research. This question should be capable of being examined in detail, and if specified as an hypothesis, capable of being confirmed or refuted. The formulation of a focussed question in systematic review contrasts with the broad scope or non-explicit question posed in a non-systematic or narrative review (Mulrow, 1987). Refinement of the question often involves making explicit several facets, including (1) the population of interest; (2) the setting in which research should be conducted; (3) the intervention; (4) the type of evidence that would provide a robust and believable answer; and (5) the outcomes that would be of interest. These form the basis of the inclusion and exclusion criteria of a systematic review, and are analogous to the inclusion and exclusion criteria of a primary piece of research.

An example of a clear research question comes from an early systematic review conducted by Mari & Streiner (1994), which examined the effectiveness of family based interventions for schizophrenia. Previous (non-systematic) reviews of this topic had yielded conflicting results and had failed to establish whether this therapy

helped those with schizophrenia, or the magnitude of any benefit (Gilbody & Sowden, 1999). These non systematic reviews had, amongst other things, failed to select the highest quality evidence, and had failed to set explicit criteria regarding what constituted family based interventions, leading to confusion of what was in fact being reviewed. Mari & Streiner (1994) began their review by specifying a clear research question and refining it as follows:

Population: Persons with schizophrenia, however diagnosed

Intervention: Any psychosocial intervention with relatives of those with schizophrenia that required more than five sessions and was not restricted to an in-patient context/environment and compared to a standard care.

Research design: randomised controlled trials

Outcomes:

- Suicide or all cause mortality;
- Relapse;
- Hospital admission;
- Drug compliance.

Deciding which research design to use

The type of research design to be included in a review is guided by the question being asked. Randomised trials provide the most robust evidence for questions relating to therapy or the effectiveness of a health policy intervention (Sackett, *et al.*, 1991). This is as true in mental healthcare as it is in other specialities (Gilbody, *et al.*, 2001e; WHO, 1991). This design maximises internal validity, such that observed differences can be attributed to the treatment under evaluation, rather than some confounding factor inherent in the patient, the clinician or the environment in which the treatment is given (Meinert, 1986). However, for questions relating to the best method of delivering and organising mental health services or about the best method of changing clinical practice for the better, then randomisation may still be appropriate, but it becomes either impossible or inappropriate to randomise individuals (Campbell, *et al.*, 2000). Instead it is individual hospitals, geographical areas, clinical teams or individual clinicians and, consequently, all the patients they treat, that should be randomised. When 'clusters' of individuals are randomised in

this way, there are profound implications for the design and analysis of studies (Campbell & Grimshaw, 1998).

In the case of interventions targeted at the individual clinician in order to improve the care of his or her patients, then there is a real danger that randomising individual patients to receive a quality improvement intervention will influence how other patients randomised to receive normal care will be managed (Ukoumunne, *et al.*, 1999b). This 'cross contamination' between subjects can potentially lead to a dilution of effect for an intervention, and can result in an effective intervention being shown to be ineffective in the context of an individualised clinical trial. For this reason, cluster based studies are often the most appropriate design when evaluating quality improvement strategies, such as those addressed within the current review (Gilbody & Whitty, 2001).

There are no known reviews that have applied a systematic methods to examine the potential for interventions to improve professional practice and the quality of mental healthcare specifically. However, there have been several reviews of quality improvement interventions across all areas of healthcare. Many of those have been conducted under the auspices of the Effective Practice and Organisation of Care group (Bero, *et al.*, 1998), within the Cochrane Collaboration. One early example is that of Grimshaw & Russell (1993), who examined the potential of guidelines to improve health and healthcare. This review utilised a range of designs – including both randomised and quasi-experimental designs in order to maximise the scope and number of individual studies that evaluated the impact of professional guidelines on the quality and delivery of healthcare. Many of the primary studies incorporated in this review fell short of the ideal of cluster randomisation in their design, or if they did use cluster randomisation, they failed to account for the effect of clustering in their design and analysis of results. The importance of cluster randomisation is not always realised by primary researchers, and the failure to account for this design feature leads to spurious conclusions – particularly type 1 errors in the analysis and interpretation of results. This has been termed unit of analysis error (Divine, *et al.*, 1992), and has led to calls for a greater acceptance and understanding of cluster-based studies (Campbell & Grimshaw, 1998).

In summary, randomised trials remain the gold standard for considering whether a treatment or quality improvement strategy helps or harms, and the cluster based study remains the most appropriate form of randomised trial in quality improvement

strategies. For this reason, a systematic review of quality improvement strategies should consider cluster randomised studies as the gold standard. However, experience suggests that this ideal is not always met (Gilbody & Whitty, 2001).

Searching for studies

A cornerstone of the systematic approach to reviewing empirical research is the explicit aim of including all possible relevant literature, and taking active steps to search for this literature. As is the case with primary research studies, flaws in the data collection can invalidate the results of a systematic review. As many relevant primary studies as possible must be collected in order to minimise random error and bias (Counsell, 1998).

For this reason, systematic reviews involve the search of a diverse range of sources in order to maximise the likelihood of finding all relevant studies. This ideally involves the following sources

- Electronic databases
- Manual searches of journal, conference proceedings and books
- Reference lists
- Existing study registries
- Pharmaceutical companies (where relevant)
- Personal contact with colleagues

The advent of electronic databases has greatly enhanced the ability of researchers to have ready access to large amounts of bibliographical information. However, a crucial step in conducting a systematic review is ensuring that the required information is retrieved efficiently from the huge quantities of data contained in bibliographic databases. Two components of a comprehensive search strategy are therefore the actual databases searched and the search terms that are chosen.

The most readily available electronic database is MEDLINE, and many systematic reviews use this as their only electronic source of primary studies (Suerez-Almar, *et al.*, 2000). There are several problems with this approach. First, MEDLINE does not contain the totality of published medical literature, but a selected portion of it. It places an undue emphasis on North American and English language publications, and ignores many important European journals. Second, MEDLINE incorrectly codes many publications, particularly with respect to their study design. For this

reason simple searches of MEDLINE only identify less than 40% of relevant studies (Adams, *et al.*, 1992; Dickersin, *et al.*, 1985; Suerez-Almar, *et al.*, 2000). Searches must therefore extend their coverage of databases to include ones that are more likely to cover Journals excluded from MEDLINE (Suerez-Almar, *et al.*, 2000). EMBASE is another electronic database, which provides a useful complement, since it has a more European focus and carries many of the journals excluded from MEDLINE. Additionally, electronic searches must incorporate terms that do not depend on the incomplete coding structure imposed by the collators of medical databases. In the case of randomised trials, this involves inclusive mixtures of free text words and medical subject headings. A typical search strategy developed in order to identify randomised trials is given in Table 24.

Locating studies relating to mental health are especially problematic, since databases such as MEDLINE and EMBASE fail to carry important mental health journals (Adams, *et al.*, 1992; Adams, *et al.*, 1994). For this reason more topic specific databases, particularly PsycLIT, published by the American Psychological Association are recommended (Hay, *et al.*, 1996).

Table 24: An example of a comprehensive electronic search strategy

SILVER PLATTER OPTIMAL SEARCH STRATEGY FOR RANDOMISED CONTROLLED TRIALS
From (Dickersin & Larson, 1996).

#1	RANDOMIZED-CONTROLLED-TRIAL in PT
#2	CONTROLLED-CLINICAL-TRIAL in PT
#3	RANDOMIZED-CONTROLLED-TRIALS
#4	RANDOM-ALLOCATION
#5	DOUBLE-BLIND-METHOD
#6	SINGLE-BLIND-METHOD
#7	#1 or #2 or #3 or #4 or #5 or #6
#8	TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#9	#7 not #8
#10	CLINICAL-TRIAL in PT
#11	explode CLINICAL-TRIALS
#12	(clin* near trial*) in TI
#13	(clin* near trial*) in AB
#14	(singl* or doubl* or trebl* or tripl*) near (blind* or mask*)
#15	(#14 in TI) or (#14 in AB)
#16	PLACEBOS
#17	placebo* in TI
#18	placebo* in AB
#19	random* in TI
#20	random* in AB
#21	RESEARCH-DESIGN
#22	#10 or #11 or #12 or #13 or #15 or #16 or #17 or #18 or #19 or #20 or #21
#23	TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#24	#22 not #23
#25	#24 not #9
#26	TG=COMPARATIVE-STUDY
#27	explode EVALUATION-STUDIES
#28	FOLLOW-UP-STUDIES
#29	PROSPECTIVE-STUDIES
#30	control* or prospectiv* or volunteer*
#31	(#30 in TI) or (#30 in AB)
#32	#26 or #27 or #28 or #29 or #31
#33	TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#34	#32 not #33
#35	#34 not (#9 or #25)
#36	#9 or #25 or #35

- Upper case denotes controlled vocabulary.
- Lower case denotes free-text terms.

The limitations of electronic databases can be in some way remedied by supplementing search strategies with hand searches of key journals – particularly those most likely to carry research that is likely to be of relevance. These can be specialist journals relevant to the area under review, or journals most likely to carry the type of research that is of interest. One of the lasting contributions of the Cochrane Collaboration has been the international voluntary effort that has been undertaken to hand search important biomedical journals, and to include possible controlled trials in a specialist register – the Cochrane Controlled trials register (The Cochrane Controlled Trails Register, 2000). Similarly, specific research groupings within the Cochrane collaboration have hand searched important journals and conference abstracts in order to identify potential controlled trials. Two important registers in sphere of mental health are those maintained by the Schizophrenia group (Adams, 1998), and the Depression and Anxiety group (McGuire, 1998) within the Cochrane Collaboration.

Appraising the quality of studies

A crucial step in the conduct of a systematic review is not just the selection of the highest quality research design, but is a fundamental consideration of the quality of the individual studies and an examination of the potential sources of bias. NHS Centre for Reviews and Dissemination (2000) outlines five reasons for examining the quality of individual studies in systematic reviews:

- To determine a minimum quality threshold (study design threshold) for the selection of primary studies
- To explore quality differences as an exploration for heterogeneity in study results
- To weight the study results in proportion to quality in meta-analysis
- To guide the interpretation of findings and to aid in determining the strength of inferences
- To guide recommendations for future research

Table 25 provides some important concepts central to the assessment of study quality (NHS Centre for Reviews and Dissemination, 2000).

Table 25: Terminology used in study quality assessment

Study quality	**The degree to which a study employs measures to minimise biases, focussing on internal validity. A set of parameters in the design and conduct of a study that reflect the validity of the outcome, related to the external and internal validity and the statistical model used**
Bias (systematic error)	**A tendency to produce results that depart systematically from the true results. Unbiased results are internally valid.**
Internal validity	**The degree to which the results of a study are likely to approximate to the truth for the group studied. It is a pre-requisite for external validity.**
External validity	**The extent to which the effects observed in a study are applicable outside of the study on a different population (in routine practice).**

Several methods are used in order to protect against biases in primary research, and which should be considered in judging the quality of that research within systematic reviews. Perhaps the most important is that of ensuring that participants in trials are allocated in a truly random manner, and ensuring that allocation is concealed, such that the person randomising the individual participant has no foreknowledge of which group the participant will be allocated (Jadad, *et al.*, 1996; Schulz, *et al.*, 1995). Failure to protect against this bias has been empirically shown to be related to the magnitude of effect obtained in primary research (Schulz, *et al.*, 1995), with poorly randomised and inadequately concealed trials showing spuriously elevated effect sizes.

Classifying studies according to their level of methodological rigour is important in identifying those studies that are of better quality, and should guide the data synthesis within a systematic review. This can be done in a number of ways, by for example, setting a minimum quality threshold above which studies will be considered within a review. Alternatively, methodological quality can be used to examine differences obtained between studies, as an exploration of heterogeneity (see below). Where heterogeneity exists, then reviewers can place greater weight on higher quality studies.

The consideration of quality can be undertaken in several ways within systematic reviews, with the use of (1) individual quality components or items; (2) quality checklists, or (3) quality scales. There are pros and cons of each of these approaches (Greenland, 1994; Juni, *et al.*, 1999). Ideally there should be both an empirical and theoretical basis for the items that are thought to represent quality.

The use of scales has been criticised due the fact that they lack psychometric rigour. For example, they are unreliable, and remain unvalidated, and contain an arbitrary number of items, with weights ascribed without any empirical basis (Juni, *et al.*, 1999; Moher, *et al.*, 1996; Moher, *et al.*, 1998). Of the scales that have been developed, that of Jadad, *et al.* (1996) contains only items that have been empirically shown to be related to effect size (Moher, *et al.*, 1999b). Further, this scale has been shown to be reliable in its application as a generic measure of study quality in randomised trials (Clark, *et al.*, 1999; Jadad, *et al.*, 1996). The contents of this particular scale are shown in Appendix 5. A widely used alternative or complementary scale is that adopted within the Cochrane Collaboration (Mulrow & Oxman, 1999), which classifies controlled trials on the basis of their likely susceptibility to bias. This is outlined in Table 26

Table 26: Examining likelihood of bias

Cochrane Handbook classification of bias (Mulrow & Oxman, 1999)
A. Low risk of bias (adequate allocation concealment)
B. Moderate risk of bias (some doubt about the results)
C. High risk of bias (inadequate allocation concealment)
The Jadad Scale (Jadad, *et al.*, 1996) measures a wider range of factors that impact on the quality of a trial. The scales includes three items:
1. Was the study described as randomised?
2. Was the study described as double-blind?
3. Was there a description of withdrawals and dropouts?

Synthesising research

The aim of data synthesis in a systematic review is to collate and summarise the results of included primary studies. There are broadly two approaches to data synthesis. The first is a non-quantitative or descriptive data synthesis, whereby all the evidence is collated in a uniform manner, including all the information about the characteristics of individual studies, together with their results. This first approach allows an overview of the totality of potentially important evidence. The second approach involves a quantitative data synthesis, whereby component studies have been judged sufficiently similar in their design and choice of outcome to justify a statistical pooling, in order to extract some overall measure of effect. The first step

of non-quantitative data synthesis is a necessary step for all systematic reviews, but the quantitative pooling is not always justified.

The descriptive or non-quantitative synthesis involves the extraction and presentation of data in a rigorous manner, according to a format that had been decided a priori, which seeks to summarise important design features and results of individual studies in a tabular form. NHS Centre for Reviews and Dissemination (2000) recommends that the following may be presented:

- Population
- Interventions
- Settings
- Outcomes measures
- Validity of the evidence
- Sample sizes and results of the studies included in the review

By presentation of a descriptive synthesis, those conducting a review should be able to assess whether participants, interventions and outcomes allow generalisation of the results of the review, or whether there are restrictions or omissions that limit the applicability of the results. Similarly, those conducting a review will then be able to make some informed comment as to whether the quality of the research is sufficient to allow believable conclusions to be drawn on the basis of the results. Lastly, such a presentation will allow a decision to be made regarding the use of a formal quantitative synthesis, and will give clues as to important differences between studies that might explain differing results across studies (heterogeneity).

A clear example of the potential of non-quantitative data synthesis to provide important insights into the general effect of an intervention and to highlight important ingredients of an effective intervention comes from the previously mentioned review by Grimshaw & Russell (1993). Widely differing guidelines implementation programmes in all spheres of healthcare were drawn together using non-quantitative data synthesis. Important differences in terms of the intervention and the outcomes studied were highlighted. These showed that a number of studies had demonstrated that guidelines had the potential to influence practice in a positive way, but the setting and choice of outcome precluded the use of formal quantitative data synthesis. A number of rigorous studies did show a positive effect, and certain features in the nature of the guideline intervention united these positive studies. These included the fact that guidelines that had been shown to influence practice

when they: are based upon an explicit consideration of the evidence; take into account local circumstances; are disseminated by an active educational intervention; and are implemented by patient specific reminders relating directly to professional activity.

Where studies are judged sufficiently similar in terms of populations, interventions then a formal statistical pooling, known as meta-analysis can at least be attempted. If the studies are too heterogeneous to combine, then subgroups of similar studies can identified within the descriptive synthesis outlined above, and can be pooled separately. When a decision has been made that a meta-analysis is both possible and appropriate, reviewers have to decide three things in advance: (1) Which comparisons should be made; (2) which outcome measure should be used; (3) which effect measure should be used to describe effectiveness.

Meta-analysis involves the statistical pooling of individual studies using some common metric between studies. Commonly used techniques allow individual studies to be weighted in some way, such that those which are judged to be of greater importance are given greater prominence and contribute to a greater extent to the overall summary effect size. The most commonly used weighting function is by sample size or study variance, such that larger studies are given greater prominence than smaller studies. The measures of effect that are combined are of a dichotomous or continuous nature, and specific techniques are available for each of these scenarios (Cooper & Hedges, 1994).

A clear example of the potential for quantitative pooling to be applied when examining the organisation and delivery of mental health care comes from a review by Marshall & Lockwood (2001). This review begins with an important policy based question, relating to the effectiveness of one model of community care that had been widely advocated in UK mental health services. Marshall & Lockwood (2001) found eleven randomised studies, which were largely underpowered. Common outcomes measures between studies were the effect of case management on ongoing contact with psychiatric services and admission to hospital. Summary pooled odds ratios were calculated showing that whilst case management was effective in increasing the ability of psychiatric teams to maintain contact with those with severe mental illness (pooled odds ratio = 0.70; 99% Confidence Interval 0.50-0.98; n=1210), this also resulted in a much higher rate of psychiatric admissions (pooled odds ratio 1.84; 99% Confidence Interval 1.33-2.57; n=1300).

Two alternative statistical models used in pooling data are the fixed effects model (Mantel & Haenszel, 1959) and the random effects model (DerSimonian & Laird, 1986), which differ in their underlying assumptions relating to the distribution of effect sizes from individual studies and their relation to some underlying 'true' effect size (Cooper & Hedges, 1994). In practice there is little difference between these two methods, and their application (Sutton, *et al.*, 1999). However, there are important differences when heterogeneity between studies is present (see on).

Examining heterogeneity

An integral step in any formal data synthesis, either quantitative or non-quantitative, is a consideration of the important differences that exist between studies (Thompson, 1995). This has been termed 'heterogeneity' and relates to the fact that different study results will be expected when different interventions are performed on different populations and in different settings. One of the main criticisms that have been levelled at the use of systematic reviews in general and meta-analysis in particular is the fact that studies that are fundamentally dissimilar are brought together when they ought not to be. This represents the vain hope of achieving something that is greater than the sum of individual pieces of research literature. Alvan Feinstein (1995) has highlighted this concern and termed meta-analysis a form of *'statistical alchemy'*. Heterogeneity is a reflection of the true nature of primary research evidence, and as such evidence should be sought for the existence of heterogeneity in outcomes between studies, and causes of these differences should be sought. Naylor (1989) outlines some important causes, which are given in table 27.

Table 27: Sources of heterogeneity

Ways in which apparently similar trials may differ (Naylor, 1989)
• Differences in inclusion and exclusion criteria
• Other pertinent differences in baseline states of available patients despite identical selection criteria
• Variability in control or treatment interventions (e.g. doses, timing, brand)
• Broader variability in management (e.g. pharmacological co-interventions, responses to intermediate outcomes including crossovers, different settings for patient care).
• Differences in outcome measures, such as follow up times.
• Variation in analysis, especially handling of withdrawals, dropouts, crossovers, etc.
• Variation in quality of design and execution, with bias of imprecision in individual estimates of treatment effect.

The preliminary non-quantitative data analysis outlined above should provide important insights into any clear differences that exist between studies, both in terms of their design, or results obtained. Several steps are recommended for a more formal examination of heterogeneity, including a visual inspection of plots of effect size, supplemented with more formal statistical tests of heterogeneity (Sutton, *et al.*, 1999). Where heterogeneity is discovered, then important causes should first be sought to explain this effect, and conclusions drawn on the likely importance of heterogeneity in the real world application of research evidence.

An example of the exploration of heterogeneity in mental health services research comes from a review of comparative studies of new (and relatively expensive) versus old (and relatively inexpensive) anti-schizophrenia drugs (Geddes, *et al.*, 2000). Preliminary analysis of the data showed that the choice of comparator was almost always haloperidol, one of the oldest typical anti-schizophrenia drugs, with perhaps the greatest propensity to cause distressing side effects. Many of the trials used comparator doses of haloperidol way in excess of those doses used in routine practice, and at dose levels that are likely to cause these side effects in most people. An overall benefit for newer drugs in terms of their tolerability was found in the meta-analysis, but substantial between study variation was noted. A plausible explanation for this between study differences was postulated as being the dose of comparator that was chosen. A formal examination of this hypothesis was made in two ways. First statistical tests of heterogeneity were made, and found to be significant. Second, studies with comparator doses above and below a clinically accepted level were pooled separately and were shown to give substantially different results. For studies with lower doses of haloperidol, there was found to be a much smaller benefit of new over older drugs.

Statistical pooling of heterogeneous data is therefore potentially misleading, and examination heterogeneity can provide important insights into the mode of action of an intervention and can reveal aspects that are of use in clinical practice. For example, Geddes, *et al.* (2000) use their meta-analysis and examination of heterogeneity to postulate that recommendations for the first line use of new anti-schizophrenia drugs have been based on naïve summaries of the randomised evidence. They suggest that a more suitable starting dose of typical drugs should be above the level where clinical efficacy is demonstrated, and below the level where adverse side effects begins to affect compliance. This dose of older anti-psychotic, rather than an expensive newer drug might form the first line treatment of

schizophrenia. They also suggest that this hypothesis might be usefully tested within a prospective randomised trial.

Examining publication bias

In addition to unexplored heterogeneity, another threat to the validity of the results of systematic reviews is the existence of publication bias. Publication bias is that which arises when research fails to be published on the basis of the direction or significance of its results (Dickersin, 1990). It has long been recognised that research with 'negative' or 'uninteresting' results is less likely to be published (Sterling, 1959). One important landmark in the recognition of the potential of publication bias to produce misleading results comes from the meta-analyses of trials of magnesium following myocardial infarction (Egger & Davey-Smith, 1995). In this case a spurious benefit from magnesium was seen from the systematic review of a data-set affected by publication bias. There are reasons to suppose that psychiatric research is just as prone to publication bias as the research produced by other specialities.

The preceding discussion on the search for trials, particularly through the use of electronic databases, places a particular emphasis on the search for published data. Much research goes unpublished, and this is particularly difficult to identify. Unfortunately there are sound reasons to suppose that psychiatric research is as prone to publication bias (Gilbody & Song, 2000; Gilbody, *et al.*, 2000). Mental health research displays many of the risk factors for the occurrence of publication bias (Dickersin, 1990), particularly: small sample size; commercial influence; and poor methodological quality (Thornley & Adams, 1998; Gilbody, *et al.*, 2000).

One method by which publication bias can be examined is through the use of funnel plots (Light & Pillemer, 1984), which rely on the fact that larger studies are generally published, irrespective of their results, whereas smaller studies are only selectively published. Funnel plots chart effect size against sample size. In the absence of publication bias, when all studies are plotted as individual data points on this graph, then a symmetrical funnel should be seen (Figure 11). Results from small studies are prone to greater random variation and will consequently be dispersed symmetrically (i.e. randomly) about some central overall effect size (i.e. that which would be obtained in a well-powered study). As sample size increases then the degree of random variation about this central axis decreases, producing a

symmetrical inverted funnel. If some smaller studies are excluded non-randomly because of their direction of effect, then the funnel will look asymmetrical about its base. Funnel plot asymmetry is highly suggestive of publication and related bias (Egger, *et al.*, 1997) and suggests that the research in the area under review should, at the least, be treated with suspicion.

One example of publication bias in the sphere of psychiatric research relates to published randomised trials of risperidone, a new atypical anti-schizophrenia drug (Kennedy, *et al.*, 2001). A funnel plot of randomised studies (figure 12) shows that the research on this compound is subject to a probable publication bias (Gilbody, *et al.*, 2000).

Given the inherent dangers of reviewing an area that is subject of a publication bias, and the danger that this can produce a biased or misleading result, then reviews should at the very least check for the existence of publication or related bias, through the use of funnel plots. Publication bias is just one of several potential causes of funnel plot asymmetry, and other causes may need to be considered (Egger, *et al.*, 1997). These include:

- Different intensity of intervention;
- Differences in underlying baseline risk;
- Poor methodological design of smaller studies;
- Heterogeneity;
- Inadequate analysis;
- Fraud;
- Chance

The use of funnel plots should therefore be seen as an extension of the general examination of heterogeneity in systematic reviews and the cause of asymmetry should be actively sought, before attributing publication bias (Gilbody, *et al.*, 2000).

Drawing conclusions from reviews

The forgoing discussion suggests a rigorous and methodically replicable framework for reviewing primary clinical research. The endpoint of such a review is a clear summary of the strengths and limitations of the existing research knowledge base, from which conclusions can be drawn. It is important that these conclusions should not make more of the data than is justified. Clear examples exist of both reviews

which have produced a clear message for health practice and policy, and which have the potential to influence the way in which healthcare is delivered (e.g. Marshall, *et al.*, 2001; Marshall & Lockwood, 2001). An equally valid endpoint of systematic review is a clear acknowledgement that the existing research is insufficiently strong to guide practice and policy, and that in conclusion more research is needed (eg Hotopf, *et al.*, 1997; Song, *et al.*, 1993). This is not a failure of systematic review methodology, but rather an honest reflection of the limitations of the research knowledge base.

Figure 11: Symmetrical funnel plot indicating no publication related bias

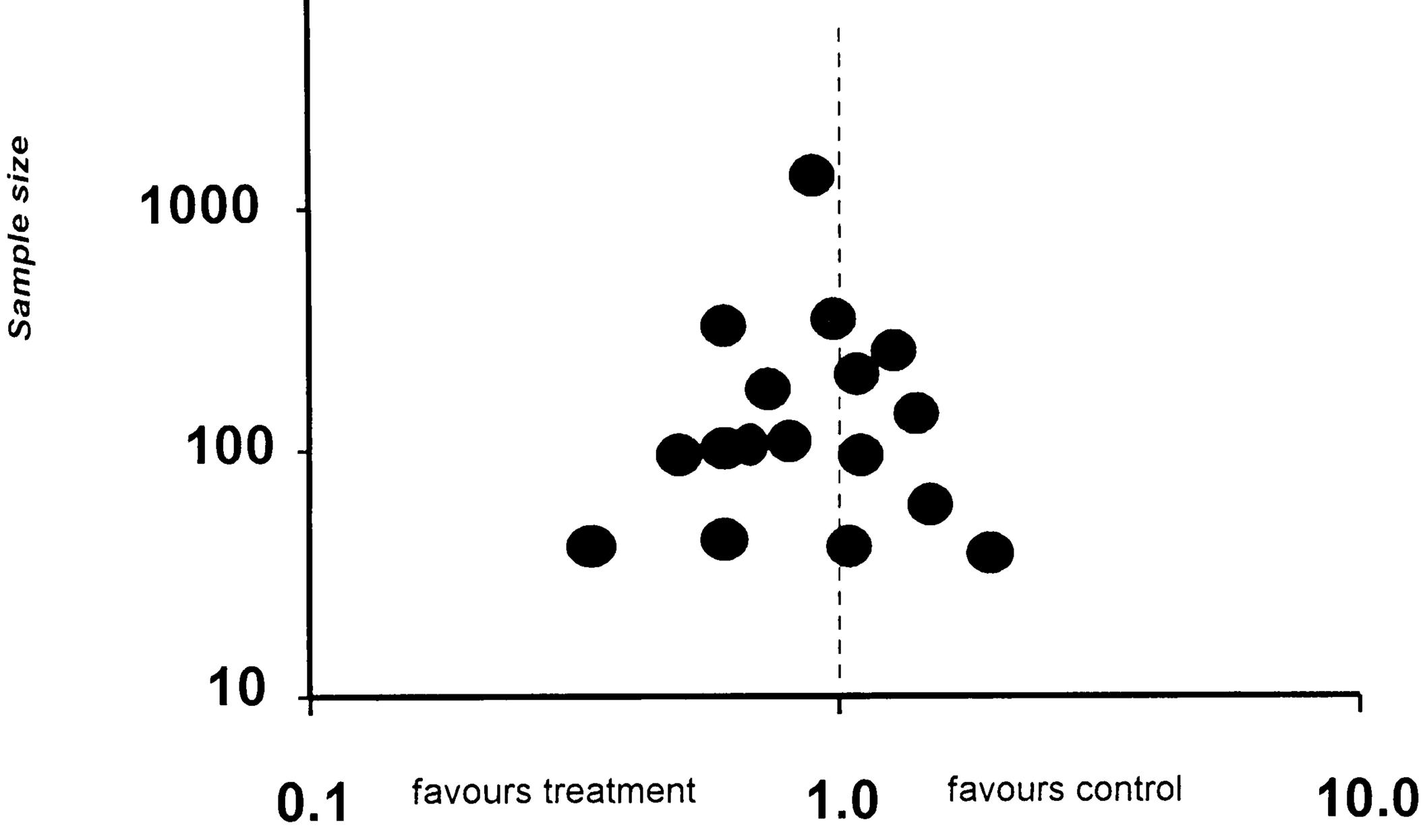
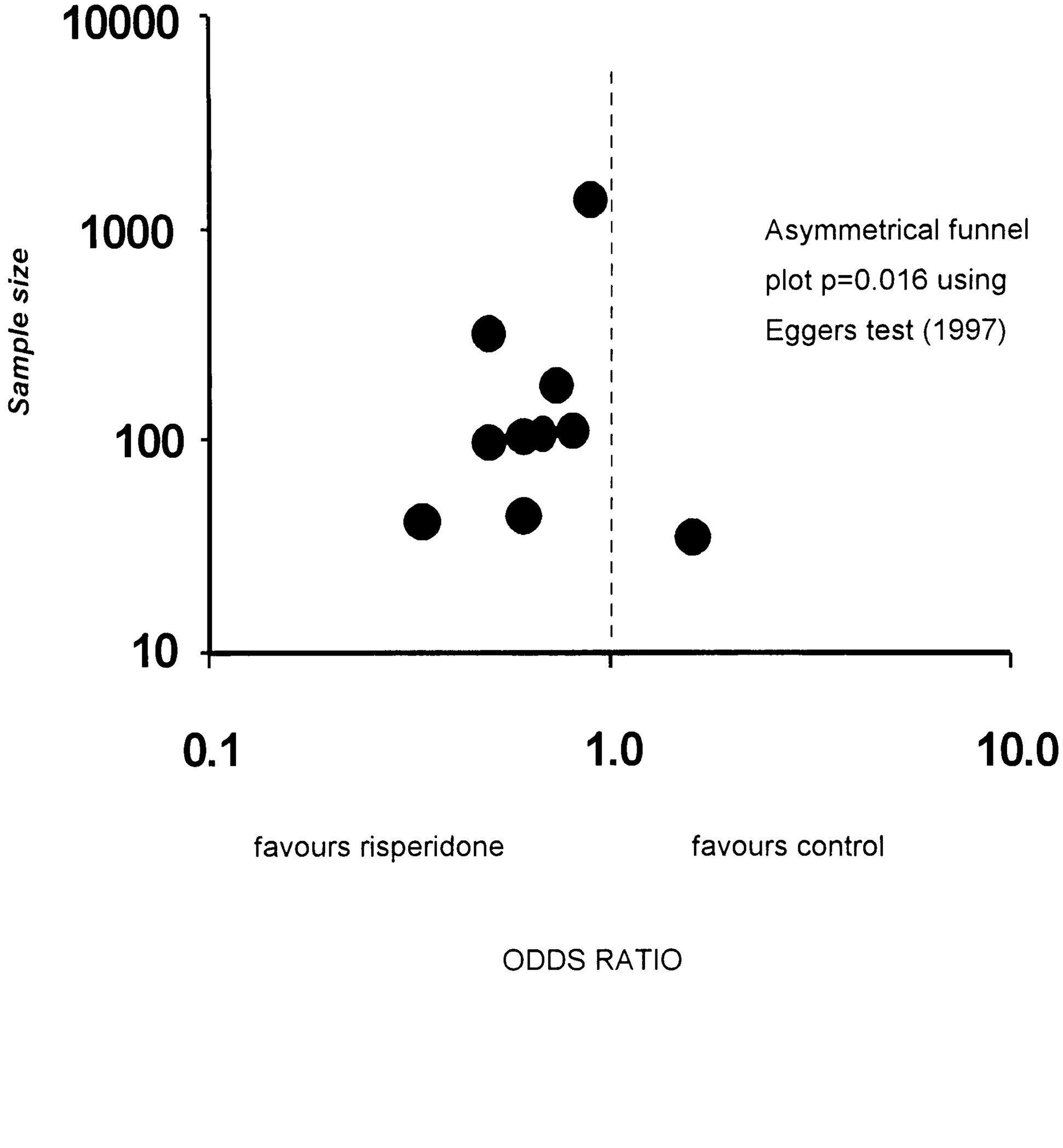


Figure 12: Asymmetrical funnel plot

Short term clinical improvement in randomised trials comparing risperidone with conventional neuroleptics (Kennedy, *et al.*, 2001; Gilbody, *et al.*, 2000)



Chapter 25 Methods of the review

Having outlined the rationale, general methods and potential difficulties of systematic reviews in general, the following section outlines the specific methods employed in the review of the use of routine outcomes measurement in mental illness.

The methods employed in this systematic review follow guidelines laid down by the NHS Centre for Reviews and Dissemination (2000), and adhere to methods outlined in the Handbook of the Cochrane Collaboration (Mulrow & Oxman, 1999). The review was conducted under the under the auspices of both the Depression and Anxiety Group, and Schizophrenia Group of the Cochrane Collaboration. As such, the review has been subject to extensive peer review in addition to that offered by the supervisors of this project, and will be published in the Electronic Cochrane Library (Gilbody, *et al.*, 2001b; Gilbody, *et al.*, 2001c). Part of this review has also been published in paper format (Gilbody, *et al.*, 2001d).

Study inclusion criteria

Patients

In order to examine the impact of routine outcomes assessment on patients with psychiatric illnesses (or with unrecognised psychiatric illness) in all settings, not just those being cared for in psychiatric settings, it was decided to make the patient inclusion criteria quite broad.

To be included, studies must have included one of the following patient populations:

- Patients in non-psychiatric settings. This includes general hospital patients and non-selected general practice patients.
- Patients with psychiatric disorders being managed by specialist psychiatric services.

Studies relating to the following patient groups were excluded from this review:

- Patient groups whose primary problem is one of substance abuse or who are managed in specialist substance abuse services.
- Child and adolescent populations
- Those with learning disabilities or dementia.

Interventions

To be included, studies must have compared the introduction of a routine form of outcome or needs assessment with a normal routine pattern of care.

Routine care (the control/comparator condition) involved usual patient-doctor interaction, with non-standardised history taking, investigation, referral, intervention and follow up. This would not usually involve the use of outcome measurement instruments by clinicians, but would have relied on the traditional channels of patient doctor communication and informal assessment of outcome using clinical history taking, psychiatric/physical examination and recording of progress in clinical notes.

The active intervention should have involved the addition of a standardised outcome assessment instrument to routine care. The outcome assessment should have been made either by the patient or by the clinician, but the active intervention will involve the information from the outcome assessment being fed back to the clinician or being incorporated into routine care procedures (such as outpatient assessment, hospital admission or routine discharge planning). Hence, standardised outcome could have been assessed in both intervention and control conditions, but the active component in an intervention involved the feeding back of this information to the clinician

Any potential form of assessment was classified as one of the four following types (for definitions see earlier chapters):

A. **An assessment tool measuring psychiatric symptoms**. This included instruments that measure the core (diagnostic) features of the disorder under evaluation.

B. **An assessment tool measuring 'Patient Based Outcome'.** These tools measure more than 'symptom severity' and assess the impact of illness on the individual - in terms of all or some the following domains: social functioning; role functioning; mental well-being; cognitive functioning (after Ware (Ware, 1995).

C. **An assessment tool measuring 'patient need'.** These tools measure unmet emotional, physical, social and financial needs of the individual patient (Brewin

& Wing, 1993; National Institute of Mental Health, 1987; Thornicroft, *et al.*, 1992), and must explicitly identify themselves as 'needs assessment tools'. Whilst there is a potential degree of overlap with 'patient based measures' (as defined above) in terms of the domains that are included, these are considered separately. Needs assessment instruments have evolved from a different tradition within mental health services research and place an explicit consideration of the identification of unmet need in their conception and use (Thornicroft, 1996), rather than the 'evaluative' approach which is inherent in the perspective of 'outcome measurement' (Jenkinson, 1994; van den Bos & Triemstra, 1999). However, their similarity to patient based assessments of outcome justified their consideration in this review.

D. Other assessment tools

Some widely used or heavily promoted measures do not fit easily not any of the above mutually exclusive categories, since they often measure combinations of all three. For example, the Health of the Nation Outcome Scale (HoNOS) (Wing, 1994) combines elements of clinical symptoms and hospital service use, together with items that might be considered 'patient based' in their focus (such as social functioning). These were included in a final 'miscellaneous' category, and if used, their content and focus will be discussed in detail within the review.

The above instruments defined in *A-D*, will hereafter be collectively referred to as *outcome and needs assessment tools.*

Design

Controlled clinical trials were included. In the absence of randomised evidence, then non-randomised or quasi-randomised controlled trials were considered. The most rigorous and robust controlled design for this intervention was considered to be the cluster based randomised trials, whereby individual clinicians or clinical teams form the unit of randomisation (Ukoumunne, *et al.*, 1999b). The degree to which authors accounted for clustering in the design and analysis of their trials is discussed in the section entitled 'quality assessment' outlined below.

Outcomes

Outcomes were studied as they were defined by the authors of studies, with particular attention to the impact of outcome and needs assessment tools on the following:

- Overall clinical improvement (as defined by individual studies)
- Patient based outcome (including social functioning, role functioning, mental well being and cognitive functioning)
- Hospital status, either discharge, readmission or length of stay (as defined in individual) trials;
- Intervention for an identified problem
- Resource uses
- Employment status
- Independent living
- Death (both as suicide and other causes)
- Costs (direct and indirect).

It should be noted that several of these outcomes potentially included the outcome that was the focus of the evaluation - i.e. that which was actively incorporated into routine care by being fed back to the clinician. The determination of the success or otherwise of the intervention according to this criterion requires that outcome was measured in both the intervention and control conditions using the instrument under evaluation. In this scenario, the experimental condition differed from the control condition in that this information was actively fed back to the clinician in the experimental condition. Equally, several of the other outcomes, particularly those relating to service use and patterns of referral, could be measured form administrative records and sources, and did not require that the outcome instrument under evaluation is administered to both intervention and control conditions.

Outcomes were grouped into those measured in the short term (up to 12 weeks), medium term (13 to 26 weeks) and long term (over 26 weeks).

Additionally, processes of care were described in individual studies if these are recorded as a criterion with which to evaluate the success of routine outcome assessment. Examples of potentially important processes included: (1) clinician and patient perceptions of the usefulness or acceptability of measurement instruments;

(2) self-reports of the use of outcome information in changing patient management;
(3) rates of referral to outside agencies.

Search strategy

The following bibliographic data bases were searched between September and December 1999: Medline; Embase; Cinahl; PsycLit; Cochrane Controlled Trials Register.

The search strategy combined two sets of search terms relating to the target patient population and the intervention – full details of the search strategy, and the rationale behind its development are given in Appendix 2:

Patient population: a search strategy is used which captures publications relating to all forms of mental illness using MeSH terms (Appendix 2 for development, refinement and exact details of this strategy)

Intervention: an already developed search strategy was used which has been shown to have acceptable sensitivity and precision in identifying research which relates to outcome and needs assessment (Brettle, *et al.*, 1998) (see Appendix 2 for development, refinement and exact details of this strategy).

Titles and abstracts from electronic searches were scrutinised and all potentially relevant articles were obtained. Reference lists were scrutinised for additional studies. The results of these searches are summarised in table 28.

It will be seen from table 28, that the search strategies were relatively insensitive, with search strategies identifying large numbers of studies of which only a small portion were relevant. The primary reason for this is the ubiquity of the term 'outcome' in electronic abstracts. Approximately 17,000 of the identified studies (>90%) were in fact primary studies which were of no direct relevance, but which were picked up by the electronic searches by virtue of the presence of the term 'outcome' as part of their structured abstract. Attempts to refine this search were unproductive and meant that studies which were already known to exist and fulfil the inclusion criteria were not identified in electronic searches.

The most fruitful database in terms of potentially relevant studies was MEDLINE, with other databases identifying relatively few studies that were eventually found to fulfil the inclusion criteria

In addition the following journals were hand searched:

- British Journal of Psychiatry 1976-1999 (no additional studies)
- American Journal of Psychiatry 1976-1999 (no additional studies)
- Archives of General Psychiatry (no additional studies)
- Psychological Medicine (no additional studies)
- Quality of Life Research (no additional studies)
- Journal of Psychosomatic Research (no additional studies)
- Medical Care (four additional studies).

Data extraction

The following data were extracted from studies, and were entered in a Microsoft Access database (Microsoft Corporation, 1998).

- Author & Year;
- Design;
- Population;
- Setting;
- Sample size;
- Routine outcome measure used;
- Intervention and control conditions;
- Length of follow up & outcomes studied;
- Results.

Quality assessment

Study quality was assessed in two ways.

First, studies were judged according to accepted quality assessment criteria, using the Jadad scale (Jadad, *et al.*, 1996), the criteria of Schulz (Schulz, *et al.*, 1995) and Cochrane criteria (Mulrow & Oxman, 1999). Particular attention was paid to the method of randomisation, such that those studies that described themselves as randomised, but did not describe an adequate method of randomisation and concealment of allocation were distinguished from those that did. Full definitions of items and a data extraction sheet are given in appendix 5.

Secondly, the unit of randomisation was established. Cluster randomised studies were considered to be superior to non-cluster based studies. For those studies in which the unit of randomisation was by clinician or clinical population, rather than individual patients, evidence was sought that clustering had been incorporated into the design and analysis of the study by the authors (Ukoumunne, *et al.*, 1999a).

Data analysis and synthesis

First, a non-quantitative data synthesis was applied. Study design features and results were tabulated. All results presented by authors were recalculated (where possible) from data presented in publications according to the following methods:

Dichotomous data. Discrete dichotomous outcomes, for example recognition of a specific psychosocial problem or admission to hospital, were summarised as rate ratios (also known as risk ratios or relative risks), absolute rate differences and Numbers Needed to Treat (NNTs) (Mulrow & Oxman, 1999), according to the following formulae:

	Outcome present	Outcome absent
Intervention group	a	b
Control group	c	d

Rate in intervention group = a/(a+b)

Rate in control group = c/(c+d)

Rate difference = (a/(a+b))-(c/(c+d))

Rate ratio = (a*(c+d))/(c*(a+b))

Confidence intervals for rate differences and ratios were calculated using Stats Direct version 1.7 (Buchan, 2000).

For studies that did not provide sufficient data to allow primary calculations to be made, then first authors were contacted in search of this information. If this was not

forthcoming, then the data presented by authors (such as p values or unverified rate differences) were included in tables. Where it was reported in studies that there had been losses to follow up, or if patients were randomised but not accounted for in the results, then these were assumed to have not had a positive outcome. In other words, an intention to treat analysis was reconstructed (Meinert, 1986).

Continuous outcomes: where continuous outcomes, such as scores on a psychometric scale were presented, then change and endpoint scores for each group were sought, together with their standard deviations, if available.

Economic outcomes: Where data were collected on resource use and economic outcomes, then these data were presented as reported by the authors of individual studies, together with details of the measurement of direct and indirect costs, the currency and time frame under which cost data were recorded. Where the authors of the individual studies conducted a synthesis of clinical and cost data, then these were presented. Further reanalysis of cost data was not attempted.

Once tabulated, important similarities and differences in terms of design and outcome were sought. Individual studies were judged to be overall positive or negative according to the following taxonomy:

- *positive - if the majority of major outcomes are statistically significant in favour of the intervention;*
- *borderline positive - if majority of outcomes are positive but non significant or have a unit of analysis error;*
- *mixed effect;*
- *borderline negative - if majority of outcomes are negative but non significant or have a unit of analysis error;*
- *negative - if the majority of major outcomes are negative and statistically significant.*

For those studies that were sufficiently similar in terms of their patients, settings, intervention and choice of outcome, then a formal data synthesis was attempted according to the following method.

For data that were felt to be sufficiently similar, a random effects meta-analysis using the methods outlined by (DerSimonian & Laird, 1986) was conducted using STATA version 6.0 (STATA corporation, 1999). This method can be used to pool both dichotomous and continuous data, and weights studies by their individual variance or sample size. Random effects meta-analysis assumes that the individual studies may be estimating different underlying effect sizes, and that these underlying effects are assumed to vary at random within this model. This variation, in addition to the variation caused by sampling error is incorporated into an overall estimate of effect size and attendant confidence limits. In the absence of substantial heterogeneity, then random pooled effect size approximates that obtained from a fixed effects model (NHS Centre for Reviews and Dissemination, 2000; Sutton, *et al.*, 1999). Where substantial evidence of statistical heterogeneity was found (see below), then sources of heterogeneity were sought. For unexplained heterogeneity, no formal meta-analysis was conducted.

Examination of heterogeneity

Heterogeneity between studies was examined by:

1. Looking for important differences between studies in terms of their design, following the non-quantitative synthesis of tabulated data
2. Inspection of plots of individual point estimates of outcome (Forrest plots)
3. Statistical tests of heterogeneity.

Inspection of plots of individual studies (Forrest plots) usually reveals obvious heterogeneity when the 95% confidence intervals of individual studies do not overlap (NHS Centre for Reviews and Dissemination, 2000). This can be supplemented by formal statistical tests such as Cochran's Q statistic (Cochran, 1954), which approximates to a χ^2 distribution, with *k-1* degrees of freedom, where k = number of component studies. Hence, if the Q statistic exceeds some critical value expected by *k-1* degrees of freedom, then the observed variance of the study is greater than that which would be expected to occur by chance (Sutton, *et al.*, 1999). Where substantial heterogeneity was found, then sources of this heterogeneity were sought. Where substantially different groups of studies were identified, then separate pooling of these individual groups was attempted, as above.

Publication bias

Where possible, funnel plots of effect size versus sample size were constructed for those studies that were judged to be sufficiently comparable. Evidence of asymmetry was sought by visual inspection of funnel plots and through the application of a statistical method outlined by (Egger, *et al.*, 1997), calculated using STATA version 6.0.

Chapter 26 Results of the review

Literature searches

Of the 19,614 individual studies identified by literature searches, 57 were felt to potentially fulfil pre-specified inclusion criteria, and full copies were obtained for further inspection. Additional studies were obtained by correspondence (Pignone, *et al.*, 2001) following the publication of an earlier version of this review (Gilbody, *et al.*, 2001d). Of these, twenty-four studies fulfilled the inclusion criteria. The flow of studies through the review is summarised in figure 13, according to guidelines laid down in the QUOROM statement on the reporting of systematic reviews and meta-analyses (Moher, *et al.*, 1999a).

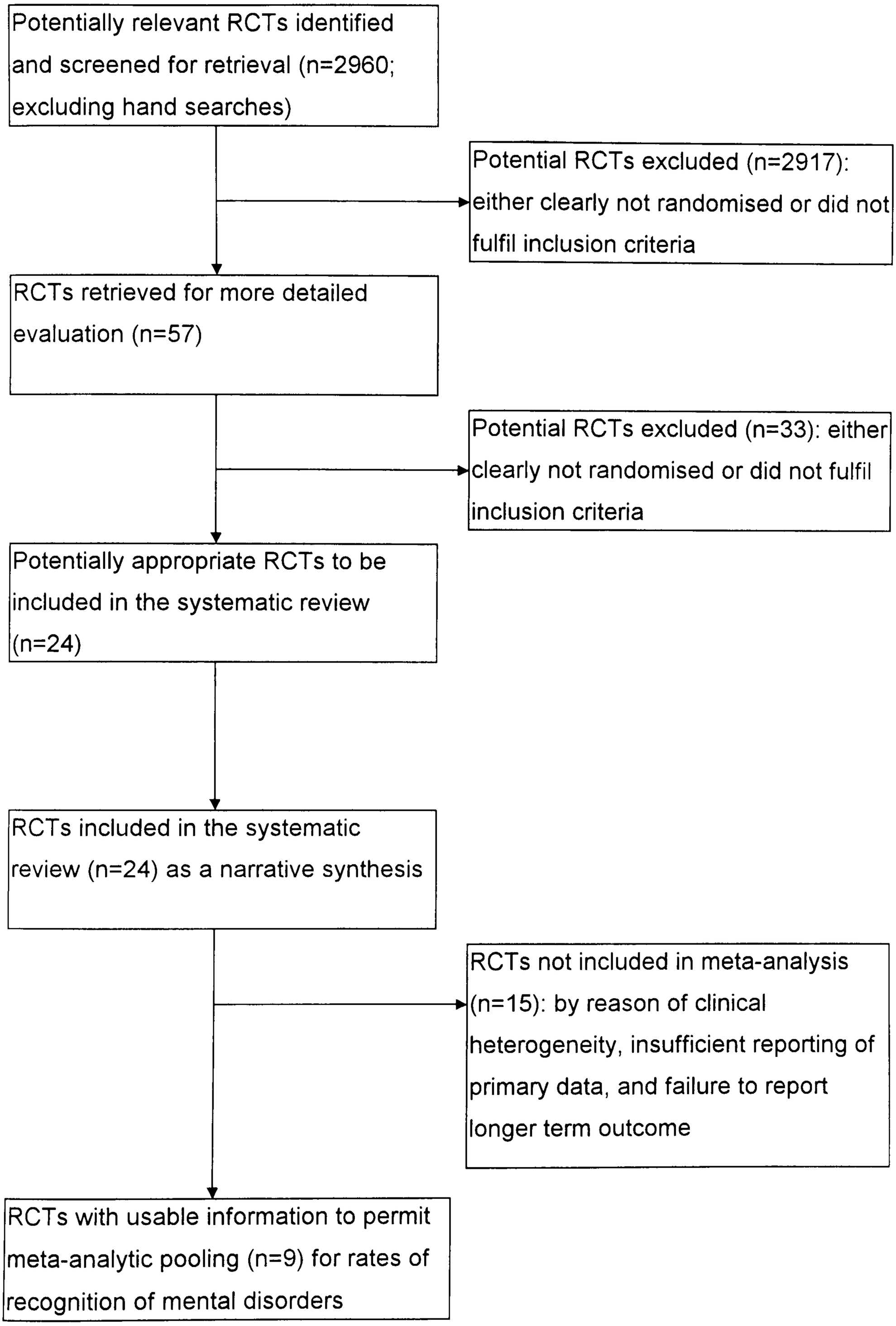
The literature searches failed to find any trials of the routine use of outcome measures in psychiatric settings. Twentyfour studies conducted in non-psychiatric settings were identified. Eleven studies were conducted in primary care settings, eight in general medical outpatients, one in general medical inpatients, one in the emergency room and one in the antenatal clinic, one in a rheumatology clinic, and one in a neurology clinic. Details of these studies are provided in Table 29 (pages 221-231).

Table 28: Utility of search strategies and databases in identifying relevant studies for the review

Database/source	Number of citations	Potentially relevant citations	Citations included in review
MEDLINE	8728	92	7
EMBASE	3270	36	2
CCTR	719	56	3
Cinahl	2160	8	0
PsycLit	4737	12	1
Reference lists and correspondence with authors	NA	24	15

Figure 13: QUOROM Trail flow diagram

(Moher, *et al.*, 1999a)



Primary care and general hospital studies

In total twenty-four randomised and pseudo randomised studies were obtained which examined the use of standardised instruments as outcomes measures in routine primary care and general hospital care settings. Specific strengths and weaknesses and facets of their design and results will now be considered in turn.

Study design & methodological quality

All studies described themselves as 'randomised', with very few giving specific details of method of randomisation and concealment of allocation. Failure to specify these facets of design are important since they have been shown to be sources of bias in randomised studies (Schulz, *et al.*, 1995). Two studies that did give details of method of randomisation, used an inadequate and not truly random allocation according to odd/even patient reference numbers (Johnstone & Goldberg, 1976), or according to alternate allocation (Weatherall, 2000). Two studies were quasi randomised, with patients seen in the first half of the study being allocated to control, and in the second period being allocated to the active intervention (Gold & Baraff, 1989; Street, *et al.*, 1994).

In the majority of studies, the unit of randomisation was the individual patient, with individual clinicians seeing both intervention and control patients (i.e. using the outcome measure for some patients and not using the outcome measure for others). This raises problems of 'cross contamination' between subjects and controls and a Hawthorne effect, whereby practice is changed for both subjects and controls by virtue of participation in a study (Roethlisberger & Dickinson, 1939). The implications of this facet of study design are explored in more detail in the discussion section. Nine studies used individual clinicians or practices as the unit of randomisation (Calkins, *et al.*, 1994; Goldsmith & Brodwick, 1989; Mazonson, *et al.*, 1994; Reilfer, *et al.*, 1996; Rubenstein, *et al.*, 1989; Rubenstein, *et al.*, 1995; Street, *et al.*, 1994; Wasson, *et al.*, 1992a; Whooley, *et al.*, 2000), so that cross contamination was avoided by single clinicians receiving either the control or experimental condition, but not both for their individual patients. None of these studies accounted for their clustering in their analysis of results, making them prone to a 'unit of analysis error' (Divine, *et al.*, 1992).

Sample size varied between 52 and 2209, and three studies (Dowrick & Buchan, 1995; Weatherall, 2000; Williams, *et al.*, 1999), included a power calculation or discussion of the sample size required to detect a specified difference in outcomes between treatment and control groups.

Setting

Eleven studies were conducted in primary care settings, eight in general medical outpatients, one in general medical inpatients, one in the emergency room and one in the antenatal clinic, one in a rheumatology clinic, and one in a neurology clinic.

Patients

There were broadly two types of patient populations who underwent randomisation: (1) 'unselected populations', where patients were included, irrespective of their baseline score on the instrument under evaluation or pre-existing probability of having some deficit or disorder as measured on the instrument under evaluation, and (2) 'high risk populations', whereby patients were only randomised if they scored above a certain level on the instrument under evaluation, or were known to have a pre-existing but unrecognised deficit or disorder such as depression. For example, an unselected population was recruited by (Hoeper, *et al.*, 1984), who administered the General Health Questionnaire (see below) to all attenders at a general practice outpatients, and randomised these patients to either have their GHQ score fed back to the clinician or to be withheld - irrespective of their score on the GHQ. Conversely, Magruder Habib, *et al.* (1990) recruited only patients with a likely pre-existing diagnosis of depression using a two stage procedure. All outpatient attenders were first given the Zung SDI (Zung, 1965), and those with high scores were then screened using a standardised diagnostic interview schedule. Only those with a confirmed diagnosis of hitherto unrecognised depression were then randomised to have their Zung SDI score fed back to clinicians in the course of the interview.

Some studies, by the nature of the services under evaluation (e.g. US veterans administration hospitals - Magruder Habib, *et al.*, 1990), included a greater proportion of elderly patients, or were specifically targeted at elderly patients (Whooley, *et al.*, 2000).

Outcome instrument used

The most commonly used instruments were self-completed scales designed to detect depression and anxiety (Beck Depression Inventory - BDI (Beck & Ward, 1961); General Health Questionnaire – GHQ (Goldberg, 1972); Zung SDI (Zung, 1965). Eight studies (Calkins, *et al.*, 1994; Goldsmith & Brodwick, 1989; Kazis, *et al.*, 1990; Rubenstein, *et al.*, 1989; Rubenstein, *et al.*, 1995; Street, *et al.*, 1994; Wagner, *et al.*, 1997; Wasson, *et al.*, 1992a) investigated the use of generic health status measures: the Short Form (SF) 36 (Ware, *et al.*, 1993); the functional status questionnaire - FSQ (Jette, *et al.*, 1986); the Dartmouth COOP (Wasson, *et al.*, 1992b); and the Sickness Impact Profile – SIP (Bergner, *et al.*, 1981). One study (Mathias, *et al.*, 1994) combined an anxiety questionnaire - the anxiety components of the Symptom Check List -90 (Derogatis, 1994; Fifer, *et al.*, 1994), with a generic health status questionnaire (the SF36) (Ware, *et al.*, 1993). Another study (Reilfer, *et al.*, 1996) used a self administered diagnostic interview schedule (Broadhead, *et al.*, 1995), which gave diagnoses for depression; generalised anxiety disorder; panic disorder; alcohol or drug abuse; obsessive-compulsive disorder; and suicidal ideation, which were then fed back to the clinician. Instruments were generally administered in the waiting room by research assistants prior to consultation.

Active intervention and choice of control

The active intervention broadly involved the feedback of instrument test results to the clinician - generally in the form of a sheet containing summary scores and an explanation of the importance of high scores in terms of the likely presence of a psychological disorder. For example, (German, *et al.*, 1987) provided summary sheets with GHQ scores together with the following statement:
'it has been shown that above a critical symptom level, a psychiatrist is likely to make a psychiatric diagnosis of a non-psychotic emotional disorder. Higher levels of GHQ scores indicate increasing probability of current emotional distress. A score higher than four is regarded an a 'positive' or abnormal result'

An alternative approach was the use of visual representations of patient problems as identified by the outcome instruments used. For example (Mazonson, *et al.*, 1994) produced a one page summary sheet, known as the 'Mental Health Patient Profile', which included summary scores of the SCL-90, highlighting elevated scores, together with visual thermometer representations of the various components of the SF36.

In some studies (e.g. Mazonson, *et al.*, 1994; Rubenstein, *et al.*, 1995), feedback of outcome results was combined with an active educational programme and the availability of standardised best practice guidelines on the management. For example, in the study by Mazonson *et al.* (1994), the active educational programme involved an educational session on the importance of deficits in health related quality of life and untreated anxiety, together with a description of the psychometric instruments and their interpretations. Results of profiles from three of their own patients were then discussed in detail and educational materials on the management of anxiety were provided in the form of audiotapes and articles. Additionally, a toll free telephone number of a study team physician was provided so that further questions could be answered.

The control condition was generally the administration of the outcome measure to the patient, without the score on this scale being fed back to the clinician. One study (Linn & Yager, 1980a) employed a factorial design that combined the above, with a discussion between the researcher and clinician in order to establish the clinicians' impression regarding the presence or absence of an emotional disorder. One study (Johnstone & Goldberg, 1976) asked the clinician about the likelihood of the presence of an emotional disorder for all patients, prior to feeding back the results of the GHQ only for those randomised to receive this information. This approach potentially increases clinician awareness of the presence of emotional problems in both intervention and control conditions.

Outcome instruments were generally administered only once in each of the studies, and were used as case finding instruments, for the purposes of identifying clinical or health related quality of life problems at an assessment interview. In most cases, the instrument was fed back to the clinician prior to the index clinical encounter, so that the clinician would be aware of the results before seeing the patient. In the study by Johnstone & Goldberg (1976), the information was fed back following the clinical encounter, and in another (Linn & Yager, 1980a), the time of feedback was varied between intervention cells, with feedback of Zung SDI results either before or following the consultation. In only four studies (Calkins, *et al.*, 1994; Kazis, *et al.*, 1990; Mazonson, *et al.*, 1994; Rubenstein, *et al.*, 1989) was the outcome battery administered sequentially during the course of care or follow-up - however this was done at fixed points by research assistants, rather than at each clinical encounter. In seven studies (Callahan, *et al.*, 1994; Dowrick & Buchan, 1995; German, *et al.*,

1987; Johnstone & Goldberg, 1976; Mathias, *et al.*, 1994; Reilfer, *et al.*, 1996; Shapiro, *et al.*, 1987), the instrument was administered on further occasions, but only as a research exercise in order to determine the outcome of the study, rather than as an intervention where the instrument was used as part of ongoing patient management (i.e. routine outcome measurement).

Trial endpoints & follow up

The most commonly collected trial endpoints were:

- The detection of depression, anxiety or an emotional problem by the clinician during the course of the clinical interview, and;
- The initiation of treatment or intervention for depression anxiety or an emotional problem.

In a number of instances, this was established by the use of clinician questionnaires or interviews following a patient consultation, whereby the clinician was asked if they believed there was an emotional disorder present (e.g. Johnstone & Goldberg, 1976). In others, it was established by case note review, whereby written evidence was sought to determine whether the clinician had noted an emotional disorder as being present, or if they had initiated any interventions for an emotional problem (e.g. Magruder Habib, *et al.*, 1990). Interventions were fairly consistently and broadly defined in studies as: referral to a mental health specialist; prescription of psychotropic medication; discussion of depression with the patient and noting the presence of depressive symptoms.

Eleven studies employed a follow up period beyond the initial consultation, which included the sequential measurement of scores on the actual outcome measure under evaluation, with follow up periods of between three and twelve months (Calkins, *et al.*, 1994; Callahan, *et al.*, 1994; Dowrick & Buchan, 1995; German, *et al.*, 1987; Johnstone & Goldberg, 1976; Kazis, *et al.*, 1990; Mathias, *et al.*, 1994; Mazonson, *et al.*, 1994; Reilfer, *et al.*, 1996; Rubenstein, *et al.*, 1989; Shapiro, *et al.*, 1987). For example (Johnstone & Goldberg, 1976) administered the GHQ to GP attenders and measured the changes in these scores at twelve months in both intervention and control groups. Similarly, Dowrick & Buchan (1995) assessed the effect of feedback of the BDI on subsequent BDI scores in both intervention and control groups. Lastly, (Kazis, *et al.*, 1990) administered the FSQ every four months

to rheumatology patients and measured endpoint scores for this measure at twelve months in both the intervention and control groups.

In one study (Street, *et al.*, 1994), the primary study endpoint was the quality of the clinical encounter and patient satisfaction with the clinical encounter following the administration of the SF36.

Study results

Effect of routine outcome measurement on recognition of emotional problems and minor psychiatric disorders

The earliest study is that by Johnstone & Goldberg (1976) which showed a large effect for the detection of depression through feedback of the GHQ, increasing the rate of detection of depression in unselected patients seen by a single general practitioner by 11%. However, this study suffers a number of problems, including inadequate randomisation, differential case ascertainment and difficulties generalising beyond the practice style of a single motivated general practitioner. Insufficient data were reported in this study to allow the reported absolute difference in the detection of depression between groups to be corroborated.

A subsequent study by Hoeper, *et al.* (1984) sought to replicate these results in sequential attenders in US primary care. No effect was found for feedback, with 16% of sequential unselected patients being found to have 'mental disorders' identified by their clinicians, irrespective of whether scores on the GHQ were fed back to the clinician or not. A subgroup analysis conducted by the authors of those with GHQ scores >4 (where this specific information and the fact that it 'indicated probable mental illness' was fed back to the clinician) showed no differential effect between controls and those receiving feedback (29% vs. 30%, relative risk of detection of depression following feedback = 1.02, 95% CI 0.81 to 1.29).

Despite being superficially similar, the studies by Johnstone & Goldberg (1976) and Hoeper, *et al.* (1984) have important differences in terms of participating clinicians, mode of feedback of outcome measure and identification of psychiatric morbidity. Johnstone & Goldberg (1976) studied the effect of feedback on 1000+ consultations with one single GP, whereas Hoeper et al includes 14 clinicians and therefore potentially reflects a wider range of practice styles. Johnstone & Goldberg, (1976)

administered the GHQ prior to the consultation and asked the clinician about the likelihood of there being an emotional disorder following the consultation. For patients allocated to the experimental group, the GHQ score was then fed back to the clinician and the clinician was then allowed to change his mind about whether there was a likely psychiatric illness. It was this final clinician diagnosis which was taken as 'case ascertainment' in the experimental group, whereas case ascertainment in the control group was by retrospective analysis of initial GHQ scores at twelve month follow up, with scores >4 defined as 'cases'. The effects of different case ascertainment methods between experimental and control conditions is potentially reflected in statistically significant differences in baseline scores on the GHQ between groups, with more severe disorders being identified in the experimental condition. In Hoeper et al's study, the fourteen clinicians received the GHQ scores in the experimental group, before making any rating of mental illness. Clinician rating of mental illness was the criterion for case ascertainment in both intervention and control conditions in Hoeper's study.

A further study of US outpatients (German, *et al.*, 1987) again shows no difference in the rate of detection of depressive illness between those who had their pre consultation GHQ fed back to clinicians and those who did not. A number of subgroup analyses were performed by the authors which suggested that the non significant results mask some potentially important increases in the rates of detection amongst those over 65 (63% vs. 41%), and amongst black and male patients. Further subgroup analysis according to GHQ score suggests that the rate of detection was increased most amongst those with moderately raised GHQ scores, rather than amongst those with high scores. This raises the possibility that the GHQ is useful in resolving clinical uncertainty amongst this group and those high scorers are detected, irrespective of whether their GHQ scores are fed back.

Linn et al's (1980) study involves a complex factorial design which allocates 150 unselected patients to one of six groups which receive either no feedback or one of five combinations of feedback before the clinical encounter, feedback after the clinical encounter and 'clinician sensitisation' to the presence of emotional problems (an interview with the researcher and discussion of the possibility of an emotional problem being present). Resultant small numbers of patients in each cell make conclusions difficult to interpret in this under-powered study, although pooling groups who received some sort of feedback and comparison with groups who received no feedback increases the rate of detection of depression (8% vs. 25%,

relative risk of detection of depression following feedback = 3.13, 95% CI 1.24 to 8.33).

One study by Williams *et al.* (1999) used a three arm intervention, comparing: (1) CES-D Questionnaire, (2) Single item question 'Have you felt depressed or sad much of the time in the past year?' and (3) usual care. The results of the first two arms were combined by the authors in all analyses and showed a non-significant positive result on the rate of recognition of depression (39% vs. 29%, relative risk of detection of depression following feedback = 1.34 95% CI = 0.79 to 2.43).

Three studies (Callahan, *et al.*, 1994; Magruder Habib, *et al.*, 1990; Moore, *et al.*, 1978) use a 'high risk' approach, targeting feedback at a selected population of primary care attenders with a probable or confirmed diagnosis of depression (Zung score >50; HDRS score >15 or diagnosis by diagnostic interview schedule). All these studies showed a positive effect for feedback.

One study by Mazonson *et al.* (1994) specifically employed routine outcome measurement and active clinician education to increase the rate of recognition and improve the outcome of anxiety in primary care. This combined intervention served to increase the rate of recognition of anxiety disorders (defined as 'chart notations') from 19% in the control arm to 32% in the intervention arm (relative risk of recognition of an anxiety disorder = 1.72, 95% CI 1.25 to 2.37).

Of the studies which employ broader measures of health related quality of life as their principle outcome measure (Calkins, *et al.*, 1994; Goldsmith & Brodwick, 1989; Kazis, *et al.*, 1990; Rubenstein, *et al.*, 1989; Rubenstein, *et al.*, 1995; Street, *et al.*, 1994; Wagner, *et al.*, 1997; Wasson, *et al.*, 1992a), four report the effect of these measures alone in improving the overall rate of recognition of emotional problems (Calkins, *et al.*, 1994; Kazis, *et al.*, 1990; Rubenstein, *et al.*, 1989; Rubenstein, *et al.*, 1995). Three of the four studies (Calkins, *et al.*, 1994; Kazis, *et al.*, 1990; Rubenstein, *et al.*, 1989) show no differences for any subscale of the FSQ or AIMS (including mental health) at 12 months. In contrast, a later study by Rubenstein *et al.* (1995) reports that feedback of the FSQ increases both the rate of recognition of depression and anxiety. Symptoms of anxiety or depression were recorded by physicians in 30% of case notes over a six month study period by clinicians receiving feedback, compared to 21% amongst those not receiving feedback of results (relative risk of detecting anxiety or depression following feedback = 1.42,

95% C. I. 0.98 to 2.08). The rate of recognition of anxiety problems was increased by the largest magnitude (13% vs. 4%, relative risk of recognition of anxiety following feedback = 3.33, 95% C. I. 1.40 to 7.92), whilst the rate of recognition of depression was subject to a non significant increase in recognition (23% vs. 20%, relative risk of recognition of depression following feedback = 1.17, 95% C. I. 0.78 to 1.77). The major limitation of this study is, however, the fact that whilst it is a cluster randomised trial (clinicians are the unit of randomisation), it is analysed according to individual patients without reference to intra-class correlation coefficients. It is therefore subject to a unit of analysis error and the chance of a type 1 error cannot be excluded.

Statistical pooling of studies intended to increase the detection of depression.

Several studies involved sufficiently similar interventions and endpoints to allow the possibility of a quantitative synthesis of study outcome to be examined (Callahan, *et al.*, 1994; German, *et al.*, 1987; Gold & Baraff, 1989; Hoeper, *et al.*, 1984; Johnstone & Goldberg, 1976; Linn & Yager, 1980a; Magruder Habib, *et al.*, 1990; Moore, *et al.*, 1978; Whooley, *et al.*, 2000; Williams, *et al.*, 1999; Zung, *et al.*, 1983).

Two studies (Johnstone & Goldberg, 1976; Zung, *et al.*, 1983) provide insufficient raw data to allow the size of the reported result to be confirmed or to be entered in a formal meta-analysis. Another study (Linn & Yager, 1980a) provides data on six separate arms of a trial, each with a different variant on time and mode of feedback of outcomes data to clinicians. Potential inclusion of this study was not felt to be justified. One further study by Williams *et al.* (1999) used a three arm intervention, comparing: (1) CES-D Questionnaire, (2) Single item question 'Have you felt depressed or sad much of the time in the past year?' and (3) usual care, and the pooling of essentially different interventions was not felt justified. The study by Whooley *et al.* (2000) followed up only those patients who screened positive for depression, rather than all those randomised, making the effect of feedback on the whole study population impossible to assess. The study by Gold & Baraff (1989) was a non-randomised study, and its inclusion in the presence of randomised data was not felt to be justified. The justification for the exclusion of these studies and the effect of their reintroduction of potentially useable data on the overall result of the meta-analysis is examined below in a sensitivity analysis and examination of sources of heterogeneity.

Visual inspection of a Forrest plot for those remaining studies shows substantial between study variations between studies (see figure 14). Evidence of between study heterogeneity is further suggested by the application of statistical tests for heterogeneity (Q 'non-combinability' for relative risk = 23.4; df = 4; p = 0.0001). Of note is the observation that larger studies produce non-significant results (German, *et al.*, 1987; Hoeper, *et al.*, 1984), whereas smaller size studies produce more marked effect sizes in favour of feedback. The differential effect according to sample size is confirmed by Funnel plot analysis (Figure 15), where the funnel is found to be substantially asymmetrical (p=0.024 using Egger's test (Egger, *et al.*, 1997).

Figure 14: Forrest plot for studies examining the effect of feedback on the rate of recognition of depression

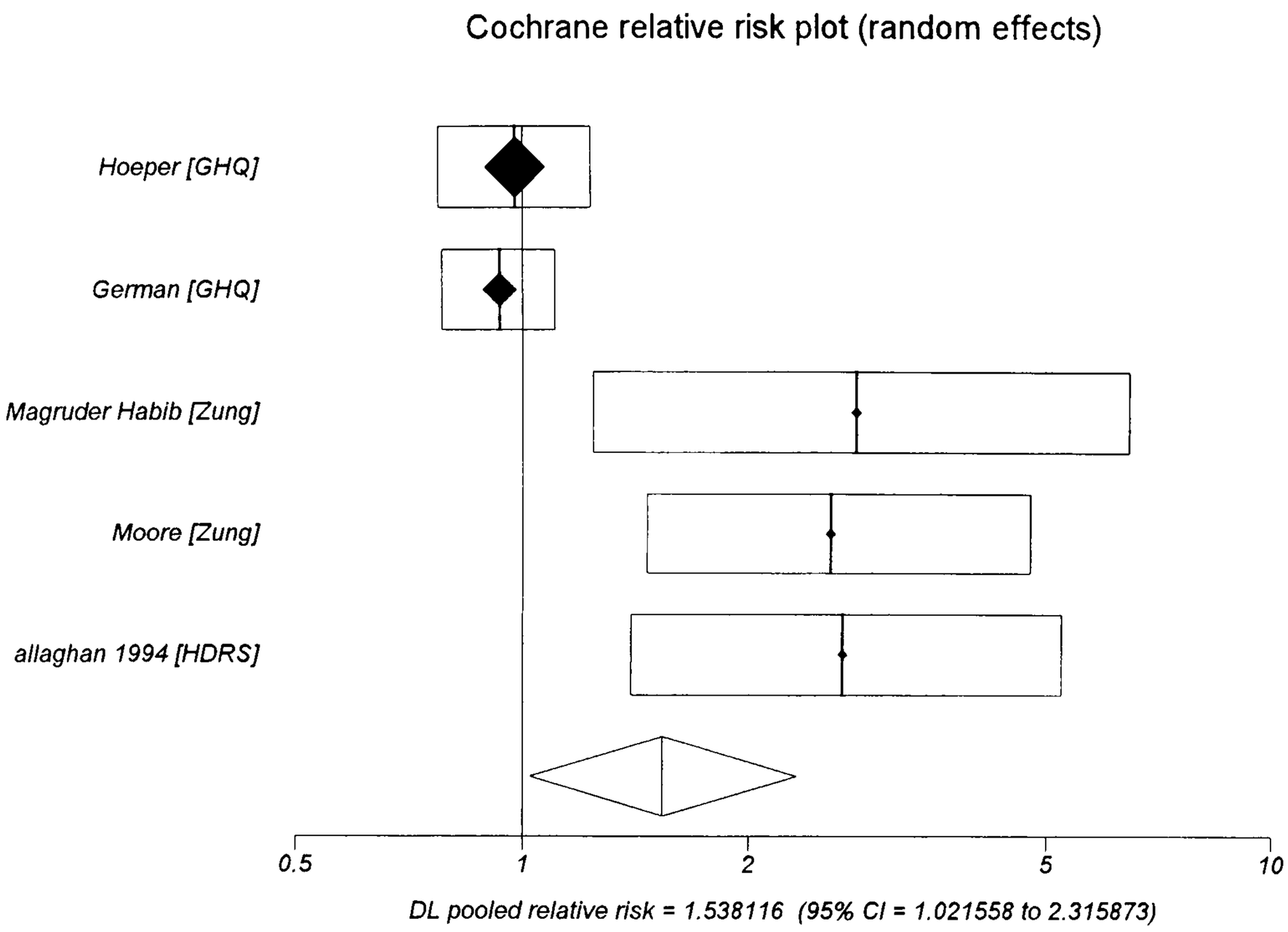
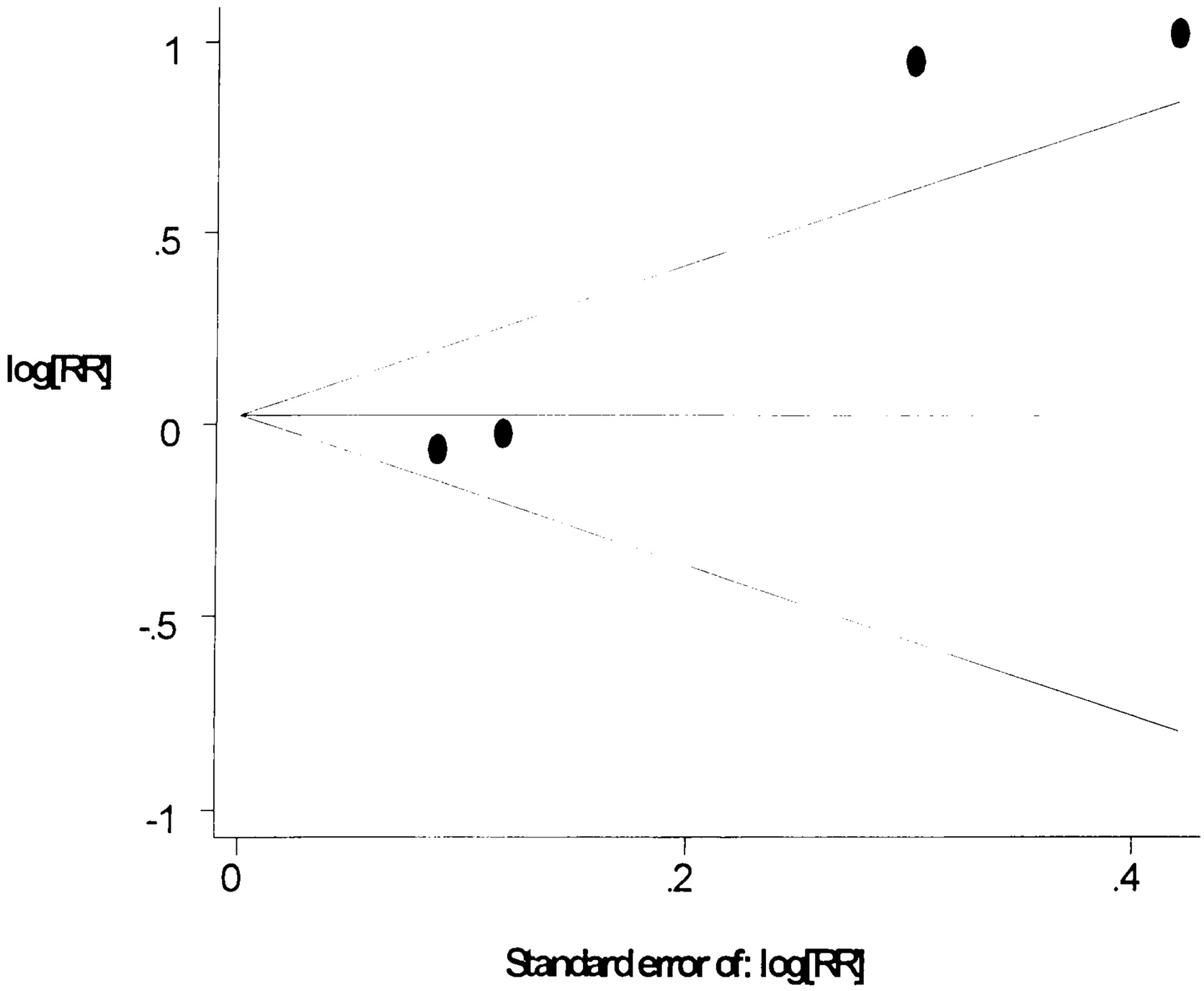


Figure 15: Funnel graph of studies examining the effect of feedback on the rate of recognition of depression



It was noted in the previous discussion that broadly two different strategies were employed in the provision of feedback in the trials included in this review: Firstly, an *unselected* form of feedback, whereby outcomes measures were administered to all patients seen in a clinical service, and their results fed back to the clinician, irrespective of their score. Secondly, a *high risk* form of feedback, whereby outcomes measures were administered, and only those with high scores were then included in a randomised trial to have their high scores fed back to the clinician or not. Examination of the Forrest plot (Figure 14) and funnel plot (Figure 15) shows that the larger of the four trials, with largely negative results employ an unselected strategy, whereas the two smaller trials with positive results employ a high risk strategy. There are plausible reasons why these two forms of feedback are likely to have fundamentally different effects in routine practice, since clinicians are potentially more likely to act on the results of positive results, when only these are fed back. These differential effects are a likely explanation of the heterogeneity shown in Figure 14. For this reason two separate meta-analyses were undertaken for unselected and high-risk studies.

Meta-analysis of unselected feedback studies

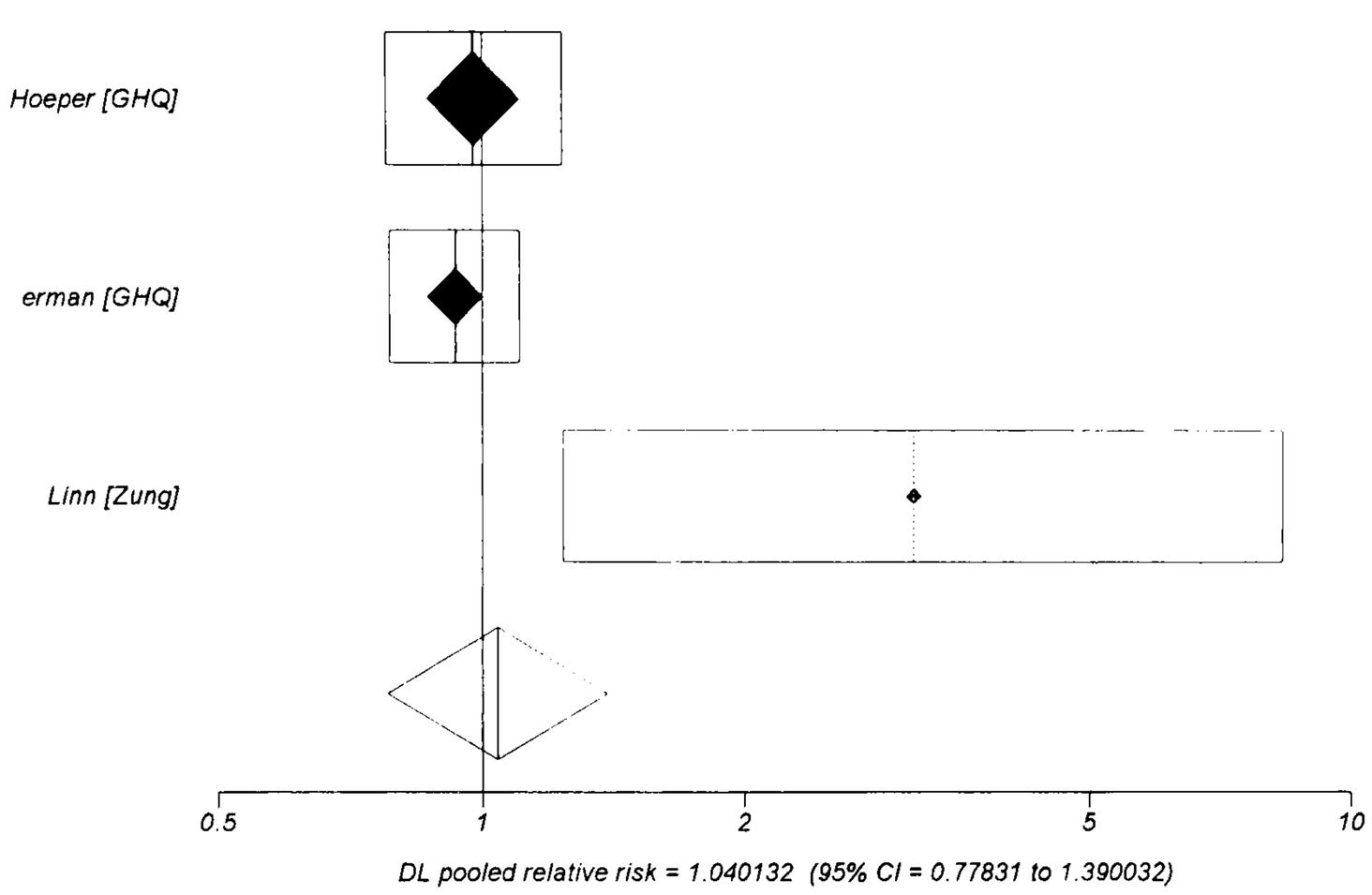
Meta-analytic pooling of the two studies by German *et al.* (1987) and Hoeper *et al.* (1984) suggests that unselected feedback is ineffective in increasing the rate of recognition of depression (DerSimonian-Laird pooled relative risk of detection of depression = 0.947, 95% CI = 0.825 to 1.088), and that there is homogeneity in the results of these two studies (Q = 0.109, df = 1, P = 0.74).

One further study (Linn & Yager, 1980a) also used an unselected approach, but was excluded from the main analysis due to the questionable validity of pooling five separate arms, which each used a different variation of the timing and mode of feedback of questionnaire results. This study was reintroduced into the preceding meta-analysis in order to test the robustness of the overall result to the inclusion of this positive study (figure 16). It was found that this study increased the level of heterogeneity within the analysis (Q = 5.59, df = 2, p = 0.0612), but that the overall negative result was robust to the inclusion of this study (DerSimonian-Laird pooled relative risk = 1.04, 95% CI = 0.78 to 1.39).

Similarly, the small sized non-randomised study by (Gold & Baraff, 1989) used an unselected approach. The introduction of this study did not alter the overall

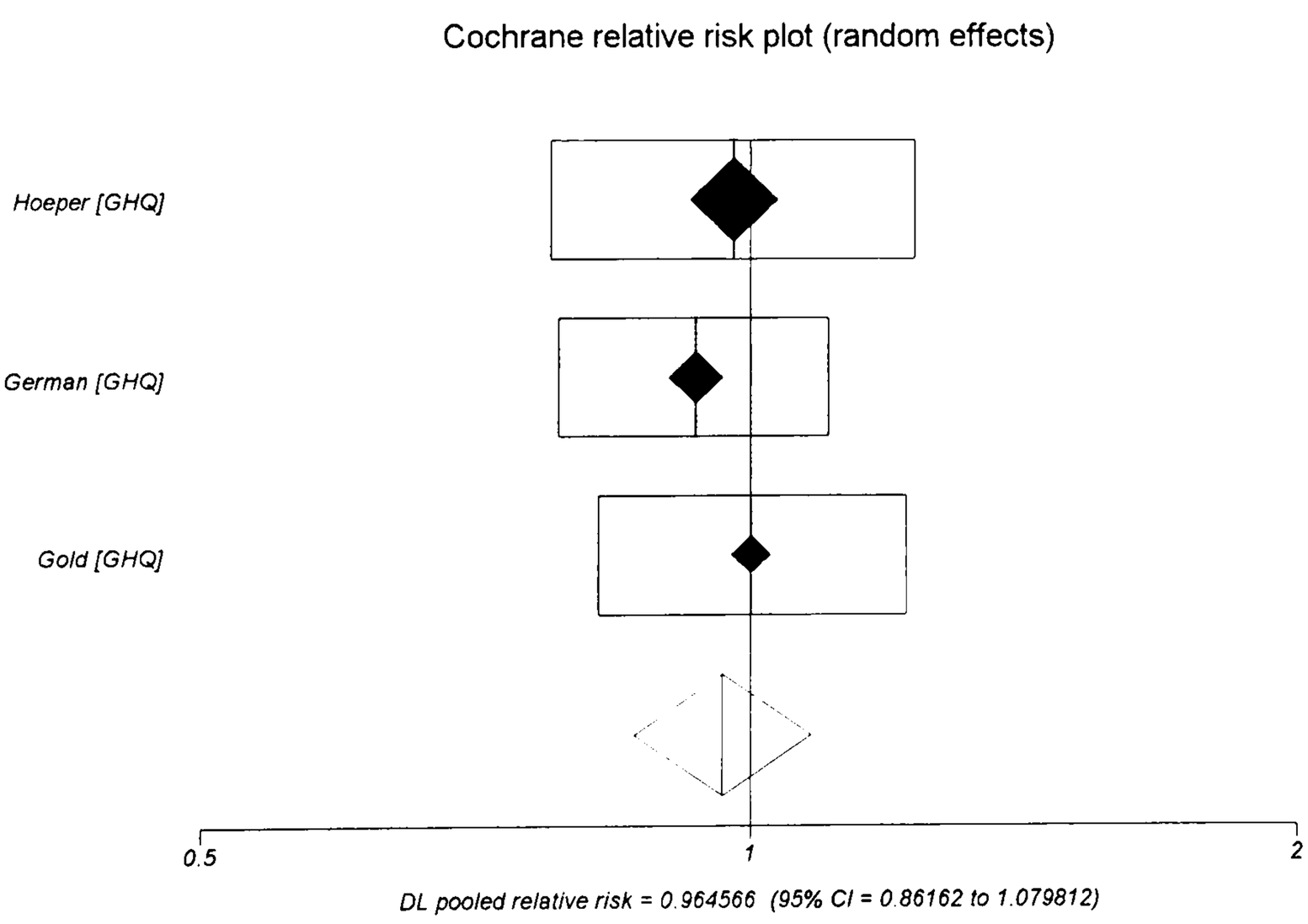
conclusion of the meta-analysis (figure 17) - DerSimonian-Laird pooled relative risk = 0.97, 95% CI = 0.86 to 1.08 (Q = 0.298, df = 2, p = 0.861).

Figure 16: Meta-analysis of studies employing unselected feedback, with the inclusion of Linn et al, as a sensitivity analysis



Q ("non-combinability" for relative risk) = 5.586 (df = 2) P = 0.0612.

Figure 17: Meta-analysis of studies employing unselected feedback, with the inclusion of Gold et al, 1989, as a sensitivity analysis

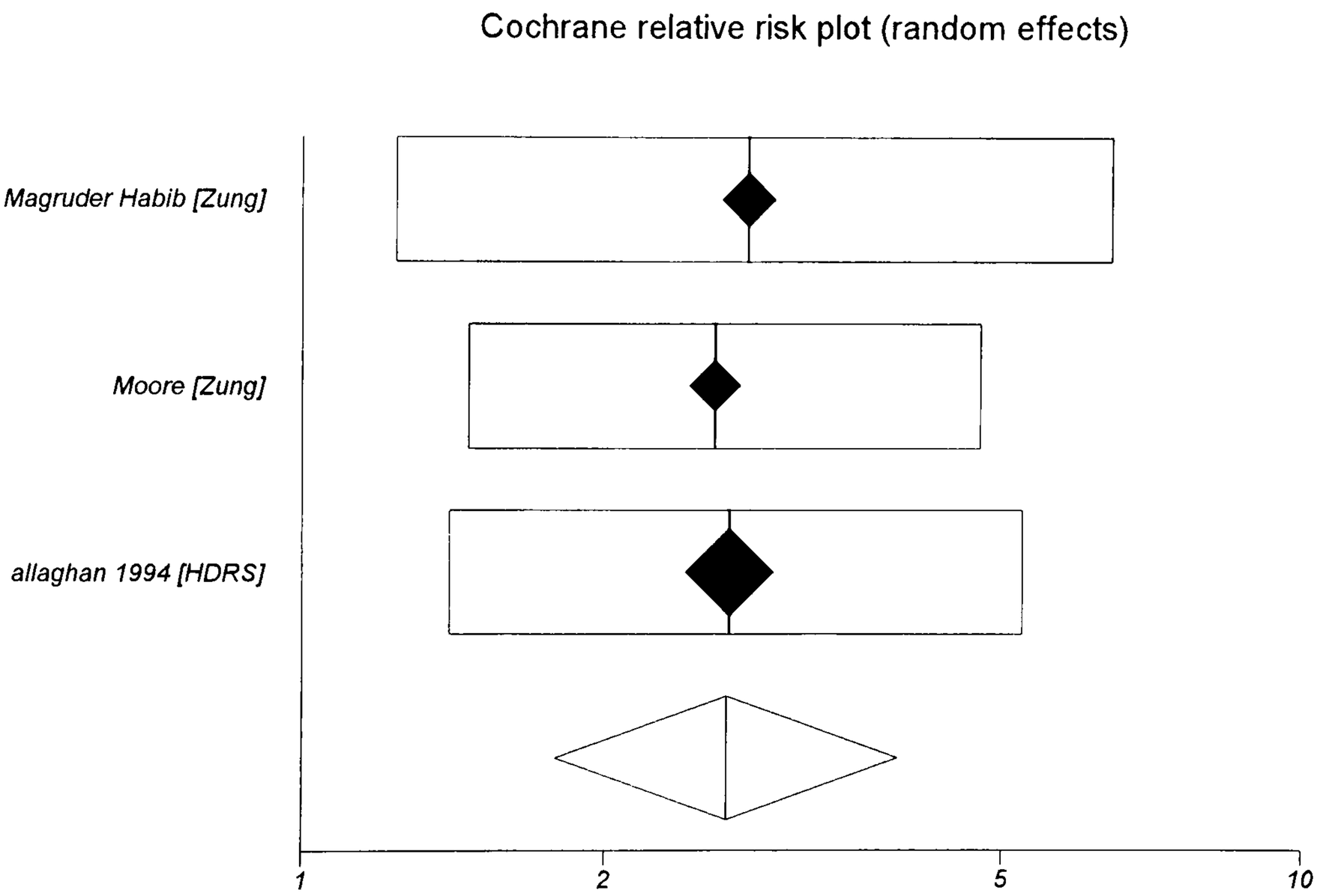


Q ("non-combinability" for relative risk) = 0.298 (df = 2) P = 0.861

Meta-analysis of high risk feedback studies

Meta-analysis of the three studies by Moore *et al.* (1978), Magruder Habib *et al.* (1990) and Callahan *et al.* (1994) show that this high risk strategy was a largely effective in increasing the rate of recognition of depression (DerSimonian-Laird pooled relative risk = 2.641, 95% CI = 1.78 to 3.94, Q = 0.02, df = 1, p = 0.887) – see figure 18. This intervention increased the rate of detection of depression by 27% (DerSimonian-Laird pooled risk difference = 0.270, 95% CI = 0.144 to 0.397), with an equivalent number needed to treat (NNT) of 4 (95%CI 3 to 7) suggesting that the results of four high scoring questionnaires need to be presented to clinicians in order that one extra case of depression is detected.

Figure 18: Meta-analysis of studies employing *high-risk* feedback



Q ("non-combinability" for relative risk) = 0.020 (df = 1) P = 0.887.

Effect of routine outcome measurement on initiation of treatment for emotional problems

Nine studies investigated the effect the feedback of questionnaire results on the rate of intervention for emotional problems (Callahan, *et al.*, 1994; Dorwick, 1995; German, *et al.*, 1987; Lewis, *et al.*, 1996; Linn & Yager, 1980a; Magruder Habib, *et al.*, 1990; Mazonson, *et al.*, 1994; Weatherall, 2000; Whooley, *et al.*, 2000; Williams, *et al.*, 1999) and all but two (Callahan, *et al.*, 1994; Magruder Habib, *et al.*, 1990) showed non significant results. Heterogeneity of methods and definition of an active intervention meant that overall pooling was not justified.

Interestingly, whilst Linn et al (1980) showed that feedback increased the rate at which clinicians recognised depression, but the likelihood of making an intervention (judged from case note review) was not altered (RR 0.93, 95% CI 0.83 - 1.05). German, *et al.* (1987) similarly showed no effect of feedback for all patients on management (RR=1.02, 95% CI 0.93 - 1.13). The subgroup analyses carried out by German, *et al.* (1987), which suggested a greater recognition of depression amongst the elderly, men and blacks when feedback is received, did not show any increase in the rate of intervention amongst these groups.

The study by Mazonson, *et al.* (1994), which specifically targeted the recognition and intervention for anxiety showed a marked increase in the rate of mental health referrals (10% vs. 3%, relative risk of outside referral for an anxiety problem = 2.94, 95% CI 1.33 to 6.51). This increased rate of intervention was not accompanied by an increased rate of initiation of psychotropic prescriptions (13% vs. 13%).

Effect of routine outcome measurement on subsequent outcome of emotional disorders

Eleven studies examined the effect of outcome measurement on the actual outcome of the patient over time (Callahan, *et al.*, 1994; Dowrick & Buchan, 1995; Johnstone & Goldberg, 1976; Kazis, *et al.*, 1990; Lewis, *et al.*, 1996; Mazonson, *et al.*, 1994; Reilfer, *et al.*, 1996; Rubenstein, *et al.*, 1989; Rubenstein, *et al.*, 1995; Whooley, *et al.*, 2000; Williams, *et al.*, 1999). Results from Johnstone & Goldberg (1976), using retrospective patient recall, showed that patients with hidden psychiatric morbidity, on whom GHQ feedback was given, have a shorter duration of illness (2.8 months vs. 5.3 months). Final 12 month GHQ scores of patients found to be positive at their index episode were broadly similar for those on whom feedback was given compared to controls. However, a subgroup analysis suggests that feedback was associated with improved GHQ scores amongst those with a 'severe' but unrecognised disorder at inception.

No overall effect of outcome measurement on outcome was detected in nine of the eleven studies. For example, the study by Dowrick & Buchan (1995), who re-administered the Beck Depression Inventory at 12 months and found there to be no significant difference between those in whose scores were fed back and controls. This study suggests that unrecognised depressive symptoms resolve over a twelve month period, irrespective of whether feedback was employed or not. Similarly, Lewis, *et al.* (1996) show a lack of overall effect of GHQ feedback on subsequent GHQ scores.

Of the two studies that showed a positive effect of routine outcomes measurement, the study by Mazonson *et al.* (1994) involving an intensive educational and feedback intervention targeted at anxiety problems found no overall improvement in either total scores on the anxiety components of the SCL-90, nor the mental health component of the SF36. The only positive effect that was found in this study was on a self report scale of anxiety, used in conjunction with the SF36 and the SCL-90. The other positive study by Rubenstein *et al.* (1995) resulted in a small, but statistically significant change in the mental health component of the FSQ (endpoint mean change difference = 4.5 points, 95%CI 0.5-8.3, on a 100 point scale). Of the four component scales of the FSQ (activities of daily living; mental health; social activities; work performance), mental health was the only scale to show a between group difference at the end of a six month study period. As mentioned previously,

this cluster-randomised trial was prone to a unit of analysis error and the possibility of a spurious positive result cannot be excluded.

Effect of routine outcome measurement on consulting behaviour

Johnstone & Goldberg (1976) examined the effect of feedback of outcome data on subsequent GP consultation over 12 months and found that the increased rate of recognition of depression and improved outcome was not followed by an increased number of consultations with their general practitioner. There had, however, been a change in the pattern of consultation behaviour. Feedback had increased the proportion of consultations that had been labelled 'psychiatric' in their content by the general practitioner. This overall trend is replicated by the more recent and rigorous study by Lewis *et al.* (1996), which also showed that rates of psychiatric and non-psychiatric referrals were unchanged as a result of feedback.

The study by Mazonson *et al.* (1994) reported brief data on non-mental health utilisation and consulting behaviour. There was no difference in the rate of non-psychiatric hospitalisations between feedback and control groups (9% vs. 10%), however there was an average of 0.6 more primary care visits (for any reason) between feedback and control groups. (3.3 visits over six months vs. 2.7 visits, p=0.054).

In contrast the study by Reilfer *et al.* (1996) showed a reduction in health utilisation in the intervention group (referrals to non mental health specialists reduced 0.9 vs 2.1 visits, $p<0.005$).

Effect of routine outcome measurement on patient satisfaction with care and patient - doctor communication

The study by Street *et al.* (1994) examined the effect of the administration and feedback of the generic health status questionnaire, the SF36 on the patient satisfaction and communication in the ante-natal clinic.

Their patient survey showed that patients generally wanted to be asked about 'health status overall', and listed the components of the health status which they wanted to be asked about. All patients wanted to be asked about 'pain' and 'perceptions of health', fewer expressed a preference to be asked about 'social

functioning' and 'mental health' (<70%). The administration of the SF36 increased the patients' satisfaction with care, but feedback of these instruments did not affect the degree to which physicians were perceived as having asked about 'health status overall'. No data were presented to examine the degree to which feedback of SF36 results increased the degree to which mental health problems were discussed or detected.

Another study, by Reilfer *et al.* (1996) showed no change in either clinician or patient satisfaction with care following the administration and feedback of the diagnostic interview schedule. Similarly, the study by Williams *et al.* (1999) showed no effect on patient satisfaction with the care they received, although clinicians who received feedback, generally said that they had found the information useful (although no direct comparison with control physicians was possible).

The study by Mazonson *et al.* (1994) included a patient interview amongst those who received treatment for anxiety. Feedback seemed to increase the tendency of clinicians to be more proactive in raising the problem of anxiety and need for treatment. Amongst those who had their scores fedback and received treatment, 67% reported that their physicians had been proactive in initiating treatment, whereas amongst those whose scores were not fed back, only 33% reported that the physicians had taken the first step in suggesting treatment.

Other outcomes

No study examined the costs and resource use associated with routine outcome measurement. No study examined patients' views about the usefulness or acceptability of standardised instruments for detecting psychiatric disorders.

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Callahan *et al.* (1994)	RCT Individual patients randomised	Elderly US primary care patients with a score above 15 on the Hamilton Depression Rating Scale (HDRS) N=175	HDRS	Int.: Three additional appointments made over a three-month period with the primary care physician. Clinicians provided with written patient specific materials, including HDRS scores; an interpretation of their meaning; a list of all medications and a specific instruction that drugs causing depression should be reviewed; and a written instruction that the presence of depression should be examined and managed appropriately – clinical algorithm provided. (n=100) Cont.: No written feedback and no extra visits scheduled (n=75)	Diagnoses of depression. Discontinuation of drugs causing depression. Initiation of antidepressants. Psychiatric referrals. Depression scores Functional status scores (Symptom Impact Profile – SIP) Follow up at six months	Increased diagnosis of depression in Int. group (int. 32/100 Vs cont. 9/75) More frequent discontinuation of depressant drugs (int. 23/100 vs cont. 17/75) Increased rate of antidepressants in Int group (int. 26/100 vs cont. 6/75) No difference in rate of psychiatric referrals (int. 12/100 vs cont. 10/75) No difference in HDRS scores at six months No difference in SIP scores between groups.
Dowrick & Buchan (1995)	RCT Individual patients randomised	Consecutive GP attenders (n=116) in Liverpool, UK, with depression score above 14 on the BDI.	Beck depression Inventory (BDI)	Int.: BDI administered pre consultation and depression scores disclosed to GP (n=52). Cont. 1: BDI administered, but not fed back to GP (n=64).	Diagnoses of depression and BDI scores at 6 & 12 months	Disclosure had no discernible effect on BDI scores

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
German, *et al.*, (1987) & Shapiro, *et al.*, (1987)	RCT Individual patients randomised	US adult & elderly general medical outpatient attenders (n=1242). Separate interventions for high (n=488) and low (n=754) GHQ scorers.	GHQ (administered by a research assistant)	Int.: GHQ administered pre consultation and results fed back to clinician, together with an indication that score was high and suggested 'psychiatric diagnosis'. (n=165) Cont.: GHQ administered, but not fed back. (n=323)	Detection of depression by clinicians. *Presence of depression according to diagnostic interview (DIS).* Treatment initiated for depression. GHQ scores at six months	No difference in detection rate amongst under 65s (int 57% vs cont 58%). Greater detection of depression in over 65s with feedback (int 63% vs cont 43%). No differences in management of depression in under 65s (46% vs 46%), but greater proportion of over 65s receiving intervention following feedback (42% vs 32%). GHQ scores at six months not reported.
Gold & Baraff (1989)	Pseudo-RCT	US emergency department attenders. Patients with existing or recognised psychiatric disorders excluded	GHQ	Int.: 28 item GHQ administered to 357 patients and results fed back to emergency physicians. Cont.: GHQ administered to 242 patients, but not fed back.	Psychiatric diagnosis made by clinician Psycho-social referrals made.	No overall improved recognition of psychiatric illness (40% vs. 40%). Moderately increased rate of recognition of psychiatric disorders for only those patients with GHQ>10 (57% vs. 66%) Increased rate of psychosocial referrals following feedback (23% vs. 5%).
Hoeper *et al.* (1984)	RCT Individual patients randomised	Adult US primary care patients (n=2309)	GHQ	Int.: GHQ administered by researcher and scores fed back to clinician, with information that a score >5 indicated mental illness. Cont.:GHQ administered, but not fed back to clinicians.	Physician diagnoses of mental illness at reference visit (info elicited as part of the study)	No difference in rate of detection of mental disorders (Int = 16.0% vs Cont. = 16.8%) No difference in rate of detection amongst those with high GHQ scores (int = 30% vs cont = 29%)

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Johnstone & Goldberg (1976)	RCT Individual patients randomised Odd/even allocation	Sequential attenders to a single UK general practitioner (n=1093). Those with psychiatric morbidity (GHQ>5) which had not been hitherto recognised by the GP (Hidden Psychiatric Morbidity) followed up.	GHQ	Int.: GHQ administered and clinician asked about likelihood of psychiatric morbidity. GHQ then fed back to clinician. Those with unrecognised depression and high scores at initial interview (hidden psychiatric morbidity) followed up (n=60). Cont.: GHQ administered and clinician asked about the likelihood of psychiatric morbidity. GHQ folded and placed in the patient note envelope. Those with unrecognised depression and high scores at initial interview (hidden psychiatric morbidity) followed up (n=59).	*For those with hidden psychiatric morbidity, the following were studied:* Diagnosis & severity of depression during 12 months follow up (incl GHQ scores). Length of depressive episodes. Pattern of consultation over 12 months	GHQ feedback increases the rate of detection of hidden psychiatric morbidity by 11% & reduces length of illness. Feedback of GHQ facilitates a more psychological, rather than somatic, pattern of consulting. No difference in overall GHQ scores at 12 months. Subgroup analysis according to initial GHQ score shows that high scorers on GHQ benefit preferentially form feedback. Low scores resolve spontaneously, irrespective of feedback.
Linn & Yager (1980a) & Linn & Yager, (1980b)	RCT Individual patients randomised	New referrals to US medical outpatients (n=150) - mean age 56	Zung self rating depression scale (SDS)	Int 1.: SDS administered prior to consultation and results placed at front of notes, together noramative values. Physician also asked about depression post consultation. Int 2.: SDS fed back to clinician following consultation. *Int 3.: SDS provided pre consultation, but clinician's impression of depression not elicited.* Int 4.: SDS given to clinican following consultation, no impression of depression sought. Int 5.: no screening by SDS, but impression of depression sought. Cont.: no screeing by SDS, no physicaain opinion sought.	Depression noted in charts Initiation of treatment for depression	Depression is generally under recognised. Screening and feedback of SDS increased the frequency of notation of depression (8 vs 25%). Increased notation of depression occurs irrespective of the time of feedback (pre or post consultation). Sensitisation to depression has no effect. Screening has a much smaller effect on the initiation of treatment for 'depression'.

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Lewis *et al.* (1996)	RCT Individual patients randomised	UK General practice attenders at a single practice with GHQ-12 score >2.	GHQ-12 & computerised assessment of psychiatric symptomatology	Int1.: GHQ administered & placed in notes with no interpretation or instruction on the presence of mental disorder (n=227 patients). Int 2.: Patient asked to complete a computerised assessment and the results of this assessment fed back to the clinician (n=227 patients. Cont..: No feedback given (n=227 patients). NB. A random sample of 200 patients with GHQ<2 had their GHQ results also placed in the notes, so that GPs would be blind to the presence of likely psychiatric disorder in Int 1 & 2.	Consultation rates & clinician attribution of encounters as due to psychological or physical problems Prescription of a pscyhotropic drug Rates of outside mental health referrals to outside agencies GHQ scores at 6 weeks, 3 & 6 months.	No differences in consultation rates, but more identified as 'psychological' for GHQ group (p=0.09). No differences in the rate of pscyhotropic prescriptions. No differences in the rate of referral to outside agencies. Moderate improvement (5% 95% CI –3 to 14%) in GHQ scores at six weeks for computerised feedback. No between group differences over longer term.
Magruder Habib *et al.* (1990)	RCT Individual patients randomised	Male adult US veterans (mean age 60) attending a US general internal medicine OP clinic with Zung SDS score >50	Zung self rating for depression scale (SDS)	Int: SDS administered and fed back to physicians at first clinic assessment visit - placed at front of clinic notes (n=48) Cont.: SDS administered but not fed back to clinicians (n=52)	Recognition of depression Initiation of management of depression Scores on SDS at 3, 6, 9 & 12 months	Greater recognition of depression in intervention group (56% vs 35% @ 12 months). More frequent intervention in feedback group (56%vs 42% @ 12 months). Feedback facilitated recognition for those with a high somatic score on SDS subscale.

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Mathias *et al.* (1994) & Mazonson *et al.* (1994)	RCT Primary care group practices randomised	US Primary care patients with hitherto unrecognised anxiety.	SCL-90 (anxiety sub-scales only) SF36	Int: Physicians (n=40) given an educational package which included teaching sessions on the importance and causes of anxiety problems. These received structured feedback of anxiety scores (SCL-90) and functional status (SF36) scores from n=357 patents. Feedback was given at consultation, at two further points in the follow up (11 weeks and 5 months) Cont.: Physicians (n=35) received no feedback from n=216 patients who had completed the SCL and SF36 questionnaires.	Recognition and treatment for anxiety problems. Changes in anxiety scores at 3 and 5 months Changes in SF36 scores at 3 and 5 months Self reported global improvement in anxiety and functional status	Increased recognition and treatment for anxiety symptoms (35.6% vs. 20.8% p<0.001). Increased referral to mental health sector (9.5% vs. 3.2%, p<0.001), but no difference in the prescription of pscyhotropics. No differences in change for anxiety scores (p=0.89) No differences in change for SF36 (total and mental health scores) Self reported global anxiety and functional status both improved in intervention group (46.3% vs. 37.0% report improvement for anxiety).
Moore *et al.* (1978)	RCT Individual patients randomised	General Practice attenders with SDS scores >50	Zung self rating depression scale (SDS)	Int.: SDS administered and score fed back ('mildly' or 'severely depressed') Cont.: SDS administered, but no feedback to clinician	Notation of depression following index visit	Feedback increased recognition of depression for high risk patients (22% vs. 56%)
Reilfer *et al.* (1996)	RCT Internal medicine firms randomised	Randomly selected patients attending a US urban internal medicine clinic (n=358)	Diagnostic interview schedules (16 item Symptom Driven Diagnostic Interview Schedule)	Int.: Patients (n=185) given screening questionnaire. Results of diagnostic codes elicited (depression; generalised anxiety disorder; panic disorder; alcohol or drug abuse; obsessive-compulsive disorder; suicidal ideation) and fed back to the clinician prior to the clinical encounter. Cont.: Questionnaire administered to patients (n=172), but results not fed back.	Functional status at 3 months – using the Short Form 36. Zung self rated depression and Sheehan anxiety scores at 3 months for those screened positive for depression. Health care utilisation over 3 months Satisfaction with care.	65% of all patients screened positive for at least one disorder. No statistical difference in SF36 scores. No statistical difference in Zung depression scores No statistical difference in Anxiety scores. Reduction on health utilisation in Int. group (referrals to non mental health specialists reduced 0.9 vs 2.1 visits, p<0.005). No change in patient satisfaction with care. NB Clustering not accounted for in the analysis of the data

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Weatherall (2000)	Pseudo RCT (odd even allocation) of individual patients	Elderly inpatients, in New Zealand (n=100)	Geriatric depression rating scale	Int: GDS administered, together with the Mini Mental State Examination. Scores written in the notes (by hand) and an interpretation of the significance of scores given Cont: An Activity of Daily living questionnaire administered in place of the GDS.	Rate of prescription of antidepressants Follow up at discharge and three months	No difference in rate of antidepressant prescription (int 6/46 vs cont 3/47 RR = 1.4; 95% CI = 0.72 to 2.09)
Williams *et al.* (1999)	RCT Individual patients randomised	Sequential attenders at a US family medicine clinic (n=969)	CES-D Questionnaire or Single item question 'Have you felt depressed or sad much of the time in the past year?'	Int 1: CES-D self administered, scored by researcher and results fed back to clinicians as either 'positive' or negative'. N=323 Int 2.: Single item question asked and answer yes or no fed back to clinician. N=330 Cont.: Usual care. N=316 NB all clinicians were given a copy of the 'Quick reference guide for clinicians on the management of depression (Depression Guideline Panel, 1993)	Sensitivity and specificity of the instruments. Recognition of depression from case note review – corroborated by DSM-III-R interview schedule. Severity of depression from DSM-III-R symptom counts. Treatment for depression (referral, antidepressants Patient and physician satisfaction with care and use of questionnaires Functional status from the SF36	CES-D sensitivity = 88% & specificity 75% Single item questionnaire sensitivity = 85% & specificity = 66% Interventions 1 and 2 were combined in the reported analysis making the effects difficult to interpret further. Authors report: Increased rate of recognition of depression (int. 30/77 vs cont 11/38; RR 1.34 95% CI = 0.79 to 2.43) No difference in rate of intervention – outside referral or antidepressant prescription (exact figures not given) No difference in prevalence of depression at three months

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Whooley *et al.* (2000)	RCT Primary care clinics randomised	Sequential US family practice attenders over 65 years (n=2,346)	Geriatric Depression Scale (GDS) administered by a research assistant	Int: GDS administered, and scored by research assistant. Scores fed back to physicians, with an indication that the score suggested moderate (score 6-10) or severe (11+) depression. In additsion, clinic attenders screened positive were offered a series of organised educational sessions. Cont: GDS admisitered, but scores not fed back. Educational sessions not offered (usual care).	Physician diagnosis of depression (case note review, by blinded researcher). Prescription of antidepressants. Healthcare ustilisation (number of clincia visits and hospitailisations) Depression scores of the GDS. Outcomes all measured at two years. NB only those with screen positive depression followed up (n=331)	Baseline prevalence of depression 14.1% (GDS >5). *No difference in detection of depression (Int 56/162 vs cont 58/169 RR = 1.00 95% CI 0.79 to 1.26).* No difference in the rate of prescription of antidepressants (int 59/162 vs cont 72/169; RR = 0.87 95% CI 0.69 to 1.09) *No difference in mean number of clinic visits (p=0.5) or hospitalisation (p=0.8).* No significant between group difference in GDS scores at two years (based upon 69% follow up). Proportion of participants with GDS>5 - int 41/97 vs cont 54/109 (RR 0.85 95% CI 0.63 to 1.14) NB clustering not accounted for in analysis
Zung *et al.* (1983)	RCT Individuals patients randomised	US patients with undetected depression attending a family medicine centre (n=143)	Zung self rating for depression scale (SDS)	Int.: Patients' (n=102) SDS results attached to the front of the medical record and the clinician verbally informed of the positive result and asked to evaluate the patient carefully for the presence of depressive disorder. Cont.: Patients' (n=41) SDS results not fed back to the clinician.	Notation of depression in the medical notes. SDS scores at 4 weeks & clinical improvement (operationally defined as a decrease of at least 12 points from baseline.	Increased notation of depression in charts for identified group (15% vs 68%). Direct comparisons of SDS scores between Intervention and Control groups not possible due to incomplete reporting of the data.

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Calkins *et al.* (1994)	RCT Physicians randomised	60 US general hospital physicians, with eight patients for each physician randomly selected (497 patients)	Functional Status Questionnaire (FSQ)	Int.: Physicians given a seminar on the importance of FSQ test results. FSQ administered to their patients every four months, and results included in the patients records. Cont.: FSQ administered as above, with no physician training and no report feedback	Six summary scales of the FSQ (activities of daily living; mental health; work performance; social activity; quality of interaction), measured at four, eight and twelve months	No significant difference on any subscale, including mental health.
Goldsmith & Brodwick (1989)	RCT Clinicians randomised, stratified by clinical experience	Sequential US family practice attenders - paid $5 to participate. (n=62)	Sickness Impact Profile (SIP)	Int: Physicians given instruction in the SIP. SIP administered by research assistant and fed back prior to consultation. Cont: SIP administered, but results not fed back.	Use of rehabilitative services, and follow up by the physician for rehabilitative problems. Physicians and patients' perceptions of the value of the SIP.	No effect on patient care for the following: return visits to the family physician; referrals to other physicians; use of rehabilitative services. All physicians and patients gave some indication that the SIP was potentially of use. Physicians universally commented that the SIP was too long and difficult to assimilate into the clinical encounter. The results of the SIP were discussed in only 1/3 of consultations.
Kazis *et al.* (1990)	RCT Individual patients randomised	US Outpatients with rheumatoid arthritis (n=1920)	Arthritis Impact Measurement scales (AIMS), which includes a battery of questions relating to anxiety and depression, in addition to arthritis specific questions and ADLs.	Int.: AIMS administered and fed back to the clinician, at least four times over a 12-month period. Substantial change scores and scores outside of population norms were highlighted. Cont.: AIMS administered, but not fed back.	Patient satisfaction with care and health status scores at 12 months. Process measures of physician impressions of the usefulness of the questionnaires also reported	No significant difference in patient satisfaction No significant difference in endpoint depression or anxiety scores on the AIMS

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Rubenstein *et al.* (1989)	RCT Physicians randomised	US internists in community internal medicine practices (n=76), and their patients who visited at least four times per year (n=510)	Functional Status Questionnaire (FSQ) - includes a five item mental health scale	Int.: FSQ administered to patients (n=253) and fed back to clinicians (n=39) every four months in the form of a summary sheet, with major deficits on domains highlighted. Clinicians encouraged to integrate FSQ results and deficits into the clinical encounter as a form of problem identification. Cont.: FSQ administered to patients (n=257), with no feedback to clinicians (n=37) and no clinician education.	Clinicians perception of usefulness of FSQ results Scores on FSQ items at four, eight and 12 months.	48% of clinicians in the experimental group reported using the questionnaire to change therapy. No differences for any subscale of the FSQ (including mental health) at 12 months.
Rubenstein *et al.* (1995)	RCT Individual clinicians randomised	US adult internal medicine outpatient attenders	Functional Status Questionnaire (FSQ) - includes a five item mental health scale	*Int: Physicians (n=40) given an educational package that included teaching sessions on the importance and causes of functional status deficits (including depression). These received structured feedback of FSQ scores from 309 patients.* Cont.: Physicians (n=33) received no feedback from their 248 patients who had completed the FSQ	Patient willingness to complete FSQ instruments. Case note review of recognition of and interventions for identified functional status deficits (including depression or anxiety) FSQ scores at six months	64% patients willing to complete questionnaires and undergo randomisation. Non significant increase in recognition of depressive symptoms (int. vs cont.: 23% vs 20%), Significant increase in the recognition of anxiety symptoms (13% vs 4% p<0.001). Total number of interventions for FSQ problems increased (3.3 vs 2.5 per patient, p<0.05) Mental health scores improved in the feedback group and deteriorated in the cont. group (endpoint mean change difference = 4.5 points (95%CI 0.5-8.3) on a 100 point scale)

Table 29: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author & Year	Design	Population, setting & sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up & outcomes studied	Results
Street *et al.* (1994)	quasi-RCT Individual clinicians allocated to intervention or control	Pregnant women (n=53) attending obstetric outpatients in USA	SF 36	Int.: SF36 administered over the phone by researcher & summary scores included in medical charts at next attendance. Clinicians provided with scores on each of 8 dimensions on the SF36 and a definition of each dimension. Cont.: SF36 administered as above, but not fed back to clinicians	Patient expectation of the clinical encounter Patient satisfaction with care.	Patients were keen to be asked about the dimensions of care included on the SF36 (incl mental health). The provision of summary scores did not influence the pattern of consultation or coverage of these items.
Wagner *et al.* (1997)	RCT Individual patients randomised	Routine patients with epilepsy (n=163), being treated by two US neurologists	SF36, including subscales on role limitations due to emotional problems, and mental health	Int.: Clinicians received a training session on the importance and interpretation of SF36 scores. Patients competed SF36 and summary scores, and profiles were presented in an individualised profile. Cont.: Clinician training programme given and SF36 administered, but SF36 scores not fed back	Physician perceptions of usefulness of scores. Patient satisfaction with care. No follow up beyond the study index encounter	Physicians generally felt the data to be useful. No change in patient satisfaction between intervention and control groups (46% vs 50% ns)
Wasson *et al.* (1992a)	RCT Individual clinicians randomised in blocks according to patient demographics	US HMO in internal medicine specialists (n=56) and their patients (n=1522)	Dartmouth COOP – which includes items on physical condition; emotional condition; daily work; social activities; health change; overall condition.	Int.: Clinicians educated about the nature and interpretation of COOP charts, and COOP chart given to patients prior to consultation, and taken into the consultation by the patients. Cont.: COOP administered, but not given to the clincian.	Clinician self reported use of the charts. Process of care, including test ordering; new medications; patient advice and referral. Patient satisfaction with care	Clinicians reported that charts provided new information on 15-30% of patients. No difference in process of care measures. Overall patient satisfaction unchanged. Nothing could be established about the specific role of the COOP in affecting the management of mental health problems

Chapter 27 Discussion of the main results of the review

This review set out to examine the effect of routine outcome measurement on the actual outcome of those with mental health problems. However, it has only succeeded in identifying evidence relating to one aspect of this: *viz*, the effect of routine outcome measurement on the detection and management of minor psychiatric disorders in general practice and the general hospital. There is no robust research evidence on the effect of outcome measurement on the management of patients in psychiatric settings. The significance of the available research will now be examined, together with a discussion of the reasons for and implications of the paucity of research into routine outcome measurement in psychiatric settings.

Methods of the review

Traditional (non-systematic) review articles in this area (Meakin, 1992; Wright, 1994a) have produced contradictory recommendations without any clear indication as to how their authors have arrived at their conclusions. The present review, in contrast, produces a series of conclusions with a clear and explicit outline of the methods by which those conclusions were arrived at. This demonstrates the major advantage of systematic reviews over traditional review articles. The present research is also novel in that it represents the application of a systematic review methodology to an area that has hitherto not been widely examined in this way – *viz* quality improvement strategies for mental healthcare.

The review has used both quantitative and non-quantitative methods to summarise this research, demonstrating that there is a place for the application of techniques such as meta-analysis, alongside a systematic description of the relative strengths, limitations and results of individual pieces of primary research. A number of methodological aspects of this review deserve further discussion.

Literature searches.

Large amounts of literature needed to be searched in order to obtain only a relatively small number of relevant studies. This demonstrates the difficulties that are inherent in searching for literature in this area, and the need to search multiple databases, and to use broad search strategies, with the expectation that searches

will still be relatively insensitive. Medline proved to be the most fruitful of the databases searched, although important studies would have been missed if this database had been the only source of references. The present study has built upon existing search strategies for literature pertaining to outcomes (Brettle, *et al.*, 1998), and adapted them for use in other important databases. Further research is needed in order to establish the precision and recall of these strategies in identifying mental health outcomes studies.

It is possible that further research exists which was not identified within the literature search strategies. Some of this emerged following the publication of this review in a paper journal, and highlights the advantages of publication in peer reviewed journals in enhancing the quality and comprehensiveness of the review . Similarly, data were left unreported in several studies, but which potentially could have been included in this review. The present review is therefore still likely to incomplete, but will be published and updated in line with existing and emerging data as reviews within the Cochrane library. For example, the study by Lewis, *et al.* (1996) contains unreported data on the rates of recognition of depression by general practioners, which the first author has pledged to make available, but which were not available at the time of writing.

Examination of heterogeneity and publication bias

The research included in the present review was subject to a large degree of heterogeneity. This became apparent when the methods and results of individual studies were described in a systematic way. For example, some studies were so radically diverse in their choice of population, setting, and intervention as to be too heterogeneous to consider for inclusion in a quantitative synthesis. One of the leading authorities on the examination of heterogeneity within meta-analyses has asserted that it is not just sufficient to test for heterogeneity, but the point is to look for causes (Thompson, 1995).

Important sources of heterogeneity that might not have been predicted in advance were those relating to the mode of administration and feedback of outcomes measure (the 'unselected' versus 'high risk' approach). The present study illustrates the complementary nature of quantitative and more qualitative approaches to the examination and exploration of sources of heterogeneity. The use of separate

statistical pooling for divergent approaches to feedback can be defended on both a statistical and an intuitive basis. The clinical implications of 'unselected' versus 'high risk' feedback are explored in more detail in the following sections.

An important strength of the present study was also the steps that were taken to test the robustness of some of the meta-analyses that were performed. Where the inclusion or exclusion of some methodologically heterogeneous studies might be subject to debate, the robustness of the overall meta-analytic result to the presence or absence of these studies was tested. The results of the meta-analysis of unselected feedback of psychological outcomes measures to non-specialists can, with some certainty, be said to be a consistent and robust finding. It will be interesting to know how this result stands up to the inclusion of further data that will be included in further versions of this review.

It was noted in the method section that an important though often overlooked step in the conduct of a review is the examination of publication bias. The present review has highlighted two problems in the examination of the influence of publication bias: the difficulty in applying tests for publication bias, and the difficulty in interpreting the tests that are used.

All published forms of research are potentially subject to publication bias, and there are reasons why psychiatric research is likely to be just as susceptible as research in other areas and specialities (Gilbody & Song, 2000). Conventional tests for publication bias, such as the funnel plot, rely upon two criteria being satisfied: First, studies must be sufficiently similar in terms of participants and interventions to justify a formal statistical pooling in the form of a meta-analysis. Secondly, the published literature must include a sufficient number of studies with a wide range of sample sizes, providing a mix of smaller studies and one or more larger studies with which to construct a funnel plot.

When applying this method of analysis to the group of studies that included the detection of mood disorders as an outcome following feedback, then the second criterion was fulfilled, with a range of study sizes between 80 and 1996. However, for reasons outlined previously, there was felt to be substantial heterogeneity between studies, making the overall application of meta-analysis difficult to justify. When a funnel plot was applied, then the asymmetrical plot that was obtained was

likely to be a reflection of underlying heterogeneity, where this was also a function of sample size. Egger, *et al.*, (1997) urges caution in making the assumption that asymmetrical funnel plots are only indicative of publication bias, and the present review provides an interesting example of this. Petticrew, *et al.*, (1999) have also demonstrated the potential for heterogeneity to produce asymmetrical funnel plots, where differences in effect size were related to the underlying quality of observational research in the area of heart disease.

Cluster randomised studies

Studies designed to evaluate quality improvement strategies should ideally use randomisation by cluster (Ukoumunne, *et al.*, 1999b). In the case of studies designed to answer the question addressed in the present review, it should be individual clinicians or clinical teams who are randomised to receive feedback of outcomes measures, in order to prevent cross contamination between individual patients. Nine of the twenty four studies randomised by cluster, but were not correctly analysed, with analysis taking place at the level of the individual patient, without due consideration of the effect of clustering. None of the studies included in the quantitative syntheses used cluster randomisation. The clinical implications of the failure to conduct clustered studies when this was the appropriate design to use, and the inclusion of potentially clustered data in systematic reviews deserves further comment.

Previous reviews of quality improvement strategies (eg Grimshaw & Russell, 1993) have also generally found that these studies either fail to randomise by cluster when they should, or fail to analyse these data appropriately when they do randomise by cluster. The difficulties in handling clustered data stem form the fact that individuals within clusters share common socio-demographic features such as age, sex or social class – all of which are potentially related to outcome (Rice & Leyland, 1996). Traditional statistical approaches make certain assumptions, including the assumption that outcomes or events for different patients are in some way independent of each other. When randomisation by cluster occurs and outcomes are analysed at the level of the individual patient, then this assumption breaks down. Failure to recognise this fact in the analysis of data has been termed 'unit of analysis error' (Divine, *et al.*, 1992), and leads to over optimistic estimates of sample

variance, unduly narrow confidence intervals and potential type 1 errors (finding an effect or association, when one does not exists).

A number of approaches have been advocated in the inclusion of potentially misleading clustered studies in systematic reviews. Firstly one approach used by Grimshaw & Russell (1993) is to draw attention to the unsound nature of studies subject to a unit of analysis error. Another approach is to seek to correct for unit of analysis errors by seeking to find the level of correlation within clusters (expressed as the intra class correlation coefficient) from authors of studies, and to seek to correct the unit of analysis error by reanalysing the results of the study (Ukoumunne, *et al.*, 1999b). However, there remains substantial difficulty in subjecting clustered data, even when corrected, to meta-analytic pooling, since individual study variance estimates in conventional meta-analytic methods (eg DerSimonian & Laird, 1986; Mantel & Haenszel, 1959) do not allow for clustering. There remains no consensus regarding the appropriate way to proceed in the meta-analysis of potentially informative groups of studies, and this issue is currently being investigated by a methodological workgroup within the Cochrane Collaboration {Professor Mike Campbell, University of Sheffield, Personal communication November 2000}.

Attempts to deal with clustering in the present review were limited, since the authors of clustered studies in the present review did not reply to a request to provide intra-class correlation coefficient in order to correct a unit of analysis error. No studies included in the meta-analysis had utilised cluster randomisation, making the problem of how to handle these data in a quantitative analysis of academic interest only. The emergence of further studies, some of which may be clustered will necessitate re-evaluation of this approach and are likely to make the use of meta-analysis untenable for this set of studies.

The negative result that was found for many outcomes (especially the effect of feedback on the rate of recognition of mood disorders) could have also resulted from the failure to use a correct unit of randomisation i.e. individual patients rather than individual clinicians. The cross contamination which potentially might have occurred between patients might have resulted in a dilution of effect, and a spurious negative result (type 2 error). It is likely that the very act of receiving feedback of outcome measures on some patients will influence how other patients, who do not have their outcome score fed back, will be managed. The following results, which are

discussed in more detail below, must therefore be considered alongside this inherent weakness of the research surveyed in the present review.

Clinical implications of the review

Mood disorder questionnaires in non-psychiatric settings

It is perhaps surprising that the uniform administration of well validated case finding instruments, such as the GHQ, with sensitivities and specificities of over 70 and 90% respectively in their ability to detect psychiatric disorders (Goldberg, 1972; Goldberg & Williams, 1988), has not been found to influence actual clinician behaviour. Routine outcome measurement only becomes effective in increasing the rate of recognition of emotional disorders when there is some form of screening procedure, whereby an instrument is administered, scored by someone other than the clinician, and only those with high scores have their results fed back to the clinician (e.g. Rubenstein, *et al.*, 1995). Routine administration combined with selective feedback is, however, unlikely to form a model for routine practice, nor does it reflect current UK practice, since this strategy is likely to require that an additional person be employed in order to administer score and feedback outcomes measure to the clinician.

There are a number of possible explanations for the observed result. First, it is *predictive value* (rather than sensitivity and specificity) which is of most interest to clinicians in the context of routine care - i.e. the proportion of those predicted by the test as having the disease who turn out to have the disease (Sackett, *et al.*, 1991). Crucially, positive predictive value increases according to the prevalence of a disorder in the population tested. Whilst unrecognised emotional disorders form a significant portion of the clinical caseload in non-psychiatric services, this is rarely going to exceed 15%. The consequence is that of those patients with a positive screening result, only 50% will turn out to actually have an emotional disorder (i.e. be 'true positives') (Hoeper, *et al.*, 1984). Equally, the workload and outside referral rate is likely to rise dramatically if all positive test results are acted upon when positive predictive value is much lower than quoted sensitivities and specificities. Clinicians may intuitively recognise this fact and will be unwilling to act on positive test results (Goldberg, 1986).

A major limitation of the research presented in this review is the fact that case definition of an emotional disorder (such as depression or anxiety) is generally based upon a questionnaire score above a certain cut off point, rather than some

gold standard, such as a standardised research interview. Thus, the principle trial endpoint - rates of recognition of emotional disorders - uses this imperfect form of case definition. Research shows that questionnaires consistently overestimate the true prevalence of clinically important emotional disorders (i.e. those worthy of intervention) e.g. (Feldman, *et al.*, 1987). It should perhaps therefore be less surprising that clinicians in this review uniformly ascribed far fewer patients as having emotional problems than did questionnaires. However, the negative result for feedback suggests that questionnaire results, in effect, add nothing to the clinical encounter. Calls for the routine application of such questionnaires in non-psychiatric settings (Wright, 1994a) are therefore not supported

A second explanation is that non-psychiatrists do not feel best equipped to deal with emotional disorders, even when these are uncovered using screening questionnaires. Screening is therefore a necessary, but not sufficient, condition in facilitating the appropriate management of these psychological problems. Supporting this conclusion is the observation that feedback is most effective when it is accompanied by an educational programme and the provision of a dedicated outside referral agency who will readily assume responsibility for management (Mazonson, *et al.*, 1994). The results of the present review also complement recent research which shows that simple educational interventions, such as the provision of guidelines on the detection and management of depression in primary care have little impact (Thompson, *et al.*, 2000).

Worthy of further research is also the suggestion that some patient groups might benefit from the routine administration of psychiatric screening questionnaires more than others. For example the subgroup analysis by German, *et al.*, (1987) suggests that the elderly may benefit more from routine administration and feedback of psychiatric questionnaires, as do men. Depressive disorders in these populations often present with non-specific somatic complaints (Goldberg & Bridges, 1988), which can prevent or delay the detection of mood disorders. However, whilst routine outcome measurement may increase the rate of detection of depression, this does not generally translate into increased rates of intervention. The ultimate goal of routine outcome measurement is to improve outcome, and the research strongly suggests that there is no benefit in this respect

Do available studies examine 'routine' outcome measurement?

A key aim of the review was to examine the use of standardised instruments as outcome measures in routine care settings, and several of the studies in fact identify themselves as examining this question. However, as discussed in section 1 of this thesis, the measurement of 'outcome' is generally taken to mean the measurement of some facet of health status over time. In the context of routine care this would involve the serial application of the instrument, so that changes in the score might be incorporated into patient management in some way. However, all the studies in the current review involve the single administration of an instrument at an initial index episode, with no further application by the clinician at subsequent consultations. The use of outcome instruments in this context is essentially a form of screening (Fitzpatrick, 1994; Fitzpatrick, *et al.*, 1992a).

Screening tests can only be justified if the instrument is (1) accurate; (2) results in a more effective treatment than would otherwise be the case and; (3) does so with a favourable ratio of costs to benefits (Cochrane & Holland, 1971; Mant & Fowler, 1990). The accuracy of an instrument is traditionally determined by the examination of sensitivity, specificity and predictive value. Several of the authors justified the choice of their instrument with reference to its sensitivity and specificity as determined in prior validation. Only one examined or published these key psychometric properties within the populations that were recruited or randomised (Williams, *et al.*, 1999). However, it is *predictive value* which is of most interest to clinicians in the context of routine care - i.e. the proportion of those predicted by the test as having the disease who turn out to have the disease (Sackett, *et al.*, 1991). Predictive value increases as the incidence of disease in the population under investigation increases and this is essentially what is happening when the instrument is administered to all patients and only those with positive score have their results 'fed back'. This is a likely explanation of the improved recognition by clinicians when feedback occurs with only 'high risk' patients as opposed to feedback with all patients. Further research might seek to evaluate the routine use of outcome measures using basic psychometric criteria such as sensitivity, specificity and predictive value.

The second criterion which must be fulfilled for a screening instrument is that its use should result in effective treatment. The evidence outlined in the present review shows that this is under researched, and the research that has been conducted is

not generally supportive. Routine feedback generally does not change clinical management and when actual outcome is studied, then this is generally not shown to improve (e.g. Dowrick & Buchan, 1995). The last criterion to be satisfied is that the benefits of screening should outweigh cost. Cost can include the costs (monetary, time and forgone opportunity) incurred through the introduction of routine outcome measurement, and no studies in this review measured this. Additionally, cost involves the harm which might be done through routine outcome measurement in terms of the initiation of treatment for those wrongly identified as having some psychological disorder ('false positives'), or the initiation of resource intensive referral or intervention for those who might be identified as having some emotional problem, but which might be self limiting. Further research is needed in all these respects and in the absence of such research, then it would be imprudent to recommend the introduction of routine outcome measurement in routine care settings.

The use of generic patient based measures.

Despite the enthusiasm for recently introduced generic health status measures, such as the SF-36, there is no robust research evidence to support their value as routine measures of outcome in psychiatric settings. However, there is some tentative research evidence to support their use to facilitate the recognition of mental health problems in non psychiatric settings (Rubenstein, *et al.*, 1995). As is discussed above, the adoption of routine outcome measure in individual patient care is not without cost, and there is at present insufficient evidence to justify this. It is possible that benefit cannot and will not ever be demonstrated for the routine use of these measures in individual patient decision making, since this is a purpose for which generic instruments are not designed. In particular, the psychometric properties of such measures are such that scores on these instruments are un-interpretable at an individual patient level (McHorney & Tarlov, 1994). Generic outcomes measures are essentially designed to evaluate healthcare and to identify need at a *population* level (Ware, 1995), and extrapolation of use beyond this is not justified.

Routine measurement of outcome in psychiatric settings.

National mental health research and policy initiatives, such as the development and adoption of the Health of the Nation Outcome Scales (HoNOS) (Wing, 1994) are dependant upon individual clinicians collecting these data in the context of routine care (Stein, 1999). For clinicians to be willing to collect such data for each and every patient there must be some value in terms of improving the management of the individual patient. No such evidence was found to support its implementation in the context of routine care

Section 4 Overall discussion of the use outcomes measures in psychiatry

The thesis began by presenting an overview of the wider *outcomes movement* in healthcare; examining the origins of this movement and the implications of this shift towards outcomes measurement and the introduction of more patient based measurement instruments. The original research presented in this thesis has largely been an exploration of this outcomes movement and patient based outcomes measurement within psychiatric research and practice.

Surveys of psychiatric research found that outcomes measurement in psychiatry is dominated by the measurement of symptoms, with little explicit adoption of patient based measures. Interestingly, it was found that a minority of trials has for some time incorporated the measurement of domains of patient based outcome – such as social functioning.

A survey of the measurement of outcome within a less well-known or less widely used research design – outcomes research - was conducted. Outcomes research is purported to bridge the gap between psychiatric research and practice, since it incorporates those outcomes collected in the context of routine practice in order to provide an alternative to randomised trials. With notable exceptions, similarly limited sets of outcomes were found to be used in outcomes research as were found in clinical trials. The primary problem with outcomes research is the time and expense involved in the collection of a diverse and comprehensive set of outcomes in routine care settings.

The difficulties inherent in collecting outcomes data in the context of routine care settings was further explored in a large-scale survey of UK consultant psychiatrists. This survey presented the first overview of current UK practice, and found that clinicians do not routinely measure outcome (patient based or otherwise) in the context of their routine practice. Substantial practical and attitudinal barriers were identified to the collection of standardised outcomes that will need to be addressed if current UK mental health policy is to be implemented.

Lastly, the first systematic review was undertaken in order to examine what evidence, if any, exists to support the benefits of routine outcomes measurement in improving the quality of care that is offered to those with psychiatric illness. There is no evidence to support the routine collection of outcomes measures in routine psychiatric care settings. When evidence to support the use of routine outcomes measures in non psychiatric care settings is explored, largely in the form of psychiatric case finding instruments, then a substantial body of research shows this to be an *ineffective strategy*.

The implications of the original research presented in this thesis will now be considered, with reference to psychiatric practice, policy and research.

Implications for mental health practice

Clinicians are increasingly encouraged to incorporate research evidence, such as the results of randomised trials, into their everyday practice (Sackett, *et al.*, 1991). A key finding of the surveys of how outcome is measured in clinical trials and what clinicians actually collect and use in their own practice is that there is little correspondence between practice and research. Outcome in clinical trials, particularly in drug trials, is measured using complex psychopathological rating scales. These are rarely used in clinical practice, and it is doubtful that clinicians who have little familiarity with these instruments can interpret the meaning of small changes on these rating scales. Small changes on symptom based psychopathology rating scales are the major criterion for success or otherwise of interventions in randomised trials. The uptake of new technologies, such as new drug entities, therefore happens for reasons other than the results of evidence from randomised trials, when this evidence is based upon unfamiliar outcomes that are difficult to interpret. A greater correspondence between research and practice will therefore require either clinicians to begin using the outcomes instruments that are used in clinical trials or researchers to begin collecting and reporting those outcomes that are of genuine interest to clinicians. From the results of the survey of clincians, these measures are unlikely to be complex psychopathological rating scales.

The survey of clinical practice showed a general reluctance amongst clinicians to collect outcomes and gave insight into the reasons behind this. It was clear from some of the comments made by clinicians that they perceived standardised outcomes measures to be 'research tools', rather than instruments that could be easily incorporated into their routine practice. Unfortunately, the surveys presented in this thesis did not explore what outcomes clinicians would like to see collected in evaluative research. This is topic for further research, and is explored in more detail below.

The rhetoric of outcomes measurement outlined in section 1 and highlighted in important mental health policy formulations (Department of Health, 1991; Department of Health, 1998; Marks, 1998; Secretary of State for Health, 1999; Slade, *et al.*, 1999) has not permeated clinical practice. Standardised measures do not generally form a part of the routine care of those with psychiatric disorders such as schizophrenia, nor are they used as measures of outcome by their serial application over time in order to measure change. The development of patient based measures and measures of psychosocial need has generally not resulted in these measures and instruments being used in the day to day care of those with common mental disorders being looked after in UK mental health services. This represents a major disparity between mental health policy and actual clinical practice, which had previously been alluded to (Slade, *et al.*, 1999), but which had not otherwise been empirically demonstrated.

Substantial barriers to the routine use of outcomes were identified and include: lack of familiarity with instruments; the length of time taken to complete measures; lack of resources made available with which to adopt routine outcomes measures and a lack of faith in the basic psychometric properties and real world relevance of available measures. Importantly, some clinicians questioned the clinical and cost effectiveness of routine outcomes measurement as a technology.

Clearly, if psychiatrists are going to use standardised measures of outcome, including patient based measures, in the course of their day to day practice, then each and every one of the barriers identified in the survey will need to be

addressed. Importantly, the resources required in implementing routine outcomes measurement have not been made available in UK mental health services. This lack of investment in outcomes measurement was highlighted by a number of clinicians within the survey of the UK practice. However, in advance of the investments that would be need to be made in order to make routine outcomes measurement work, a more fundamental question about whether outcomes measurement is a worthwhile activity needs to be asked.

The research presented in this thesis explicitly demonstrates for the first time the fact that mental health policy with respect to routine outcome measurement is being formulated in the absence of robust evidence of effectiveness in influencing practice or patient outcome. When research evidence was sought in order to answer this question, then none was found to have been conducted in psychiatric care settings. An important body of research evidence was found that showed that such an approach has not proved to be useful in non-psychiatric care settings. In the absence of a robust body of research, then the value of routine outcomes measurement remains unproven. The research that would be needed in order to demonstrate this benefit was discussed in some detail in Chapter 26, and further explored below. Similarly, the reasons for the major disparity between mental health practice and policy formulation are explored in more detail below.

We can speculate as to whether the investment in outcomes measurement as a technology would result in its adoption by clinicians. Clinicians are unlikely to change their practice unless they perceive some benefit to themselves or to the patients in the care that is delivered. Similarly, patients are unlikely to comply with the collection of repetitive and complex questionnaires unless they see some benefit to the care that they receive. Whilst available instruments are perceived as unwieldy, irrelevant and uninformative, then they will continue to represent a threat to effective care, rather than a tool with which to improve the quality and outcome of care.

This view was expressed by Alvan Feinstein (1967) more than 30 years ago, when he wrote:

'The care of the patient is the ultimate specific act that characterises the clinician, and any classificatory system that cannot help in that will fail to gain acceptance'

Feinstein stressed that the unless the clinician believes that an intervention would directly help the patent in the consulting room, or at the very least, in assisting in the diagnostic or clinical process, then the intervention will not be undertaken.

Implications for mental health policy

Recent mental health policy encourages the measurement of outcome. However, policy formulations to measure outcome on a routine basis have essentially been 'top down', with little consideration of the time and resources involved. Two high cost and high profile research and development activities serve to illustrate this approach. The Health of the Nation Scale has been developed at substantial cost as a tool to evaluate the success or otherwise of health policy formulations (Wing, 1994). The HoNOS forms a core component of a battery of outcomes measures that all clinicians and Trusts are (at the time of writing) to be forced to collect as a matter of routine – this battery is known as 'the minimum data set' (Glover, *et al.*, 1997).

The survey of UK consultants in this thesis has shown that they are less than keen to collect these data, and that Trusts have little experience or success in encouraging their clinicians to collect data as a matter of course. What Trusts seem to have uniformly done is collect those administrative outcomes that are easy to collect (such as length of stay and readmission rates), and which form part of the Performance Management Framework outlined in recent health policy documents (Department of Health, 1998; Secretary of State for Health, 1999). Similarly, it is these data that are fed back to clinicians and form the mainstay of audit activities, despite the aspiration that audit would be a more patient centred approach (Frater & Costain, 1992). Recent evidence on the collection and publication of routinely collected performance data in Scotland suggests that such data are largely ignored in the planning and improvement

of clinical services (Mannion & Goddard, 2001). The survey of UK consultants provides empirical support that this observation is also true in the planning and improvement of mental health services. UK psychiatrists gave few examples of positive experiences or knowledge of routinely collected outcomes data being used in the planning or improvement of clinical services, and many believed that the data they were asked to collect was a bureaucratic exercise.

Mannion & Goddard (2001), in their exploration of the impact of routinely collected outcomes in changing clinical practice and in improving the quality of care identify several major themes which are germane with those highlighted in the present thesis. Outcomes data have little impact when they are not perceived as being *credible* in terms of their quality or relevance. Similarly, the *timeliness* of outcomes data, when there is a substantial delay between their collection and feedback hampers their impact. The absence of any programme of *training and facilitation* in the interpretation and appropriate use of outcomes data also makes their collection and feedback a bureaucratic exercise.

This distortion of the behavior of organisations that also occurs when there is a pre-occupation with a small number of easy to measure outcomes indicators was discussed in Chapter 22 (see Davies & Lampel, 1998; Smith, 1996a). It is clear from the survey of clinical practice that those measures that are collected by Trusts are those that are easy to measure, rather than those that are of importance or value. This was a widely held perception amongst clinicians. There is a very real danger that the elevation of easy to collect data, rather than clinically meaningful data, to the position of a performance indicator will adversely affect the outcome of patients, or will at best, confer little advantage. The perverse consequences of the limited focus on routinely collected outcomes measures are summarised in table 30. Davies & Crombie (1997) have highlighted the need for studies that examine the impact on organisations and individuals of the regular feedback of outcomes data.

Table 30: Perverse consequences of a limited focus on outcomes measures

(after Davies & Crombie, 1997; Smith, 1996a)

Tunnel vision	Concentration on those areas in the outcome set, to the exclusion of other important areas
Suboptimisation	The pursuit of narrow objectives within a unit or organisation at the expense of strategic co-ordination with others
Myopia	Concentration on short term issues to the exclusion of long term criteria
Ossification	A disinclination to experiment with new and innovative practices
Convergence	An emphasis on not being exposed as an outlier rather than a desire to be outstanding
Gaming	The alteration of behaviour to gain strategic advantage
Misrepresentation	Including creative accounting and fraud

The survey of UK psychiatrists identified substantial barriers to the routine use of outcomes measures by clinicians that will have to be addressed if current mental health policy is to be implemented. Most importantly, the whole value of routine outcomes measurement is called into question by the research presented in this thesis. On the basis of the research, there are very good reasons to suppose that mental health policy that involves and relies upon the routine collection of standardised measures is likely to be unsuccessful. Further research (see on) should precede the further implementation of this strategy.

Mental health policy, with respect to routine outcomes measurement, is therefore formulated either in the absence of evidence or in the face of evidence that shows it to be ineffective. The drivers of this urge to measure outcomes are therefore political and sociological. The reasons for this urge to measure outcomes were discussed in detail in section 1, and included the pressure to be seen to measure things in order to establish what works, and to be seen to be improving the quality of care that is delivered. Michael Power (1997) places the urge to collect and measure things within a wider context of increased accountability of professions and institutions, and the need to demonstrate value and worth. He outlines 'rituals of verification' which have sprung up in all spheres of the public sector, with little thought about the effectiveness or consequences of these changes.

Implications for mental health services research

Limitations of existing randomised-controlled trials are all too evident in psychiatry. These include limited external validity, and the collection uninformative outcomes. These limitations are highlighted in the survey of randomised trials in the present thesis. The need to address these shortcomings was one of the motivating forces behind the development of techniques such as *outcomes research*, whereby routinely collected data are harnessed in order to establish what works and for whom within routine care settings (Wennberg, 1991). The present thesis has highlighted both the potential and the limitations of this approach. Within the context of UK mental health services, there is little prospect of the successful adoption of *outcomes research* when clinicians are demonstrably so reluctant to collect outcomes, and when the outcomes that they do collect fall so far short of the patient centred approach advocated by proponents of outcomes research. Outcomes research in mental health, as expounded by Wells (1999) and Marginson *et al.*, (2000) will not become viable until the barriers to the collection of routine outcomes measures, outlined above, have been addressed. In the meantime, outcomes research that utilises the limited clinical data collected within UK mental health services should be interpreted with caution.

Two major strands of further research are identified as priorities by the research presented in this thesis. The first relates to the use of patient based outcomes as instruments in psychiatric research. The second relates to important research that needs to precede the implementation of outcomes measurement in routine clinical care settings.

The diversity and lack of coherence in outcomes measurement that has been demonstrated in the surveys of clinical evaluations deserves clear thought about what instruments should be used, for whom and in what settings. Psychiatric research has a strong tradition of patient based outcome measurement, as evidenced by the measurement of social functioning. However, more recently developed measures of patient based outcome, such as health profiles and health utility measures, have not been widely used. Basic research is needed to judge the potential of these measures to be used

within clinical evaluations in psychiatry. Recent methodological reviews conducted under the auspices of the NHS Health Technology Assessment Programme provide a source of guidelines as to how these questions should be tackled (Brazier, *et al.*, 1999; Fitzpatrick, *et al.*, 1998). Table 31 provides the key properties and dimensions which must be satisfied by patient based outcomes measures in order that be used within clinical trials.

Table 31: Essential properties of a patient based outcome measure
(Fitzpatrick, *et al.*, 1998)

Appropriateness	Is the content of the instrument appropriate to the question that the clinical trial is intended to address?
Reliability	Does the instrument produce results that are reproducible and internally consistent?
Validity	Does the instrument measure what it claims to measure?
Responsiveness	Does the instrument detect changes over time that mater to patients?
Precision	How precise are the scores of the instrument?
Interpretability	How interpretable are the scores of the instrument?
Acceptability	Is the instrument acceptable to the patients?
Feasibility	Is the instrument easy to administer and process?

The wide variety of standardised outcomes measures that are available and are used in clinical trials in psychiatry also deserves further consideration. Rheumatology is a speciality that found its research evidence to be bedevilled by similar problems to psychiatry, particularly the abundance of disparate measurement techniques and instruments. A key stage in the evolution of outcomes measurement in rheumatology was the construction of a core battery of outcomes measures, the use of which was widely prompted as good practice by leading researchers (Felson, *et al.*, 1993; Tugwell & Boers, 1993). Unfortunately, the choice of method and outcomes that are explored in psychiatric research is largely influenced by the main sponsor of research (the pharmaceutical industry) and its needs. This is in turn dictated by drug regulation bodies and the outcomes that they demand as sufficient evidence of efficacy in order that a drug is granted a licence. A clear indication by the

major drugs licensing bodes, such as the Food and Drug Administration and the Medicines Control Agency, that they will demand evidence of benefit in terms of patient based outcomes would encourage the adoption of these measures.

The lack of evidence to support the adoption of routine outcomes measurement is perhaps the most important finding of the current thesis. Having described the importance of the questions, and having laid out a clear argument both for and against the collection of outcomes data, then a systematic review was conducted. In advance of conducting the research contained in this review, it was taken as an article of faith by the present author that outcomes measurement was a 'good thing', and a worthy activity. It was anticipated that the important gaps in the research knowledge base that would be identified would be those surrounding the basic implementation of outcomes measurement as an activity or technology.

The argument was thought to be germane to that of guidelines in healthcare. Guidelines have come to be seen as a 'good thing', that have potential to influence practice for the better. The important questions relating to guidelines are about their construction such that they are credible and evidence based, and about how they should be adopted or implemented in order that they change practice (NHS Centre for Reviews and Dissemination, 1994; NHS Centre for Reviews and Dissemination, 1999b).

This is clearly not the case for routine outcomes measurement. Given the enormity of the task involved in encouraging reluctant clinicians to collect outcomes, then there needs to be demonstrated benefit in terms of the potential of routine outcomes measurement, as an intervention, to benefit both clinicians and patients. Chapter 27 has outlined what sort of evidence would be needed in order to demonstrate or refute the value of routine outcomes measurement, *viz* cluster-randomised trials conducted in routine care settings. In the absence of this evidence, then health practice and health policies are unlikely to shift from their current position of reluctant and ineffective collection by clinicians and healthcare systems and inevitable policy failures.

References

Aaranson, N. (1989) Quality of life assessment in clinical trials: methodologic issues. *Controlled Clinical Trials* **10**, 195-208s.

Adams, C. (1998) The Cochrane Schizophrenia Group's Register of Randomised Controlled Trials. In *The Cochrane Library*. Oxford: Update Software.

Adams, C. E. (1997) Establishing cost-effectiveness of antipsychotic drugs. *Br J Psychiatry* **171**, 486.

Adams, C. E., Gray, R., Daniels, J., Philpot, H., et al (1999) *Are UK consultant psychiatrists interested in pragmatic trials for those with schizophrenia? (Poster presentation)*. Paper presented at the Schizophrenia Trials Meeting, Stratford Upon Avon, UK, May 1999.

Adams, C. E., Lefebre, C. & Chalmers, I. (1992) Difficulty with MEDLINE searches for randomised controlled trials. *Lancet* **340**, 915-916.

Adams, C. E., Power, A., Frederick, K. & Lefebvre, C. (1994) An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. *Psychol Med* **24,** (3), 741-8.

Aday, L. A., Begley, C. E., Lairson, D. R. & Slater, C. H. (1998) *Evaluating Healthcare System: Effectiveness, efficiency and equity* (Second edn). Chicago: AHSR.

Agency for Health Care Policy Research (1993) *Depression in primary care*. Washington DC: US Department of Health and Human Services.

Aitchison, K. J. & Kerwin, R. W. (1997) Cost effectiveness of clozapine: A UK clinic based study. *British Journal of Psychiatry* **171**, 125-130.

Albrecht, G. L. (1994) Subjective health assessment. In *Measuring health and medical outcomes* (ed C. Jenkinson). London: UCL Press.

Al-Shahi, R. & Warlow, C. (2000) Using patient identifiable data for observational research and audit: overprotection could damage the public interest. *British Medical Journal* **321**, 1031-2.

Altman, D. G. (1991) *Practical Statistic for Medical Research*. London: Chapman Hall.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual - 4th Edition*. Washington DC: American Psychiatric Association.

Anderson, C. (1994) Measuring what works in health care. *Science* **263**, 1080-1082.

Anderson, R. (2001) Undermining data privacy in health information: new powers to control patient information contribute nothing to health. *British Medical Journal* **322**, 442-3.

Anonymous (1987) A proposal for more informative abstracts of clinical articles. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. *Annals of Internal Medicine* **106**, 598-604.

Anonymous (1989) Databases for healthcare outcomes. *Lancet* **396**, 195-196.

Anonymous (1992) Cross design synthesis: a new strategy for studying medical outcomes. *Lancet* **340**, 944-946.

Anonymous (1994) From research to practice. *Lancet* **344**, 417-418.

Anthony, W. & Rogers, S. (1995) Relationship between psychiatric symptomatology, work skills, and future vocational performance. *Psychiatric Services* **46**, 353-358.

Asch, D. A., Jedrziewski, K. & Christiakis, N. A. (1997) Response rates to mailed surveys published in medical journals. *Journal of Clinical Epidemiology* **50**, 1129-1136.

Awad, A. G. (1992) Quality of life of schizophrenic patients on medication and implications for new drug trials. *Hospital and Community Psychiatry* **43**, 262-265.

Bagnall, A.-M., Fenton, M., Lewis, R., Leitner, M. L., et al (2001a) Molindone for schizophrenia and severe mental illness (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Bagnall, A. M., Lewis, R., Gilbody, S. M. & Kleijnen, J. (2001b) New atypical drugs for schizophrenia. *Health Technology Assessment* **In Press**.

Baker, R. & Hall, J. N. (1988) REHAB: a new assessment instrument for chronic psychiatric patients. *Schizophrenia Bulletin* **14**, 95-113.

Bardsley, M. & Coles, J. (1992) Practical experiences in auditing outcomes. *Quality in Health Care* **1**, 124-130.

Barkham, M., Evans, C., Marginson, F., McGrath, G., et al (1998) The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Mental Health* **7**, 35-47.

Baumberg, L., Long, A. & Jefferson, J. (1995) *International workshop: culture and outcomes*. Paper presented at the European Clearing Houses on Health Outcomes., Barcelona, 9-10 June 1995.

Bebbington, P., Brugha, T., Hill, T., Marsden, L., et al (1999) Validation of the Health of the Nation Outcome Scales. *British Journal of Psychiatry* **174**, 389-394.

Bech, P., Malt, U. F., Denker, S. J., Ahlfors, U. G., et al (1993) Scales for the assessmetn of diagnosis and severity of mental disorders. *ACTA Psychiatrica Scandinavica* **87,** (372), Supplementum.

Beck, A. T. & Ward, C. H. (1961) An inventory for measuring depression. *Archives of General Psychiatry* **4**, 561-571.

Becker, M., Diamond, R. & Sainfort, F. (1993) A new patient focussed index for measuring quality of life in persons with severe and persistent mental illness. *Quality of Life Research* **2**, 239-251.

Beckingham, A. (1994) Measuring the results of your purchasing decisions. *The Health Summary* **11,** (5), 11.

Bentham, J. (1789) *An Introduction to the Principles of Morals and Legislation*. New York: Hafner.

Bergner, M. (1985) Measurement of health status. *Medical Care* **23**, 696-704.

Bergner, M. (1989) Quality of life, health status, and clinical research. *Med Care* **27,** (3 Suppl), S148-56.

Bergner, M., Bobbitt, R. A., Carter, W. B. & Gilson, B. S. (1981) The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* **19,** (8), 787-805.

Bergner, M., Bobbitt, R. A. & Kressell, S. (1976) The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. *International Journal of Health Services* **6**, 393-415.

Bergner, M. & Rothman, M. L. (1987) Health status measures: an overview and guide for selection. *Annu Rev Public Health* **8**, 191-210.

Bero, L., Grilli, R., Grimshaw, J. & Oxman, A. e. (1998) The Cochrane Effective Practice and Organisation of Care Group (EPOC) Module. In *The Cochrane Library Issue 4*. Oxford: Update Software.

Best, J. (1988) The matter of outcome. *Medical Journal of Australia* **148**, 161.

Bigelow, D. A., Gareau, M. J. & Young, D. J. (1982) A quality of life interview. *Psychosocial Rehabilitation Journal* **14**, 94-98.

Black, N. A. (1999) High Quality Clinical Databases: breaking down barriers. *Lancet* **353**, 1205-1206.

Blumberg, M. S. (1991) Potentials and limitations of database research illustrated by the QMMP AMI Medicare mortality study. *Statistics in Medicine* **10**, 637-646.

Bombardier, C. & Tugwell, P. (1982) A methodological framework to develop and select indices for clinical trials: statistical and judgemental approaches. *Journal of Rheumatology* **9**, 753-757.

Borg, W. R. & Gall, M. D. (1983) *Educational research: an introduction*. New York: Longman.

Bowling, A. (1995) *Measuring Disease*. Buckingham: Open University Press.

Bowling, A. (1997) *Measuring Health: A review of quality of life measurement scales.* (Vol. Volume 2.). Buckingham: Open University Press.

Brazier, J. (1993) The SF-36 health survey questionnaire--a tool for economists [comment]. *Health Econ* **2,** (3), 213-5.

Brazier, J., Deverill, M., Green, C., Harper, R., et al (1999) A review of the use of health status measures in economic evaluation. *Health Technology Assessment* **3,** (9).

Brettle, A. J., Long, A. F., Grant, M. J. & Greenhalgh, J. (1998) Searching for information on outcomes: do you need to be comprehensive? *Quality in Health Care* **7**, 163-167.

Brewin, C. R. & Wing, J. K. (1993) The MRC Needs for Care Assessment: progress and controversies [editorial]. *Psychol Med* **23,** (4), 837-41.

Bristow, M. F. (1999) Usage of clozapine and the new neuroleptics: a postal survey among general psychiatrists. *Psychiatric Bulletin* **23**, 478-480.

Broadhead, W. E., Leon, A. C. & Weissman, M. M. (1995) Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Archives of Family Medicine* **4**, 211-219.

Brook, R. H. & Appel, F. A. (1973) Quality of care assessment: choosing a method for peer review. *New England Journal of Medicine* **288**, 1323-1329.

Brook, R. H. & Lohr, K. (1985) Efficacy, effectiveness, variations and quality. *Medical Care* **23 (supp)**, 710-722.

Brook, R. H., Ware, J. E., Jr., Rogers, W. H., Keeler, E. B., et al (1983) Does free care improve adults' health? Results from a randomized controlled trial. *N Engl J Med* **309,** (23), 1426-34.

Brookes, R. G. (1995) *Heath Status Measurement: a perspecitve on change*. Basingstoke: McMillan.

Brown, T. L., Decker, D. J. & Connelly, N. A. (1989) Response to mail surveys on resource based recreation topics: a behavioural model and an empirical analysis. *Leisure Services* **11**, 99-110.

Buchan, I. (2000) StatsDirect (Version 1.6). Cambridge.

Cairns, J. (1996) Measuring health outcomes. *British Medical Journal* **313**, 6.

Calkins, D. R., Rubenstein, L. V. & Cleary, P. D. (1994) Functional disability screening of ambulatory patients: a randomised controlled trial in a hopital based group practice. *Journal of General Internal Medicine* **9**, 590-592.

Callahan, C. M., Hendrie, H. C., Dittus, R. S., Brater, D. C., et al (1994) Improving treatment of late life depression in primary care: a randomized clinical trial. *Journal of the American Geriatrics Society* **42**, 839-46.

Calman, K. C. (1984) Quality of life in cancer patients: an hypothesis. *Journal of Medical Ethics* **10**, 1551-1555.

Campbell, M. K. & Grimshaw, J. M. (1998) Cluster randomised trials: time for improvement. *British Medical Journal* **317**, 1171.

Campbell, M. K., Mollinson, J., Steen, N., Grimshaw, J. M., et al (2000) Analysis of cluster randomised trials in primary care: a practical approach. *Family Practice* **17**, 192-196.

Carlwood, P., Mason, A., Goldacre, M., Clearly, R., et al (Eds.). (1999) *Health Outcome Indicators: Severe Mental Illness. Report of a working group to the department of Health*. Oxford: National Centre for Health Outcomes Development.

Chalmers, I. & Altman, D. G. (Eds.). (1995) *Systematic Reviews*. London: BMJ.

Clark, H. D., Wells, G. A., Huet, C., McAllister, F. A., et al (1999) Assessing the quality of randomised trials: reliability of the Jadad scale. *Controlled Clinical Trials* **20**, 448-52.

Clifford, P. (1998) M is for Outcome: the CORE outcomes initiative. *Journal of Mental Health* **317**, 1167-1168.

Cochran, W. G. (1954) The combination of estimates from different experiments. *Biometrics* **10**, 154-173.

Cochrane, A. L. (1972) *Effectiveness and efficiency: random reflections on health services*. London: Nuffield Provincial Hospitals Trust.

Cochrane, A. L. & Holland, W. W. (1971) Validation of screening procedures. *British Medical Bulletin* **27**, 3-8.

Codman, E. (1914) The product of a hospital. *Surgery, Gynaecology and Obstetrics* **18**, 491-496.

Cohen, J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213-220.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.

Collins, E. J., Hogan, T. P. & Himansu, D. (1991) Measurement of therapeutic response in schizophrenia. *Schizophrenia Research* **5**, 249-253.

Cook, T. D. & Campell, D. T. (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Cooper, H. & Hedges, L. V. (1994) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Cooper, J. E., Kendall, R. E., Gurland, B. J., Sharpe, L., et al (1972) *Psychiatric Diagnosis in New York and London*. London: Oxford University Press.

Counsell, C. (1998) Formulating questions and locating primary studies for inclusion in systematic reviews. In *Systematic Reviews: synthesis of best evidence for healthcare decisions* (eds C. D. Mulrow & D. Cook). Philadelphia: American College of Physicians.

Croghan, T., Melfi, C., Dobrez, D. & Kniesner, T. (1999) Effect of mental health specialty care on antidepressant length of therapy. *Medical Care* **37,** (4 Suppl Lilly), AS20-3.

Crombie, I. K. & Davies, H. T. O. (1997) Beyond health outcomes: the advantages of measuring process. *Journal of Evaluation in Clinical Practice* **4**, 31-38.

Curtis, R. & Beevor, A. (1995) Health of the Nation Outcome Scales (HoNOS). In *Measurement for Mental Health: contributions from the College Research Unit* (ed J. Wing). London: Royal College of Psychiatrists.

Davies, A. E., Doyle, M. A. & Lansky, D. (1994) Outcomes assessment in clinical settings: consensus statement on principles and best practices in project management. *Journal of Quality Improvement* **20**, 6-16.

Davies, H. (1997) What's a healthy outcome? *Health Service Journal*, (10th April), 22.

Davies, H. T. O. & Crombie, I. K. (1997) Interpreting Health Outcomes. *Journal of Evaluation in Clinical Practice* **3**, 187-199.

Davies, H. T. O. & Lampel, J. (1998) Trust in performance indicators. *Quality In Health Care* **7**, 159-162.

Delamonthe, T. (1994) Using outcomes research in clinical practice. *British Medical Journal* **308**, 1583-1584.

Denzin, N. & Lincoln, Y. (Eds.). (1994) *Handbook of Qualitative Research*. London: Sage Publications.

Department of Health (1989) *Working for Patients*. London: HMSO.

Department of Health (1991) *The Health of the Nation: a strategy for England*. London: HMSO.

Department of Health (1992) *Assessing the Effects of Health Technologies: Principles, Practice and Proposals. Advisory Group on Health Technologies*. London: HMSO.

Department of Health (1998) *Modernising Mental Health Services: safe, sound and supportive*. London: HMSO.

Department of Health (2000) Medical and Dental Workforce - detailed statistics 2000. http://www.doh.gov.uk/public/stats3.htm#workforce (Accessed.

Depression Guideline Panel (1993) *Depression in primary care: Quick reference guide for clinicians. Clinical practice Guideline Number 5*. Rockville, MD: AHCPR.

Derogatis, L. R. (1994) *Symptom Checklist-90-R (SCL-90-R) administration, scoring and procedures manual* (3rd Edition edn). Minneapolis: National Computer Systems.

DerSimonian, R. & Laird, N. (1986) Meta-analysis in Clinical Trials. *Controlled Clinical Trials* **7**, 177-188.

Deyo, R. A. & Carter, W. B. (1992) Strategies for improving and expanding the application of health status measures in clinical settings. A researcher-developer viewpoint. *Med Care* **30,** (5 Suppl), Ms176-86.

Deyo, R. A. & Patrick, D. L. (1989) Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Med Care* **27,** (3 Suppl), S254-68.

Dickersin, K. (1990) The existence of publication bias and risk factors for its occurrence. *JAMA* **263**, 1385-1389.

Dickersin, K., Hewitt, P., Mutch, L., Chalmers, I., et al (1985) Pursuing the literature: comparisons of MEDLINE searching with a perinatal trials database. *Controlled Clinical Trials* **6**, 306-17.

Dickersin, K. & Larson, K. (1996) Establishing and maintaining an international register of RCTs. In *The Cochrane Library*. Oxford: Update Software.

Dilman, D. (1991) The design and administration of mail surveys. *Annual Review of Sociology* **17**, 225-249.

Divine, G. W., Brown, J. T. & Frazer, L. M. (1992) The unit of analysis error in studies about physicians' patient care behaviour. *Journal of General Internal Medicine* **7**, 623-629.

Doessel, D. P. & Marshall, J. V. (1985) A rehabilitation of health outcome in quality assessment. *Soc Sci Med* **21,** (12), 1319-28.

Donabedian, A. (1966) Evaluating the quality of medical care. *Milbank Mem Fund Q* **44,** (3), Suppl:166-206.

Donabedian, A. (1980) *Explorations in Quality Assessment and Monitoring.* (Vol. Volume 1. The Definition of Quality and Approaches to its Assessment.). Ann Arbour, MI.: Health Administration Press.

Donabedian, A. (1989) The end result of health-care: Ernest Codman's contribution to quality assessment and beyond. *The Millbank Quarterly* **67**, 233-56.

Donoghue, J., Tylee, A. & Wildgust, H. (1996) Cross sectional database analysis of antidepressant prescribing in general practice in the United Kingdom, 1993-5. *British Medical Journal* **313**, 861-2.

Dorwick, C. (1995) Does testing for depression influence diagnosis or management by general practioners? *Family Practice* **12**, 461-465.

Dowrick, C. & Buchan, I. (1995) Twelve month outcome of depression in general practice: does detection or disclosure make a difference? *British Medical Journal* **311**, 1274-1276.

Drummond, M. F., O'Brien, B., Stoddard, G. L. & Torrance, G. W. (1997) *Methods for the Economic Evaluation of Health Care Programmes.* (Vol. 2nd). Oxford: Oxford University Press.

Duggan, L., Fenton, M., Dardennes, R. M., El-Dosoky, A., et al (2001) Olanzapine for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Dunn, G. (1996) Statistical methods for measuring outcomes. In *Mental Health Outcome Measures* (eds G. Thornicroft & M. Tansella). Berlin: Springer Verlag.

Ebrahim, S. (1995) Clinical and public health perspectives and applications of health-related quality of life measurement. *Soc Sci Med* **41,** (10), 1383-94.

Egger, M. & Davey-Smith, G. (1995) Misleading meta-analysis: lessons from "an effective, safe, simple" intervention that wasn't. *BMJ* **310**, 752-754.

Egger, M., Davey-Smith, G., Schneider, M. & Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629-634.

Eisele, W., Slee, V. N. & Hoffman, R. G. (1956) Can the practice of internal medicine be audited? *Annals of Internal Medicine* **44**, 144-166.

Elkington, J. (1966) Medicine and the quality of life. *Annals of Internal Medicine* **64**, 711-714.

Ellwood, P. M. (1988) Shattuck lecture - outcomes management. A technology of patient experience. *New England Journal of Medicine* **318,** (23), 1549-56.

Enndicot, J., Spitzer, R. L., Fleis, J. L. & Cohen, J. (1976) The global assessment scale: a procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry* **33**, 766-771.

Epstein, A. M. (1990) The outcomes movement--will it get us where we want to go? *N Engl J Med* **323,** (4), 266-70.

Faden, R. & Leplege, A. (1992) Assessing quality of life. Moral implications for clinical practice. *Med Care* **30,** (5 Suppl), Ms166-75.

Farquar, M. (1995) Definitions of quality of life: a taxonomy. *Journal of Advanced Nursing* **22**, 502-508.

FDA (1997) *The Food and Drug Administration Modernisation Act*. Rockville: FDA.

Feinstein, A. (1967) *Clinical Judgement*. Baltimore, MD: Williams and Wilkins.

Feinstein, A. R. (1995) Meta-analysis: statistical alchemy for the 21st century. *Journal of Clinical Epidemiology* **48**, 71-9.

Feldman, E., Mayou, R., Hawton, K., Ardern, M., et al (1987) Psychiatric disorders in medical in-patients. *Quarterly Journal of Medicine* **63**, 405-412.

Felson, D. T., Anderson, J. J., Boers, M., Bombardier, C., et al (1993) The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis Rheum* **36**, 729-740.

Fienberg, S. E., Loftus, E. F. & Tanur, J. M. (1985) Cognitive aspects of health survey methodology: An overview. *Health and Society* **63,** (3), 547-564.

Fifer, S. K., Mathias, S. D., Patrick, D. L., Mazonson, P. D., et al (1994) Untreated anxiety among adult primary care patients in a health maintainance organisation. *Archives of General Psychiatry* **51**, 740-750.

Financial Times Healthcare (2000) *The Medical Directory on CD-ROM*. London: FT Pharmaceuticals.

Fitzpatrick, R. (1994) Applications of health status measures. In *Measuring health and medical outcomes* (ed C. Jenkinson). London: UCL Press.

Fitzpatrick, R., Davey, C., Buxton, M. J. & Jones, D. R. (1998) Evaluating patient based outcome measures for use in clinical trials. *Health Technology Assessment* **2,** (17).

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., et al (1992a) Quality of life measures in health care. I: Applications and issues in assessment. *BMJ* **305,** (6861), 1074-7.

Fitzpatrick, R., Hinton, J., Newman, S., Scambler, G., et al (Eds.). (1984) *The Experience of Illness*. London: Tavistock Publications.

Fitzpatrick, R., Ziebland, S., Jenkinson, C., Mowat, A., et al (1992b) Importance of sensitivity to change as a criterion for selecting health status measures. *Quality in Health Care* **1**, 89-93.

Folstein, M. F., Folstein, S. E. & McHugh, P. R. (1975) 'Mini Mental State': a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**, 189-98.

Fowler, F. J. (1993) *Survey Research Methods* (Second Edition edn). California: SAGE.

Frater, A. (1992) Health outcomes: a challenge to the status quo. *Quality in Health Care* **1**, 87-88.

Frater, A. & Costain, T. (1992) Any better? Outcome measures in medical audit. *British Medical Journal* **304**, 519-520.

Fuchs, V. (1974) *Who shall live?* New York: Basic Books.

Galbraith, J. K. (1958) *The Affluent Society*. New York.

Gardiner, M. J. & Altman, D. G. (1989) *Statistics With Confidence*. London: BMJ Publishing.

Geddes, J., Freemantle, N. & Harrison, P. (2000) Atypical anti-psychotics in the treatment of schizophrenia: systematic overview and meta-regression. *British Medical Journal* **321**, 1371-1376.

Geddes, J. R., Freemantle, N., Mason, J., Eccles, M. P., et al (2001) Selective serotonin reuptake inhibitors (SSRIs) for depression (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Geigle, R. & Jones, S. B. (1990) Outcomes measurement: a report from the front. *Inquiry* **27,** (1), 7-13.

General Accounting Office (1992) *Cross design synthesis: a new strategy for medical effectiveness research*. Washington, DC.: GAO.

German, P. S., Shapiro, S. & Skinner, E. A. (1987) Detection and management of mental health problems of older patients by primary care providers. *Journal of the American Medical Association* **257**, 489-496.

Gilbody, S. & Sowden, A. (1999) Systematic reviews in mental health. In *Evidence Based Counselling and Psychological Therapies* (eds N. Roland & S. Goss). London: Routledge.

Gilbody, S. M., Bagnall, A. M., Duggan, L. & Tuunainen, A. (2001a) Risperidone versus other atypical antipsychotic medication for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Gilbody, S. M., House, A. O. & Sheldon, T. A. (2001b) Outcome and needs assessment for schizophrenia and related disorders (Cochrane Review). In *The Cochrane Library - Issue 1*. Oxford: Update Software.

Gilbody, S. M., House, A. O. & Sheldon, T. A. (2001c) Routine outcomes assessment to improve the detection and management of depression, anxiety and related disorders (Cochrane Review). In *The Cochrane Library - Issue 1*. Oxford: Update Software.

Gilbody, S. M., House, A. O. & Sheldon, T. A. (2001d) Routinely administered questionnaires for depression and anxiety: a systematic review. *British Medical Journal* **322**, 406-409.

Gilbody, S. M. & Petticrew, M. (1999) Rational decision making in mental health: the role of systematic reviews in clinical and economic evaluation. *Journal of Mental Health Policy and Economics* **2**, 99-107.

Gilbody, S. M. & Song, F. (2000) Publication bias and the integrity of psychiatry research. *Psychological Medicine* **30**, 253-258.

Gilbody, S. M., Song, F., Eastwood, A. J. & Sutton, A. J. (2000) The causes, consequences and detection of publication bais in psychiatry. *ACTA Psychiatrica Scandinavica* **102**, 241-249.

Gilbody, S. M., Wahlbeck, K. & Adams, C. E. (2001e) Randomised controlled trials in schizophrenia: a critical review of the literature. *ACTA Psychiatrica Scandinavica* **In Press**.

Gilbody, S. M. & Whitty, P. A. (2001) Improving the delivery and organisation of mental health services: beyond the conventional RCT. *British Journal of Psychiatry* **Article In Press**.

Gill, T. M. & Feinstein, A. R. (1994) A critical appraisal of quality of life measures. *Journal of the American Medical Association* **272**, 619-626.

Glanze, W. D. (Ed.). (1990) *Mosby's Medical, Nursing, and Allied Health Dictionary*. St Louis: Mosby Co.

Glover, G., Knight, S., Melzer, D. & Pearce, L. (1997) The development of a new minimum data set for specialist mental health care. *Health Trends* **29**, 48-51.

Glover, J. A. (1938) The incidence of tonsillectomy in school children. *Proceedings of the Royal Society of Medicine* **xxxi**, 1219-36.

Glover, J. A. (1948) The paediatric approach to tonsilectomy. *Archives of Diseases of Childhood*,, 1-6.

Gold, I. & Baraff, L. J. (1989) Psychiatric screening in the emergency department: its effect on physician behaviour. *Annals of Emergency Medicine* **18**, 875-880.

Goldberg, D. (1972) *The Detection of Psychiatric Illness by Questionnaire*. Oxford: Oxford University Press.

Goldberg, D. (1986) The use of the general health questionnaire in clinical work. *British Medical Journal* **293**, 1188-1189.

Goldberg, D. & Bridges, K. (1988) Somatic presentation of psychiatric illness in primary care setting. *Journal of Psychosomatic Research* **32**, 137-144.

Goldberg, D., Eastwood, M. R. & Kedwood, H. B. (1970) A standardised pscyhiatric interview for use in community surveys. *British Journal of Preventative and Social Medicine* **24**, 18-23.

Goldberg, D. & Huxley, P. (1980) *Mental Illness in the Community*. London: Tavistock.

Goldberg, D. P. & Williams, P. (1988) *The user's guide to the General Health Questionnaire*. Windsor: NFER-Nelson.

Goldsmith, G. & Brodwick, M. (1989) Assessing the functional status of older patients with chronic illness. *Family Medicine* **21**, 38-41.

Golligher, J. C. (1987) Judging the quality of life after surgical operations. *Journal of Chronic Diseases* **40**, 631-633.

Goyder, J. (1987) *The silent minority: Non-respondents on sample surveys*. Cambridge: Polity Press.

Green, S. B. & Byar, D. P. (1984) Using observational data from registries to compare treatments. *Statistics in Medicine* **3**, 351-370.

Greenfield, S. & Nelson, E. C. (1992) Recent developments and future issues in the use of health status assessment measures in clinical settings. *Med Care* **30,** (5 Suppl), Ms23-41.

Greenland, S. (1994) Quality scores are useless and potentially misleading. *American Journal of Epidemiology* **140**, 300-301.

Grimshaw, J. M. & Russell, I. T. (1993) Effect of clinical guidelines on medical practice. A systematic review of rigorous evaluations. *Lancet* **342**, 1317-22.

Guthrie, E. (2000) Psychotherapy for patients with complex disorders and chronic symptoms: The need for a new research paradigm. *British Journal of Psychiatry* **177**, 131-137.

Guyatt, G., Feeny, D. & Patrick, D. (1991) Issues in quality-of-life measurement in clinical trials. *Control Clin Trials* **12,** (4 Suppl), 81s-90s.

Guyatt, G. H., Feeny, D. H. & Patrick, D. L. (1993a) Measuring health-related quality of life. *Ann Intern Med* **118,** (8), 622-9.

Guyatt, G. H., Sackett, D. L. & Cook, D. J. (1993b) Users' guides to the medical literature. II. How to use articles about Therapy or Prevention. Evidence-Based Medicine Working Group. *Journal of the American Medical Association* **270**, 1232-7.

Guyatt, G. H., Veldhuyzen Van Zanten, S. J., Feeny, D. H. & Patrick, D. L. (1989) Measuring quality of life in clinical trials: a taxonomy and review. *Cmaj* **140**, (12), 1441-8.

Hall, J., Masters, G., Tarlo, K. & Andrews, G. (1984) *Measuring outcomes of health services*. Westmead Centre, NSW: Department of Community Medicine.

Hall, J., Masters, G., Tarlo, K. & Andrews, G. (1986) *Report to the National Committee on Health and Vital Statistics on outcome data in health*. Canberra: AGPS.

Hamilton, M. (1967) Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology* **6**, 278-296.

Hamilton, M. (1972) Rating scales in depression. In *Depressive Illness* (ed P. Kielholtz). Berlin: Huber.

Hamilton, M. (1976) The role of rating scales in psychiatry. *Psychological Medicine* **6**, 347-49.

Harvey, L. (1987) Factors affecting response rates to mailed questionnaires: a comprehensive literature review. *Journal of the Market Research Society* **29**, 341-353.

Hay, P. J., Adams, C. E. & Lefebre, C. (1996) The efficiency of searches for randomised trials in the International Journal of Eating Disorders. A comparison of handsearching, EMBASE and PsycLIT. *Health Libraries Review* **13**, 91-96.

Hays, R. D., Sherbourne, C. D. & Mazel, R. M. (1993) The RAND 36-Item Health Survey 1.0. *Health Econ* **2,** (3), 217-27.

Hedges, L. V. (1982) Estimation of effect size from a series of independent experiments. *Psychological Bulletin* **92**, 490-499.

Heinrichs, D. W., Hanlon, E. T. & Carpenter, W. T. (1984) The quality of life scale: an instrument for rating the schizophrenic deficit syndrome. *Schizophrenia Bulletin* **10**, 388-98.

Herberlein, T. A. & Baumgartner, R. (1978) Factors affecting response rates to questionnaires: A quantitative analysis of the published literature. *American Sociology Review* **43**, 447-462.

Higginson, I. (1994) Quality of care and evaluating services. *International Review of Psychiatry.* **6**, 15-14.

Hoeper, E. W., Nycz, G. R., Kessler, J. D. & Pierce, W. E. (1984) The usefulness of screening for mental illness. *Lancet* **1**, 33-35.

Hong, W. W., Rak, I. W., Ciuryla, V. T., Wilson, A. M., et al (1998) Medical-claims databases in the design of a health-outcomes comparison of quetiapine ('Seroquel') and usual-care antipsychotic medication. *Schizophrenia Research* **32**, 51-8.

Hopkins, A. & Constantin, D. (Eds.). (1990) *Measuring the Outcomes of Medical Care*. London: Royal College of Physicians.

Hotopf, M., Churchill, R. & Lewis, G. (1999) Pragmatic randomised trials in psychiatry. *British Journal of Psychiatry* **175**, 217-223.

Hotopf, M., Lewis, G. & Normand, C. (1997) Putting trials on trial--the costs and consequences of small trials in depression: a systematic review of methodology. *J Epidemiol Community Health* **51,** (4), 354-8.

House of Commons (1990) *The National Health Service and Community Care Act*. London: HMSO.

Hunt, S. M., McEwen, J. & McKenna, S. P. (1985a) Measuring Health Status: a new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practioners* **35**, 185-188.

Hunt, S. M., McEwen, J. & McKenna, S. P. (1985b) Measuring health status: a new tool for epidemiologists and clinicians. *Journal of the Royal College of General Practionners* **35**, 185-188.

Hylan, T., Crown, W., Meneades, L., Heiligenstein, J., et al (1999) SSRI antidepressant drug use patterns in the naturalistic setting: a multivariate analysis. *Medical Care* **37,** (4 Suppl Lilly), AS36-44.

Iezzoni, L. I. (1997) Assessing quality using administrative data. *Annals of Internal Medicine* **127**, 666-674.

Jadad, A. R., Moore, R. A. & Carroll, D. (1996) Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* **17**, 1-12.

Jencks, S. F. (1985) The recognition of mental distress and diagnosis of mental disorder in primary care. *Journal of the American Medical Association* **253**, 1903-1907.

Jenkinson, C. (1994) Measuring health and medical outcomes: an overview. In *Measuring health and medical outcomes* (ed C. Jenkinson). London: UCL Press.

Jenkinson, C. (1995) Evaluating the efficacy of medical treatments: possibilities and limitations. *Soc. Sci. Med.* **41**, 1395-1401.

Jette, A. M., Davies, A. R. & Calkins, D. R. (1986) The Functional Status Questionnaire: its reliability and validity when used in primary care. *Journal of General Internal Medicine* **1**, 143-149.

Johnstone, A. & Goldberg, D. (1976) Psychiatric screening in General Practice. *Lancet* **1**, 605-612.

Jordan, J., Wright, J., Wilkinson, J. & Williams, R. (1998) Assessing local health authorities needs in primary care: understanding and experience in three English districts. *Quality in Health Care* **7**, 83-89.

Juni, P., Witschi, A., Bolsch, R. & Egger, M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* **282**, 1054-1060.

Kahn, K. L., Kosecoff, J., Chassin, M. R., Flynn, M. F., et al (1988) Measuring the clinical appropriateness of the use of a procedure. Can we do it? *Med Care* **26,** (4), 415-22.

Kane, R. A. & Kane, R. L. (1981) *Assessing the Elderly: A Practical Guide to Measurement*. Toronto: Lexington Books.

Kane, R. L. (1987) Functional assessment for geriatric patients - or the clinical Swiss army knife. *Medical Care* **supplement**, S95-99.

Kaplan, R. M. & Coons, S. J. (1992) Relative importance of dimensions of health related quality of life for patients with hypertension. *Progress in Cardiovascular Nursing* **7**, 29-36.

Karnofsky, D. A., Ablemann, W. H. & Craver, L. F. (1948) The use of the nitrogen mustrds in the paliative treatement of carcinoma. *Cancer* **1**, 634-656.

Katching, H. (1983) Methods for measuring social adjustment. In *Methods in Evaluation of Psychiatric Treatment* (ed T. Helgason), pp. 205-208. Cambridge: Cambridge University Press.

Katz, M. M. & Lyerly, S. B. (1963) Methods for measuring adjustment and social behaviour in the community. 1. Rationale, discriminative validity and scale development. *Psychological Reports* **13**, 503-535.

Katz, S., Ford, A. B. & Moskovitz, R. W. (1963) Studies of illness in the aged: a standardised measure of biological and social function. *Journal of the American Medical Association* **185**, 914-919.

Kay, S. R. (1991) *Positive and negative syndromes in schizophrenia*. New York: Brunner-Mazel.

Kazis, L. E., Anderson, J. J. & Meenan, R. F. (1988) Health status information in clinical practice: the development and testing of patient profile reports. *J Rheumatol* **15,** (2), 338-44.

Kazis, L. E., Callahan, L. F., Meenan, R. F. & Pincus, T. S. O. (1990) Health status reports in the care of patients with rheumatoid arthritis. *Journal of Clinical Epidemiology* **43**, 1243-53.

Kelly, H., Russell, E. M., Stewart, S. & McEwan, J. (1996) Needs assessment: taking stock. *Health Bulletin* **54**, 115-18.

Kennedy, E., Song, F. & Gilbody, S. (2001) Risperidone versus typical antipsychotic medication for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Kindig, D. A. (1977) *Purchasing Population Health*. Ann Arbour, MI.: The University of Mitchigan Press.

Klein, R. (1995) *The New Politics of the NHS*. (Vol. Third). London: Longman.

Kmietowicz, Z. (2001) Registries will have to apply for right to collect patients' data without consent. *British Medical Journal* **322**, 1199.

Lakahni, A. (1994) *Central Health Outcomes Unit*. London: Department of Health.

Lam, J. A. & Rosenheck, R. (1999) Street outreach for homeless persons with serious mental illness: is it effective? *Medical Care* **37**, 894-907.

Larsen, D., Attkinson, C. C. & Hargreaves, W. A. (1979) Assessment of clinet/patient satisfaction: development of a general scale. *Evaluation and Programme Planning* **2**, 197-207.

Last, J. M. (Ed.). (1994) *A Dictionary of Epidemiology*. (Vol. 3rd). Oxford: Oxford University Press.

Lehman, A. F. (1983a) The effect of psychiatric symptoms on quality of life assessments among the chronically menatally ill. *Evaluative Programme Planning* **6**, 143-151.

Lehman, A. F. (1983b) The well being of chronic mental patients: assessing their quality of life. *Archive of General Pscyhiatry* **40**, 369-373.

Lehman, A. F. (2001) Measures of quality of life for people with severe mental disorders. In *Mental Health Outcome Measures* (eds M. Tansella & G. Thornicroft), Second edn. London: Gaskell.

Lembke, P. A. (1952) Measuring the quality of medical care through vital statistics based on hospital service areas: 1 comparative study of appendicectomy rates. *American Journal of Public Health* **42**, 276-86.

Leslie, D. L. & Rosenheck, R. A. (2000) Comparing quality of mental health care for public sector and privately insured populations. *Psychiatric Services* **51**, 650-655.

Lewis, G., Sharp, D., Bartholomew, J. & Pelosi, A. J. (1996) Computerized assessment of common mental disorders in primary care: effect on clinical outcome. *Family Practice* **13**, 120-6.

Lewis, R., Bagnall, A.-M. & Leitner, M. L. (2001a) Sertindole for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Lewis, R., Bagnall, A.-M. & Leitner, M. L. (2001b) Ziprasidone for schizophrenia and severe mental illness (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Liang, M. H. & Katz, J. N. (1992) Measurement of outcome in rheumatoid arthritis. *Baillieres Clin Rheumatol* **6,** (1), 23-37.

Light, R. J. & Pillemer, D. B. (1984) *Summing up: the science of reviewing research.* Cambridge, Massachusetts, and London: Harvard University Press.

Lilford, R. & Royston, G. (1998) Decision analysis in the selection, design and application of clinical and health services research. *Journal of Health Services Research* **3,** (3), 159-166.

Linn, L. S. & Yager, J. (1980a) The effect of screening, sensitisation and feedback on notation of depression. *Journal of Medical Education* **20**, 942-953.

Linn, L. S. & Yager, J. (1980b) Screening for depression in relationship to subsequent patient and physician behaviour. *Medical Care* **20**, 1233-1245.

Lohr, K. N. (1988) Outcome measurement: concepts and questions. *Inquiry* **25,** (1), 37-50.

Long, A., Dickson, P., Hall, R., Carr-Hill, R., et al (1993) The outcomes agenda. *Quality in Health Care* **2**, 154-87.

Long, A. L. & Dixon, P. (1996) Monitoring outcomes in routine practice: defining appropriate measurement criteria. *Journal of Evaluation in Clinical Practice* **2**, 71-78.

Luborsky, L. (1962) Clinicians judgements of mental health. *Archives of General Psychiatry* **7**, 407-417.

Magruder Habib, K., Zung, W. W. & Feussner, J. R. (1990) Improving physicians' recognition and treatment of depression in general medical care. Results from a randomized clinical trial. *Medical Care* **28,** (3), 239-250.

Makover, H. B. (1951) The quality of medical care: methodological survey of the medical groups associated with the Health Insurance Plan of New York. *American Journal of Public Health* **41**, 824-829.

Mannion, R. & Goddard, M. (2001) Impact of published clinical outcomes data: case study of NHS hospital trusts. *British Medical Journal* **323**, 260-264.

Mant, D. & Fowler, G. (1990) Mass screening: theory and ethics. *British Medical Journal* **300**, 916-18.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from reterospective studies of disease. *Journal of the National Cancer Institute* **22**, 719-730.

Marginson, F. R., McGrath, G., Barkham, M., Mellor Clark, J., et al (2000) Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *British Journal of Psychiatry* **177**, 123-130.

Mari, J. J. & Streiner, D. L. (1994) An overview of family interventions and relapse on schizophrenia: meta-analysis of research findings. *Psychol Med* **24,** (3), 565-78.

Marks, I. S. O. (1998) Overcoming obstacles to routine outcome measurement. The nuts and bolts of implementing clinical audit. *British Journal of Psychiatry* **173**, 281-286.

Marshall, M., Gray, A., Lockwood, A. & Green, R. (2001) Case management for people with severe mental disorders (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Marshall, M. & Lockwood, A. (2001) Assertive community treatment for people with severe mental disorders (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., et al (2000) Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry* **176**, 249-252.

Mathias, S. D., Fifer, S. K., Mazonson, P. D., Lubeck, D. P., et al (1994) Necessary but not sufficient: the effect of screening and feedback on outcomes of primary care patients with untreated anxiety. *Journal of Internal Medicine* **9**, 606-615.

Mayfield, D., Millard, G. & Hall, P. (1974) The CAGE questionnaire. *American Journal of Psychiatry* **131**, 1121-3.

Mays, N. & Pope, C. (Eds.). (1996) *Qualitative Research in Health Care*. London: BMJ Publishing.

Mazonson, P. D., Mathias, S. D., Fifer, S. K., Beusching, D. P., et al (1994) The mental health patient profile: does it change primary care physicians practice patterns? *Journal of the American Board of Family Practice* **9**, 336-345.

McCallum, J. (1993) What is an outcome and why look at them? *Critical Public Health* **4**, 4-6.

McColl, E., Jacoby, A., Tomas, L., Souter, J., et al (2001) The conduct and design of questionnaire surveys in healthcare research. In *The Advanced Handbook of Methods in Evidence Based Healthcare* (eds A. Stevens, K. Abrams, J. Brazier, et al.). London: Sage.

McDaniel, R. W. & Bach, C. A. (1994) Quality of life: a concept. *Nursing Research* **3**, 18-22.

McDowell, I. & Jenkinson, C. (1996) Developing standards for health measures. *Journal of Health Services Research and Policy* **1**, 238-246.

McDowell, I. & Newell, C. (1996) *Measuring Health: A guide to rating scales and questionnaires*. Oxford: Oxford University Press.

McGuire, H. (1998) The Cochrane Depression Anxiety and Neurosis Group's Register of Randomised Controlled Trials. In *The Cochrane Library*. Oxford: Update Software.

McHorney, C. A. & Tarlov, A. R. (1994) The use of health status measures for individual patient level applications: problems and prospects. *Quality of Life Research* **3**, 43-44.

McHorney, C. A. & Ware, J. E., Jr. (1995) Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. *Med Care* **33,** (1), 15-28.

McHorney, C. A., Ware, J. E., Jr. & Raczek, A. E. (1993) The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* **31,** (3), 247-63.

McKibbon, K. A. & Walker-Dilks, C. J. (1995) The quality and impact of MEDLINE searches performed by end users. *Health Libraries Review* **12**, 91-200.

Meakin, C. J. (1992) Screening for depression in the medically ill. *British Journal of Psychiatry* **160**, 212-216.

Medical Research Council (2000) *Personal information in medical research*. London: Medical Research Council.

Meenan, R. F. (1982) The AIMS approach to health status measurement: Conceptual background and measurement properties. *Journal of Rheumatology* **9**, 785-788.

Meenan, R. F., Anderson, J. J., Kazis, L. E., Egger, M. J., et al (1984) Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum* **27,** (12), 1344-52.

Meinert, C. L. (1986) *Clinical Trials: Design, conduct and analysis*. Oxford: Oxford University Press.

Melfi, C., Chawla, A., Croghan, T., Hanna, M., et al (1998) The effects of adherence to antidepressant treatment guidelines on relapse and recurrence of depression. *Archives of General Psychiatry* **55,** (12), 1128-32.

Mellor-Clarke, J., Barkham, M., Connell, J. & Evans, C. (1999) Practice based evidence and the need for a standardised evaluation system: Informing the design of the CORE system. *European Journal of Psychotherapy, Counsellling and Health* **3**, 357-374.

Meltzer, H. Y. (1999) Outcome in schizophrenia: beyond symptom reduction. *Journal of Clinical Psychiatry* **60 (suppl 3)**, 3-8.

Melzack, R. (1975) The McGill pain questionnaire: major properties and scoring methods. *Pain* **1**, 277-299.

Microsoft Corporation (1997a) Microsoft Excel (Version 2000): Microsoft.

Microsoft Corporation (1997b) Microsoft Publisher (Version 2): Microsoft.

Microsoft Corporation (1998) Microsoft Access 97 (Version 3): Microsoft.

Milio, N. (1983) *Primary care and the publics health*. Lexington, MA.: Lexington Books.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., et al (1999a) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* **354**, 1896-8.

Moher, D., Cook, D. J., Jadad, A. R., Tugwell, P., et al (1999b) Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technology Assessment* **3,** (12).

Moher, D., Jadad, A. R. & Tugwell, P. (1996) Assessing the quality of randomized controlled trials. Current issues and future directions. *International Journal of Technology Assessment in Health Care* **12**, 195-208.

Moher, D., Pham, B., Jones, A., Cook, D. J., et al (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* **352**, 609-613.

Moore, J. T., Silimperi, D. R. & Bobula, J. A. (1978) Recognition of depression by family medicine residents: the impact of screening. *Journal of Family Practice* **7**, 509-513.

Mor, V. & Guadagnoli, E. (1988) Quality of life measures: a psychometric tower of Babel. *Journal of Clinical Epidemiology* **41**, 1055-1058.

Moser, K. & Kalton, G. (1971) *Survey methods of social investigation*. London: Gower and Aldershot.

Mosteller, F. (1989) Finale panel: comments on advances in health status measurement. *Medical Care* **27 (suppl)**, 282-286.

Mulrow, C. D. (1987) The medical review article: state of the science. *Ann Intern Med* **106**, 485 - 488.

Mulrow, C. D. & Oxman, A. D. (1999) Cochrane Collaboration Handbook [updated June 1999]. In *The Cochrane Library [database on disk and CDROM]* (ed T. C. Collaboration). Oxford: Update Software.

National Committee on Quality Assurance (1995) *National Committee on Quality Assurance report card pilot project*. Washington DC: NCQA.

National Institute of Mental Health (1987) *Toward a model plan for a comprehensive community based mental health system*. Washington, DC.: National INstitute of Mental Health.

Naylor, C. D. (1989) Meta-analysis of controlled trials. *Journal of Rheumatology* **16**, 424-426.

Naylor, C. D. (1995) Grey zones of clinical practice: some limits to evidence based medicine. *Lancet* **345**, 840-842.

Nelson, E. & Berwick, D. (1989) The measurement of health status in clinical practice. *Medical Care* **27**, S77-S90.

Nelson, E. C., Landgraf, J. M., Hays, R. D., Wasson, J. H., et al (1990) The functional status of patients: how can it be measured in physician's offices? *Medical Care* **28**, 1111-1126.

NHS Centre for Reviews and Dissemination (1994) *Implementing Practice Guidelines*. (Vol. 1). York: University of York.

NHS Centre for Reviews and Dissemination (1999a) *Drug treatments for schizophrenia*. (Vol. 5). York: University of York.

NHS Centre for Reviews and Dissemination (1999b) *Getting evidence into practice*. (Vol. 4). York: University of York.

NHS Centre for Reviews and Dissemination (2000) *Undertaking Systematic Reviews of Research on Effectiveness: CRD report 4 (second edition)*. York: University of York.

Nightingale, F. (1863) *Notes on Hospitals*. London: Longman, Roberts and Green.

NSW Health Department (1992) The NSW Health Outcomes Program, New South Wales. *Public Health Bulletin* **3**, 12.

Nunnally, J. (1967) *Psychometric Theory* (Second Edition edn). New York: McGraw Hill.

Office of Health Technology Assessment - US Congress (1994) *Identifying Health Technologies that Work: Searching for Evidence OTA-H-608*. Washington DC: US Government Printing Office.

O'Leary, D. (1987) The Joint Commision looks to the future. *JAMA* **258**, 951-2.

Oliver, J., Huxley, P., Bridges, P. & Mohamid, H. (1996) *Quality of Life and Mental Health Services*. London: Routeledge.

Opit, L. J. (1990) The measurement of health service outcomes. In *Oxford Textbook of Public Health Medicine* (eds W. H. Holland, R. Detels & G. Knox), Vol. 3, pp. 159-172. Oxford: Oxford medical Publications.

Oppenheim, A. N. (1992) *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Orley, J., Saxena, S. & Herrman, H. (1998) Quality of life and mental illness. *British Journal of Psychiatry* **172**, 291-293.

Overall, J. E. & Gorham, D. R. (1962) The brief psychiatric rating scale. *Psychological Reports* **10**, 799-812.

Parkerson, G., Broadhead, W. & Chiu-Kit, J. (1990) The Duke Health Profile: a 17 item measure of of health and dysfunction. *Medical Care* **28**, 1056-1069.

Parsons, T. (1951) *The Social System*. Glencoe, IL: Free Press.

Patrick, D. & Erickson, P. (1993) *Health Status and Health Policy*. New York: Oxford University Press.

Patrick, D. L. & Bergner, M. (1990) Measurement of health status in the 1990s. *Ann Rev Public Health* **11**, 165-83.

Revicki, D. A. & Murray, M. (1994) Assessing health related quality of life outcomes of drug treatements for psychiatric disorders. *CNS Drugs* **1**, 465-476.

Rice, N. & Leyland, A. (1996) Multi-level models: applications to health data. *Journal of Health Services Research and Policy* **3**, 154-64.

Roethlisberger, F. J. & Dickinson, W. J. (1939) *Management and the Worker.* Cambridge, MA: Harvard University Press.

Rosenberg, M. (1979) *Conceiving the Self.* New York: Plenum Press.

Rosenfeld, L. S. (1957) Quality of medical care in hospitals. *American Journal of Public Health* **41**, 856.

Rosenheck, R., Leda, C., Frisman, L. & Gallup, P. (1997) Homeless mentally ill veterans: race, service use, and treatment outcomes. *American Journal of Orthopsychiatry* **67**, 632-638.

Rosenheck, R., Stolar, M. & Fontana, A. (2000) Outcomes monitoring and the testing of new psychiatric treatments: work therapy in the treatment of chronic post-traumatic stress disorder. *Health Services Research* **35**, 133-152.

Rosenheck, R. A. (1996) *Department of Veterans Affairs national mental health programme performance monitoring system: fiscal 1995 report*. West Haven, CT.: Northeast Programme Evaluation Centre.

Rosenheck, R. A., Druss, B., Stolar, M., Leslie, D., et al (1999a) Effect of declining mental health service use on employees of a large corporation. *Health Affairs* **18**, 193-203.

Rosenheck, R. A., Fontanna, A. & Stolar, M. (1999b) Assessing quality of care: administrative indicators and clinical outcomes in post traumatic stress disorder. *Medical Care* **37**, 180-188.

Rosental, R. & Rubin, D. B. (1982) Further meta-analytic procedures for assessing cognitive gender differences *Journal of Educational Psychology*, (74).

Rosser, R. (1983) A history of the development of health outcome indicators. In *Measuring the social benefits of medicine.* (ed G. Teeling Smith). London: OHE.

Rosser, R. (1993) The history of health related quality of life in 10 and 1/2 paragraphs. *Journal of the Royal Society of Medicine* **86**, 315-318.

Rost, K., Smith, G. R., Matthews, D. B. & Guse, B. (1994) The deliberate misdiagnosis of major depression in primary care. *Archives of Family Medicine* **3**, 333-342.

Rowan, K. M. (1994) Intensive Care National Audit and Research Centre: past present and future. *Care of the Critically Ill* **10**, 148-149.

Rubenstein, L. V., Calkins, D. R. & Young, R. T. (1989) Improving patient functioning: a randomised trial of functional disability screening. *Annals of Internal Medicine* **111**, 836-842.

Rubenstein, L. V., McCoy, J. M., Cope, D. W., Barrett, P. A., et al (1995) Improving patient quality of life with feedback to physicians about functional status. *Journal of General Internal Medicine* **10**, 707-614.

Rubin, D. B. (1997) Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757-763.

Ruggeri, M. (2001) Measuring satisfaction with psychiatric services: towards a multidimensional measurement of outcome. In *Mental Health Outcome Measures* (eds M. Tansella & G. Thornicroft), Second edn. London: Gaskell.

Russo, J., Trujillo, C. A., Wingerson, D., Decker, K., et al (1998) The MOS 36-Item Short Form Health Survey: reliability, validity, and preliminary findings in schizophrenic outpatients. *Medical Care* **36**, 752-6.

Sackett, D. L., Haynes, R. B., Guyatt, G. H. & Tugwell, P. (1991) *Clinical Epidemiology: A basic science for clinical medicine*. Boston, MA.: Little, Brown and Company.

Sainfort, F., Becker, M. & Diamond, R. (1996) Judgements of quality of life of individuals with severe mental disorders: Patient self report versus provider perspectives. *American Journal of Psychiatry* **153**, 497-502.

Sanders, C., Egger, M., Donnovan, J., Tallon, D., et al (1998) Reporting on quality of life in randomised controlled trials: bibliographic study. *British Medical Journal* **317**, 1191-1194.

Sartorius, N., Jablensky, A., Korten, A., Ernberg, G., et al (1986) Early manifestations and first contact incidence of schizophrenia in different cultures. *Psychological Medicine* **16**, 909-928.

Schulz, K. F., Chalmers, I., Hayes, R. J. & Altman, D. G. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama* **273**, 408 - 412.

Scroeder, S. A. (1987) Outcome assessment 70 years later: are we ready? *New England Journal of Medicine* **316**, 160-162.

Secretary of State for Health (1999) *National Service Framework - Mental Health*. London: HMSO.

Sederer, L. I. & Dickey, B. (Eds.). (1996) *Outcomes Assessment in Clinical Practice*. Baltimore: Williams and Wilkins.

Selby, P. (1993) Measuring the quality of life of patients with cancer. In *Quality of Life Assessment: Key issues for the 1990s* (eds S. R. Walker & R. M. Rosser). Dordecht: Kluwer Academic.

Seymour, J., Newell, D. & Shiell, A. (1993) *Health outcomes in medical literature*. Paper presented at the 7 years and counting: health beyond 2000, Department of Community Medicine, Westmead Hospital, NSW.

Shanks, J. & Frater, A. (1993) Health status, outcome and attributability. *Quality in Health Care* **2**, 259-62.

Shapiro, S., German, P. S., Skinner, E. A., VonKorf, M., et al (1987) An experiment to change the detection and management of mental morbidity in primary care. *Medical Care* **25**, 327-339.

Sharma, V. K., Wilkinson, G. & Fear, S. (1999) Health of the Nation Outcome Scales: a case study in general psychiatry. *British Journal of Psychiatry* **174**, 395-398.

Sheldon, T. A. (1994) Please bypass the PORT. *British Medical Journal* **309**, 142-143.

Sheldon, T. A. & Faulkner, A. (1996) Vetting new technologies [editorial]. *Bmj* **313,** (7056), 508.

Shepherd, M., Watt, D., Fallon, I. & Smeeton, N. (1989) *The Natural History of Scizophrenia: A five year outcome and prediction in a representative sample of schizophrenics*. Cambridge: Cambridge University Press.

Simon, G. E., VonKorff, M., Heiligenstein, J. H., Revicki, D. A., et al (1996) Initial antidepressant choice in primary care. Effectiveness and cost of fluoxetine vs tricyclic antidepressants. *Journal of the American Medical Association* **275**, 1897-902.

Slade, M., Thornicroft, G. & Glover, G. S. O. (1999) The feasibility of routine outcome measures in mental health. *Social Psychiatry and Psychiatric Epidemiology* **34,** (5), 243-249.

Smith, D. M. (1997) Database research: Is happiness a homongous database? *Annals of Internal Medicine* **127**, 725-756.

Smith, M. L. & Glass, G. V. (1977) Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32**, 752-60.

Smith, P. (1996a) A framework for analysing the measurement of outcome. In *Measuring Outcome in the Public Sector* (ed P. Smith). London.: Taylor and Francis Limited.

Smith, P. (Ed.). (1996b) *Measuring Outcome in the Public Sector*. London.: Taylor and Francis Limited.

Smith, R. (1991) Rationing: the search for sunlight. *British Medical Journal* **303**, 1561-1562.

Song, F., Freemantle, N., Sheldon, T. A., House, A., et al (1993) Selective serotonin reuptake inhibitors: meta-analysis of efficacy and acceptability. *Bmj* **306,** (6879), 683-7.

Spiegelhalter, D. J., Gore, S. M., Fitzpatrick, R., Fletcher, A. E., et al (1992) Quality of life measures in health care. III: Resource allocation. *BMJ* **305,** (6863), 1205-9.

Spitzer, R. L., Gibbon, M., Williams, J. B. W. & Endicott, J. (1996) Global Assessment of Functioning (GAF) Scale. In *Outcomes Assessment in Clinical Practice* (eds L. I. Sederer & B. Dickey). Baltimore: Williams and Wilkins.

Sprangers, M. A. & Aaranson, N. K. (1992) The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *Journal of Clinical Epidemiology* **45**, 743-760.

Srisurapanont, M., Disayavanish, C. & Taimkaew, K. (2001) Quetiapine for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Standing Committee on Postgraduate Medical Education (1989) *Medical Audit - the educational implications*. London: SCOPME.

STATA corporation (1999) STATA (Version 6). Texas: STATA corporation.

Stein, G. S. (1999) Usefulness of the Health of the Nation Outcome Scales. *British Journal of Psychiatry* **174**, 375-377.

Steinwachs, D. M. (1989) Application of health status measures in policy research. *Medical Care* **27,** (3 (suppl)).

Sterling, T. D. (1959) Publication decisions and their possible effects on inferences drawn tests of significance - or vice versa. *Am Stat Assoc J* **54**, 30-34.

Stevens, A. & Gillam, S. (1998) Needs assessment: from theory to practice. *British Medical Journal* **316**, 1448-1452.

Stevens, S. (1951) Mathematics, measurement and psychophysics. In *Handbook of Experimental Psychology* (ed S. Stevens), pp. 1-14. New York: Wiley.

Stewart, A. L., Greenfield, S., Hays, R. D., Wells, K., et al (1989) Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *Journal of the American Medical Association* **262,** (7), 907-13.

Stewert, A. L. & Ware, J. E. (Eds.). (1992) *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Durham, N. C.: Duke University Press.

Streator, S. E. & Moss, J. T. (1997) Non approved usage identification of off label anti-depressant use and cost in a network model HMO. *Drug Benefit Trends* **9**, 42-47.

Street, R. L., Jr., Gold, W. R. & McDowell, T. (1994) Using health status surveys in medical consultations. *Medical Care* **32,** (7), 732-44.

Streiner, D. & Norman, G. (1995) *Health Measurement Scales: A practical guide to thier development and use.* (Vol. 2.). Oxford, UK.: Oxford University Press.

Suerez-Almar, M., Belseck, E., Homick, J., Dorgan, M., et al (2000) Identifying clincial trials in the medical literature with electronic databases: MEDLINE is not enough. *Controlled Clinical Trials* **21**, 476-487.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., et al (1999) Systematic reviews of trials and other studies. *Health Technology Assessment* **2,** (19).

Tait, R. C., Pollard, C. A. & Margolis, R. B. (1987) The pain disability index: psychometric and validity data. *Archives of Physical and Medical Rehabilitation* **68**, 438-441.

Tarlov, A. R., Ware, J. E., Jr., Greenfield, S., Nelson, E. C., et al (1989) The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association* **262,** (7), 925-30.

Testa, M. A. & Nackley, J. F. (1994) Methods for quality-of-life studies. *Annu Rev Public Health* **15**, 535-59.

The Cochrane Controlled Trails Register (2000). In *The Cochrane Library, Issue 4.* Oxford: Update Software.

The Cochrane Controlled Trials Register (2000). In *The Cochrane Library, Issue 4.* Oxford: Update Software.

The Cochrane Database of Systematic Reviews (2000). In *The Cochrane Library, Issue 4.* Oxford: Update Software.

Thier, S. O. (1992) Forces motivating the use of health status assessment measures in clinical settings and related clinical research. *Medical Care* **30,** (5 Suppl), Ms15-22.

Thompson, C. (1989) *The Instruments of Psychiatric Research*. Chichester: John Willey.

Thompson, C., Kinmonth, J., Stevens, L., Peveler, R. C., et al (2000) Effects of a clinical-practice guideline and practice-based education on detection and outcome of depression in primary care: Hampshire Depression Project randomised controlled trial. *Lancet* **355**, 50-57.

Thompson, S. (1995) Why sources of heterogeneity in meta-analysis should be invetsigated. In *Systematic Reviews* (eds I. Chalmers & D. G. Altman). London: BMJ.

Thornicroft, G. (1996) Needs Assessment. In *Mental Health Service Evaluation* (eds H. C. Knunsden & G. Thornicroft). Cambridge: Cambridge University Press.

Thornicroft, G., Brewin, C. & Wing, J. (1992) *Measuring Mental Health Needs*. London: Royal College of Psychiatrists.

Thornicroft, G., Strathdeee, G., Phelan, M., Holloway, F., et al (1998) Rationale and design: PRiSM psychosis study I. *British Journal of Psychiatry* **173**, 363-370.

Thornley, B. & Adams, C. E. (1998) Content and quality of 2000 controlled trials in schizophrenia over 50 years. *British Medical Journal* **317**, 1181-1184.

Timmreck, T. (Ed.). (1992) *Dictionary of Health Services Management*. Owings Mills: National Health Publishing.

Tollefson, G. D., Beasley, C. M., Tran, P. V., Street, J. S., et al (1997) Olanzapine versus haloperidol in the treatment of schizophrenia and schizoaffective and schizophreniform disorders: results of an international collaborative trial. *American Journal of Psychiatry* **154**, 457-465.

Torgerson, D. & Rafferty, J. (1999) Measuring outcomes in economic evaluations. *British Medical Journal* **318**, 1413.

Torrance, G. W. (1987) Utility approach to measuring health related quality of life. *Journal of Chronic Diseases* **40**, 593-600.

Torrance, G. W. & Feeny, D. (1989) Utilities and quality-adjusted life years. *Int J Technol Assess Health Care* **5,** (4), 559-75.

Tugwell, P. & Boers, M. (1993) Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. *J Rheumatol* **20**, 555-556.

Tunis, S. L., Croghan, T. W., Heilman, D. K., Johnstone, B. M., et al (1999) Reliability, validity, and application of the medical outcomes study 36-item short-form health survey (SF-36) in schizophrenic patients treated with olanzapine versus haloperidol. *Mediacl Care* **37**, 678-91.

Tuunainen, A. & Gilbody, S. M. (2001) Newer atypical antipsychotic medication versus clozapine for schizophrenia (Cochrane Review). In *The Cochrane Library, Issue 2*. Oxford: Update Software.

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, A. C., et al (1999a) Methods for evaluating area-wide and organisation based interventions in health and health care: a systematic review. *Health Technology Assessment* **3,** (5).

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A., et al (1999b) Methods in health service research. Evaluation of health interventions at area and organisation level. *British Medical Journal* **319,** (7206), 376-9.

van den Bos, G. A. M. & Triemstra, A. H. M. (1999) Quality of life and as an instrument for need assessment and outcome assessment of health care in chronic patients. *Quality in Health Care* **8**, 247-252.

van Hemert, A. M., Hengeveld, M. W., Bolk, J. H., Rooijmans, H. G., et al (1993) Psychiatric disorders in relation to medical illness among patients of a general medical out-patient clinic. *Psychological Medicine* **23**, 167-173.

von Neumann, J. & Morgernstern, O. (1944) *Theory of games and economic behaviour*. Princeton: Princeton University Press.

Wagner, A. K., Ehrenberg, B. L., Tran, T. A., Bungay, K. M., et al (1997) Patient based health status measurement in clinical practice: a study of its impact in epileps patients. *Quality of Life Research* **6**, 329-341.

Walshe, K. (1995) Opportunities for improving the practice of clinical audit. *Quality in Helath Care* **4**, 231-232.

Ware, J., Brook, R., Davies, A. & al., e. (1981) Choosing measures of health status for individuals in populations. *American Journal of Public Health* **71**, 620-625.

Ware, J. E. (1985) Monitoring and evaluating health services. *Medical Care* **23**, 705-709.

Ware, J. E. (1987) Standards for validating health measures: definition and content. *Journal of Chronic Diseases* **40**, 473-480.

Ware, J. E. (1995) The status of health assessment in 1994. *Annual Review of Public Health* **16**, 327-354.

Ware, J. E., Jr. & Sherbourne, C. D. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* **30,** (6), 473-83.

Ware, J. E., Snoww, K. K., Kosinski, M. & Gandek, B. (1993) *SF-36 Health Survey: Manual and Interpretation Guide*. Boston, MA.: The Health Institute, New England Medical Centre.

Ware, J. E. & Young, J. (1979) Issues in the conceptualisation and measurement of values placed on health. In *Health: What is it worth?* (eds S. J. Mushkin & D. W. Dunlop). New York: Pergamon Press.

Wasson, J., Hays, R., Rubenstein, L., Nelson, E., et al (1992a) The short-term effect of patient health status assessment in a health maintenance organization. *Quality of Life Research* **1,** (2), 99-106.

Wasson, J., Keller, A., Rubenstein, L., Hays, R., et al (1992b) Benefits and obstacles of health status assessment in ambulatory settings. The clinician's point of view. The Dartmouth Primary Care COOP Project. *Med Care* **30,** (5 Suppl), Ms42-9.

Weatherall, M. (2000) A randomized controlled trial of the Geriatric Depression Scale in an inpatient ward for older adults. *Clinical Rehabilitation* **14**, 186-91.

Weissman, M. M. (1975) The assessment of social adjustment. A review of techniques. *Archives of General Psychiatry* **32**, 357-365.

Weissman, M. M. & Bothwell, S. (1976) The assessment of social adjustment by self report. *Archives of General Psychiatry* **33**, 1111-1115.

Wells, K. B. (1999) Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry* **156**, 5-10.

Wells, K. B., Stewart, A., Hays, R. D., Burnam, M. A., et al (1989) The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *Journal of the American Medical Association* **262,** (7), 914-9.

Wells, K. B., Sturm, R., Sherbourne, C. D. & Meredith, L. S. (1996) *Caring for Depression*. Massachusetts: Harvard University Press.

Wennberg, J. (1991) What is outcomes research? In *Health Services Research: Key to health policy* (ed E. Ginzberg), pp. 33-46. Cambridge, Massachusetts: Harvard University Press.

Wennberg, J., Bunker, J. & Barnes, B. (1980) The need for assessing the outcome of medical practice. *Annual Review of Public Health* **1980**, 277-295.

Wennberg, J. & Gittelsohn (1973) Small area variations in health care delivery. *Science* **182,** (117), 1102-8.

Wennberg, J. & Gittelsohn, A. (1982) Variations in medical care among small areas. *Sci Am* **246,** (4), 120-34.

Wennberg, J. E. (1990) Better policy to promote the evaluative clinical sciences. *Qual Assur Health Care* **2,** (1), 21-9.

Wennberg, J. E., Barry, M. J., Fowler, F. J. & Mulley, A. (1993) Outcomes research, PORTs, and health care reform. *Annals of the New York Academy of Sciences* **703**, 52-62.

WHO (1991) *Evaluation of Methods for the Treatment of Mental Disorders*. Geneva: WHO.

Whooley, M. A., Stone, B. & Soghikian, K. (2000) Randomized trial of case- finding for depression in elderly primary care patients. *J Gen Intern Med* **15**, 293-300.

Wiersma, D. (1996) Measuring social disabilities in health. In *Mental Health Outcome Measures* (eds G. Thornicroft & M. Tansella). Berlin: Springer Verlag.

Wilkin, D., Hallam, L. & Doggett, M. (1992) *Measures of needs and outcomes for primary health care*. Oxford: Oxford University Press.

Williams, A. (1985) The economics of coronary artery bypass grafting. *British Medical Journal* **291**, 326-329.

Williams, A. & Kind, P. (1992) The present state of play about QALYs. In *Measures of Quality of Life* (ed A. Hopkins). London: Royal College of Physicians.

Williams, J. W. J., Mulrow, C. D. & Kroenke, K. (1999) Case-finding for depression in primary care: a randomized trial. *American Journal of Medicine* **106**, 36-43.

Williamson, J. (1978) *Assessing and improving healthcare outcomes: the health accounting approach to QA.* Cambridge, MA.: Ballinger.

Wing, J. (1994) Measuring mental health outcomes: a perspective from the Royal College of Psychiatrists. In *Outcomes into clinical practice* (ed T. Delamonthe), pp. 147-152. London: BMJ Publishing.

World Health Organisation (1948) *Constitution*. Geneva: WHO.

Wright, A. (1994a) Should general practitioners be testing for depression? *British Journal of General Practice* **44**, 132-135.

Wright, J., Williams, R. & Wilkinson, J. R. (1998) Development and importance of health needs assessment. *British Medical Journal* **316**, 1310-1313.

Wright, L. (1994b) The long and the short of it: the development of the SF-36 general health survey. In *Measuring health and medical outcomes* (ed C. Jenkinson). London: UCL Press.

Young, J. B. & Chamberlain, M. A. (1987) The contribution of the Stanford Health Assessmetn questionnaire in rheumatology clinics. *Clinical Rehabilitation* **1**, 97-100.

Zarin, D. A., Pincus, H. A., West, J. C. & McIntyre, J. S. S. O. (1997) Practice-based research in psychiatry. *American Journal of Psychiatry* **154**, 1199-208.

Zarin, D. A., West, J. C., Pincus, H. A. & McIntyre, J. S. (1996) The American Psychiatric Association Practice Research Network (PRN). In *Outcomes Assessment in Clinical Practice* (eds L. I. Sederer & B. Dickey). Baltimore: Williams and Wilkins.

Zigmond, A. S. & Snaith, R. P. (1983) The Hospital Anxiety and Depression scale. *Acta Psychiatrica Scandinavica* **67**, 361-70.

Zung, W. W., Magill, M., Moore, J. T. & George, D. T. (1983) Recognition and treatment of depression in a family medicine practice. *Journal of Clinical Psychiatry* **44,** (1), 3-6.

Zung, W. W. K. (1965) A self rating depression rating scale. *Archives of General Psychiatry* **12**, 63-70.

Appendices to the thesis

Appendix 1: Measuring outcome in mental health research: do the methods matter? (protocol)

Background

In order to determine whether a treatment or intervention is effective, it is necessary to systematically measure the outcome of that intervention – ideally in the context of a robust clinical experiment, such as a randomised trial. Outcome in medicine has traditionally been measured in terms of mortality and biophysical parameters, such as radiological, clinical and laboratory assessments. However, there is a growing body of opinion that such traditional measures are inadequate, since they do not reflect the impact of diseases or illness on the individual, nor do they reflect the experience of illness form the patients' perspective (Jenkinson, 1995; Lohr, 1988). Patient based instruments have emerged to augment this perceived inadequacy of traditional outcome measurement (Fitzpatrick, *et al.*, 1984). These *patient based* instruments have been variously called Quality of Life measures; Health Status measures; Health Related Quality of Life measures; and functional status measures (Bowling, 1997). A unifying feature is that they tend to examine some combination of the following domains: *physical functioning; social functioning; role functioning; mental well being; and general health perceptions* (Ware, 1995). Such measures also give prominence to the patients' own evaluation of these domains, in that they are often (but not always) completed by the patient, rather than the clinician.

There is a strong tradition of measurement in psychiatry, where standardised measures have evolved, often as the basis of diagnostic classification systems which record or count clinician-elicited symptoms in order to form a psychiatric diagnosis in a reproducible manner (Hamilton, 1976). This tradition has facilitated important research into, for example, the population incidences and natural history of common psychiatric disorders (e.g. (Reiger & Kaelber, 1995; Sartorius, *et al.*, 1986)) and their aetiological risk factors (e.g. (Shepherd, *et al.*, 1989)). However, the measurement of psychiatric symptoms has also become the basis of how outcome is measured in evaluative research in psychiatry. For example, trials investigating drug treatments for depression almost uniformly use symptom based measures of depressive disease severity (such as the Hamilton Depression Rating Scale (Hamilton, 1967)) as their primary measure of outcome (Song, *et al.*, 1993), whilst patient based outcome is

largely ignored (Hotopf, *et al.*, 1997). Similarly, evaluative research in schizophrenia gives primacy to the measurement of symptoms such as delusions and hallucinations (Thornley & Adams, 1998) using instruments such as the Positive and Negative Syndrome Scale (Kay, 1991), but generally fails to measure patient based outcomes, such as health related quality of life (Awad, 1992; Collins, *et al.*, 1991; Meltzer, 1999). The research knowledge base in psychiatry therefore generally uses symptom reduction as its primary outcome and judges the value of its treatments according to this criterion.

Despite this general trend a number of patient based measures have been developed or applied in psychiatry to supplement symptom based measures. For example important *disease specific* patient based measures are available for use in psychiatric populations (e.g. (Lehman, 1983b)). Similarly, *generic* patient based measures such as the Short Form 36 (Ware, *et al.*, 1993) have been developed to be applied across different patient populations and different diseases in order to provide a common metric of health related quality of life. Measures such as the SF36 have been shown to be applicable to psychiatric populations (Russo, *et al.*, 1998), and have been adopted as outcome measures in psychiatric research (e.g. (Simon, *et al.*, 1996; Tunis, *et al.*, 1999)).

The theoretical foundation of the recent enthusiasm for patient based measures is the assertion that they examine a potentially different facet of outcome from conventional measures. Consequently, the results obtained from a patient based measure may not be the same as the results from a conventional measure. Thus, for example, a person may report improvement on a patient based outcome measure, whilst showing no improvement on a conventional measure of psychiatric symptomatology (or vice versa).

We therefore decided to examine the empirical basis of this assertion in psychiatry by examining whether a different result or answer is obtained in evaluative research when a patient based measure is used alongside a conventional symptom based measure of outcome.

Aims and objectives

The aim of the present study is to examine the inter-relationship between patient based and symptom based outcome in high quality evaluative psychiatric research.

The specific objective is to examine whether the results which are obtained from patient based measures differ in direction or magnitude from the results that are obtained from symptom based measures.

Methods

In order to examine the relationship between patient based and symptom based outcome, we sought all examples of high quality research which used a patient based measure alongside a symptom based measure. The following inclusion criteria and operational definitions were adopted:

Study inclusion criteria

Randomised trials form the least biased method by which differential outcome can be assessed between competing interventions (WHO, 1991). We therefore sought English language reports of randomised trials of psychiatric interventions that fulfilled the following criteria:

1. Truly randomised (not quasi-experimental or non-randomised control design)
2. Measuring the efficacy or effectiveness of an intervention.
3. Includes a psychiatric population (i.e. where the primary reason for inclusion in the study is the presence of a common functional psychiatric disorder such as depression, anxiety, schizophrenia or manic depressive illness).
4. Uses both a patient based measure and a symptom based measure to record outcome (see below for definitions).

Patient based outcome measures

A study was defined as having employed a patient based outcome measure if the measure was explicitly referred to as a measure of Health Status; Quality of Life; Health Related Quality of life; functional status.

Additionally, any measures which could not be clearly classified as patient based outcomes according to the above were included if they contained two or more of the following domains: *physical functioning; social functioning; role functioning; mental well-being; cognitive capacity; general health perceptions & subjective well-being*. A series of operational definitions for these domains were adapted from (Ware, 1987; Ware, 1995). The content of individual measures was first ascertained from the text of the study, and if it's eligibility was still unclear, then its content was established from the cited reference supporting the instrument or a standard text summarising the content of many patient based measures of outcome (Bowling, 1997).

Instruments that examined only one of these domains were excluded. For example, instruments that measure only social functioning were not considered to be patient based measures, since they are limited in their scope. Similarly instruments which measure symptoms of common psychiatric disorders such as depression or anxiety have been classified as patient based measures by some authors (e.g. (Sanders, *et al.*, 1998)), but are not classified as such here, since these are both limited in scope and also dominated psychiatric symptoms.

Studies which employed unpublished scales (i.e. those which did not have a supporting reference to outline its basis and psychometric properties) were not included, since they have recently been shown to be an important source of bias, and therefore potentially confounding influence (Marshall, *et al.*, 2000).

Symptom based measures

A study was defined as having employed a symptom based measure if it contained a standardised measure that was broadly made up of items which measured symptoms and signs associated with the underlying psychiatric

disorder under investigation. Core symptoms were taken from standard textbooks of psychiatry and the DSM (American Psychiatric Association, 1994), where this was unclear. Again, only studies that employed published scales were included to eliminate this source of bias (Marshall, *et al.*, 2000).

The literature search.

Electronic searches were undertaken of the following databases in order to identify potentially relevant trials: Medline; Embase; Cinahl; PsycLit; Cochrane Controlled Trials Register.

The search strategy used combined two sets of search terms relating to the target patient population and the use of patient based measures:

Patient population: a search strategy is used which captures publications relating to all forms of mental illness using MeSH terms.

Patient based measures: an already developed search strategy is used which has been shown to have acceptable sensitivity and precision in identifying research which relates to or includes patient based outcome measures (Brettle, *et al.*, 1998).

The Cochrane controlled trials register contains records of randomised trials identified by handsearches of a number of psychiatric journals. In addition we hand searched the following journals, which regularly publish evaluative research that employs patient based outcome measures: Medical Care (1972-2000); Quality of Life Research (1993-2000).

A further source of trials was the Cochrane database of systematic reviews. All reviews published by the Cochrane Schizophrenia Group and the Cochrane Depression, Anxiety and Neurosis Group were examined to check for included trials which reported the use of patient based outcomes. Lastly, we searched an ongoing database of trials maintained by the Cochrane Schizophrenia Group which includes a brief summary of the scope of the specific outcomes measures which are employed in each trial (Adams, 1998).

Data extraction

The following data were extracted from each eligible trial, and entered into standardised electronic data extraction form:

- Background information
- Patient population and underlying psychiatric problem
- Type of intervention
- Setting
- Name of patient based measure
- Content of patient based outcome measure (domains which are included)
- Mode of completion of patient based outcome measure (interviewer completed, patient completed)
- Name of symptom based measure
- Mode of completion of symptom based measure (interviewer completed, patient completed)

Study results

For both symptom based and patient based outcomes the following were extracted where possible.

Significant/non significant result

A result was said to be significant if it reported a between group difference at less than the p=0.05 significance level, or reported a 95% confidence interval which did not include a null result at some point in the trial. Results based on overall (summary) scores from a scale were considered in the first instance. Claims of significance based upon analyses of individual items or sub-components of a scale were ignored. However, several measures were found to report a 'profile' of their various component domains of outcome. For example, the SF36 is generally presented as a profile of its eight components of health related quality of life, as is the Duke health profile (Parkerson, *et al.*, 1990). Where only profiles were presented, then a significant result was

operationally defined as being present when half or more of the individual profile items were significant at less than p=0.05.

Favours treatment or control.

For those studies that reported a significant result, the direction of that result in terms of favouring either treatment or control was established. Additionally, for studies where it was not clear what represented the treatment or the control, this was recorded.

Magnitude of effect sizes.

Where possible, data were extracted pertaining to the magnitude of change and between group differences for the scale under consideration. The following data were sought:

- Control group change scores and standard deviations
- Treatment group change scores and standard deviations
- Control group endpoint scores and standard deviations
- Treatment group endpoint scores and standard deviations.
- Between group differences in change scores (and standard deviations)

Where standard errors or confidence intervals were reported, these were converted to standard deviations, according to accepted methods ((Altman, 1991; Cooper & Hedges, 1994)).

Choice of outcome measures.

Where studies employed more than one potentially relevant symptom based or patient based outcome measure, then the following hierarchical decision rules were employed to determine which outcome should be included in the data analysis.

1. If the authors clearly indicated which was the primary outcome of concern, then this was used.
2. If only one outcome presented sufficient data to be extracted, then this was used.

3. If several outcomes were presented in sufficient detail to allow data to be extracted, then the first outcome to be reported (either in the text or in tabular form) was used.
4. In cases where there was still confusion, then an outcome was selected at random; using computer generated random number tables.

Reliability of judgements

The reliability with which studies were included, outcomes chosen, and statistical significance established were examined in a subset of 10% of the potentially relevant studies by a second rater. Cohen's weighted kappa was calculated for each of these judgements (Cohen, 1968).

Data analysis

The effect of choice of outcome measure and the outcome observed in RCTs was examined in two ways. First by examining whether the choice of outcome measure influenced the chance of a treatment being shown to be statistically superior to a control condition. Second, by examining whether the magnitude of effect size was different according to whether a patient based outcome or a symptom based outcome measure was selected.

Is the choice of outcome related to the chance of a significant result?

The chance of a significant result being obtained was calculated by establishing a relative risk and attendant 95% confidence intervals from the following contingency table (Gardiner & Altman, 1989).

	Favours treatment (p<0.05)	nonsignificant result
Patient based outcome measure	a	b
Symptom based outcome measure	c	d

What is the relationship between patient based outcome measures and symptom based measures in terms of magnitude and direction of effect?

We obtained a standardised measure of effect size for both symptom based and patient based outcome measures from each trial, by calculating Cohen's d (Cohen, 1988) for the *ith* study, according the following formula:

$$d_i=(M_{exp}-M_{cont})/SD_{pooled}$$

Chosen means (M_{exp} -M_{cont}) were those which were most frequently reported in the included trials - i.e. either endpoint scores or mean change scores.

A convention was adopted such that positive standardised effect sizes indicated benefit for the treatment group and the larger the magnitude of the standardised effect, the greater the magnitude of benefit. This generally required sign changes for symptom based measures, since higher scores indicate clinical deterioration and worse outcome. Conversely, high scores for patient based measures generally indicated patient improvement, and required no sign change.

The relationship between standardised patient and symptom-based scores was examined graphically by constructing a scatterplot, and statistically using regression analysis. Since each point on this scatterplot represents the standardised estimate from a single study of variable sample size and dispersion (and therefore with variable precision of estimate) then a weighted

regression analysis was used. Studies were weighted according to the inverse of their variance (Hedges, 1982):

$$weight_i = 1/variance_i$$

The variance for each individual study was estimated using the approximation of (Rosental & Rubin, 1982), where the variance for the *i*th study was found according to the following formula:

$$variance_i = 2N_i/(8+d_i^2)$$

Where N_i is the total sample size and d_i is estimated as previously. Thus, the largest studies and those with least dispersion contribute the greatest weight to this regression model. This approach is robust when sample sizes are approximately equal and greater than 10 in control and experimental groups (Hedges, 1982). Data that deviated from this assumption were excluded from the analysis. All calculations were conducted using STATA software (STATA corporation, 1999).

Sensitivity analyses

A number of à priori specified sensitivity analyses were conducted where there were sufficient data to allow this. The relationship between patient and symptom based outcome were examined separately for:

- Different mental disorders (e.g. depression and anxiety versus Schizophrenia and related disorders)
- Outcomes measures completed by different methods (e.g. patient completed versus physician completed measures).

Appendix 2: Electronic search strategies

Search strategies for the thesis

The search strategies employed in this thesis were designed and conducted in collaboration with an experienced information expert, Ms Kate Misso, at the NHS Centre for Reviews and Dissemination between 1998 and 2000

The primary purpose of the literature searches was to collect primary research and review articles pertaining to patient based clinical outcomes in the sphere of psychiatry and mental health.

Several sections of the present thesis draw upon searches of published literature. The literature searches were therefore designed to meet each of the overlapping requirements of several different reviews in different sections of the thesis

In particular, the following were sought:

1. Background papers, such as review articles, journal editorials, or opinion pieces, which discussed the use of patient based outcomes in the sphere of mental health. These form the basis of the literature used in Section 1.

2. Outcomes research conducted in the sphere of mental health. These form the basis of the studies used in section 2.

3. Controlled trials that examine the impact of the use of routine outcomes measures in the care of those with mental illness. These form the basis of studies used in section 4.

Components of the search strategies – general comments

A search strategy usually contains two elements (NHS Centre for Reviews and Dissemination, 2000):

1. Search terms specific to the clinical area and type of intervention under consideration – such as disease type, intervention or area of speciality.
2. Search terms specific to the study type under consideration – such as a controlled trial in the case of reviews of effectiveness.

The search strategies employed in this review concentrated on the former rather than the latter, since a wide variety of research designs and publication types were potentially needed in the series of reviews contained in this thesis. An initial decision was taken to be as inclusive as possible or practical in the conduct of search strategies. Traditional search strategies which are employed in systematic reviews using well honed search terms designed to identify randomised trials with maximum specificity and sensitivity were therefore not employed.

Simple search strategies which employed the term 'outcome(s)' as a free text term produced unmanageable numbers of references, and were likely to miss studies which used one of the many synonyms for patient based outcome, such as quality of life and health status. This is likely to be due to the ubiquity of the term outcome in primary and secondary research, and the fact that this is a common subheading in the structured abstract of many journals (Anonymous, 1987). A level of refinement was therefore required.

An initial scoping review of the literature revealed that some empirical research into the utility of various search strategies and databases relating either to outcomes or mental health had been conducted.

Searching for literature on outcomes

Brettle, *et al.*, (1998) reported a series of iterations of search strategies developed in assembling databases for the now extinct UK Outcomes Clearing House, formerly based at the Nuffield Institute at the University of Leeds. Brettle, *et al.*, (1998) produced a series of 'filters' to be used by clinicians and librarians in order to retrieve information from MEDLINE pertaining to outcomes measurement. Three types of filter were produced which varied in their degree of efficiency in identifying literature on outcomes. These were termed *basic, intermediate and comprehensive* search strategies. The efficiency of search strategies is traditionally described in terms of *precision* and *recall* (McKibbon & Walker-Dilks, 1995), which are terms analogous to *sensitivity* and *specificity* in medical screening (Sackett, *et al.*, 1991). Precision refers to the proportion of citations in a given search that are relevant to the search question, whereas recall refers to the proportion of relevant citations in a given search among the total possible relevant citations. Basic searches produce fewer references, but the proportion of those that are relevant to the underlying question is high – these are therefore precise, but have lower recall. More comprehensive searches obtain a large number of references, and include a greater number of total relevant references, but contain a large number of redundant citations – these are therefore less precise, but have greater recall.

When the three search strategies were used on MEDLINE, in combination with mental health specific search strategies, an unmanageable number of references were obtained with the comprehensive strategy (see table 1). The strategy adopted in all searches was therefore based upon the 'intermediate' strategy of Brettle, *et al.*, (1998).

The three filters contained combinations of text words and medical subject (MeSH) terms relevant to outcomes measurement. Common to the three search strategies are the three MeSH healthings: health status indicators/; outcome and process assessment (health care)/; and outcome assessment (healthcare)/. These searches, together with their precision and recall are given in are given in full in table 1.

Results obtained from *basic, intermediate and comprehensive* 'outcomes' searches of MEDLINE

Search	**Number of hits**
Basic	69,367
Intermediate	21,489
Comprehensive	4,492

Unfortunately Brettle, *et al.*, (1998) do not give any guidance regarding the use of similar filters in searching other databases (such as EMBASE and PsycLIT), nor do they make recommendations about the value of other databases in identifying literature not held in MEDLINE. There are sound theoretical reasons why additional databases might need to be searched, particularly in relation to mental health (Adams, *et al.*, 1992; Adams, *et al.*, 1994). The filters described by Brettle, *et al.*, (1998) were therefore adapted for use in other databases (see on).

Search strategies for identifying literature on outcomes measurement, after (Brettle, *et al.*, 1998)

Basic	Intermediate	Comprehensive
1 health status indicators/ 2 outcome and process assessment (health care)/ 3 outcome assessment (health care)/ 4 quality of life/ 5 outcome measure$.tw. 6 health outcome$.tw 7 1 or 2 or 3 or 4 or5 or 6] *8 your search term(s) (subject specific)* 9 7 and8 **Mean recall = 57%** **Mean precision = 69%**	1 health status indicators/ 2 outcome and process assessment (health care)/ 3 outcome assessment (health care 4 quality of life/ 5 outcome measure$.tw. 6 health outcome$.tw. 7 quality of life.tw.] 8 measure$.tw. 9 assess$.tw. 10 (score$ or scoring).tw. 11 health status.tw. 12 (endpoint$ or end point$ or end-point$).tw. 13 scale$.tw. 14 functional outcome$.tw. 15 outcome$.ti. 16 or/1–15 17 (outcome$ adj3 (measure$ or assess$ or (score$ or scoring) or index or indices or scale$ or monitor$)).tw. 18 or/1–7 19 17 or 18 *20 your search term(s) (subject specific* 21 19 and 20 **Mean recall = 73%** **Mean precision = 65%**	1 health status indicators/ 2 outcome and process assessment (health care)/ 3 outcome assessment (health care)/ 4 quality of life/ 5 health status/ 6 severity of illness index/ 7 self assessment [psychology] 8 outcome measure$.tw. 9 health outcome$.tw. 10 quality of life.tw. 11 health status.tw. 12 (endpoint$ or end point$ or end-point$).tw. 13 (self-report$ or self report$).tw. 14 functional outcome$.tw. 15 outcome$.ti. 16 or/1–15 17 outcome$.tw. 18 measure$.tw. 19 assess$.tw. 20 (score$ or scoring).tw. 21 index.tw. 22 indices.tw. 23 scale$.tw. 24 monitor$.tw. 25 or /18–24 26 17 and 25 27 16 or 26 *28 your search term(s) (subject specific)* 29 27 and 28 **Mean recall = 100%** **Mean precision = 48%%**

Searching for literature on mental health

A substantial amount of work has been conducted into the databases that are required to identify research literature (primarily controlled trials) for inclusion in systematic reviews in the sphere of mental health in general, and schizophrenia in particular. Hay, *et al.*, (1996) and Adams, *et al.*, (1994) recommend the use of MEDLINE, EMBASE and PsycLIT, in searching for trials relating to eating disorders and schizophrenia. These databases were searched, in addition to three other databases that are likely to contain literature of relevance to a review of outcomes measures. Specific details of each of these databases are given below:

- Cinahl
- British Nursing Index
- Cochrane Controlled Trials Register

Search terms have also been developed which are maximally sensitive in retrieving literature related to schizophrenia and depression as distinct conditions. The reviews presented in the thesis do not restrict themselves to specific conditions, and these specialist filters were not deemed appropriate. Instead, a series of search strategies were developed in order to obtain literature relevant to psychiatry and mental health. These search strategies generally comprised medical subject headings for literature related to psychiatry and mental health.

For example, three major Medical subject headings in MEDLINE were relevant:

- "Mental-Health"/ all subheadings
- "Psychiatry"/ all subheadings
- "Mental-Disorders"/ all subheadings

In addition, free text words were added to capture literature not classified under these subheadings:

- (mental health) in ti,ab
- (mental* illness*) in ti,ab
- (mental* ill) in ti,ab
- psychiatry in ti,ab
- (mental* disorder*) in ti,ab
- psychiatric in ti,ab
- (mental* ill-health) in ti,ab

The specific 'mental health' and 'outcomes' search terms varied and were adapted for each specific database, and are given below:

MEDLINE

This database corresponds to three print indexes: Index Medicus; Index to Dental Literature and the International Nursing Index. Additional materials not published in Index Medicus are included in MEDLINE in areas of communication disorders, population and reproductive biology. MEDLINE is the National Library of Medicine's premier bibliographic database covering the fields of medicine, nursing, dentistry and the pre-clinical sciences. Each record is indexed using NLM's controlled vocabulary, MeSH (Medical Subject Heading. Coverage is from 1966 to date. It is produced by the National Library of Medicine, Bethesda MD, USA.

MENTAL HEALTH TERMS

1. explode "Mental-Health"/ all subheadings
2. explode "Psychiatry"/ all subheadings
3. explode "Mental-Disorders"/ all subheadings
4. (mental health) in ti,ab
5. (mental* illness*) in ti,ab
6. (mental* ill) in ti,ab
7. psychiatry in ti,ab
8. (mental* disorder*) in ti,ab
9. psychiatric in ti,ab
10. (mental* ill-health) in ti,ab

OUTCOMES TERMS

1. "Health-Status-Indicators"
2. "Outcome-and-Process-Assessment-(Health-Care)"/ all subheadings
3. "Outcome-Assessment-(Health-Care)"/ all subheadings
4. "Quality-of-Life"/ all subheadings
5. (outcome measure*) in ti,ab
6. (health outcome*) in ti,ab
7. (quality of life) in ti,ab
8. measure* in ti,ab
9. assess* in ti,ab
10. (score* or scoring) in ti,ab
11. index in ti,ab
12. indices in ti,ab
13. scale* in ti,ab
14. monitor* in ti,ab
15. #8 or #9 or #10 or #12 or #11 or #13 or #14
16. outcome* in ti,ab
17. #16 near3 #15
18. #1 or #2 or #3 or #4 or #5 or #6 or #7
19. #17 or #18

EMBASE

EMBASE is a major bibliographic database, which covers world-wide medical journals, with particular emphasis in the areas of drugs and toxicology. Inclusion of European material is particularly strong. Produced by Elsevier Science B. V., Amsterdam, Netherlands.

MENTAL HEALTH TERMS

1. explode "mental-health"/ all subheadings
2. explode "psychiatry"/ all subheadings
3. explode "mental-disease"/ all subheadings
4. mental health in ti,ab
5. mental* ill in ti,ab
6. mental* illness* in ti,ab
7. mental* ill-health in ti,ab
8. psychiatry in ti,ab
9. psychiatric in ti,ab
10. mental* disorder* in ti,ab

OUTCOMES TERMS

1. "health-survey"/ all subheadings
2. explode "quality-of-life"/ all subheadings
3. "outcomes-research"/ all subheadings
4. health outcome* in ti,ab
5. quality of life in ti,ab
6. outcome measure* in ti,ab
7. measure* in ti,ab
8. (score* or scoring) in ti,ab
9. index in ti,ab
10. indices in ti,ab
11. scale* in ti,ab
12. monitor* in ti,ab
13. assess* in ti,ab
14. #7 or #8 or #9 or #10 or #11 or #12 or #13
15. outcome* in ti,ab
16. #15 near3 #14
17. #1 or #2 or #3 or #4 or #5 or #6
18. #16 or #17

PsycLIT

This database provides access to the international literature in psychology and related behavioural and social sciences, including psychiatry, sociology, anthropology, education, pharmacology, and linguistics. PsycLIT contains all records from the printed Psychological Abstracts, plus materials from Dissertation Abstracts International and other sources fo publication types indexed to include journal articles, dissertations, reports, books, and book chapters. Coverage 1887 to date. Produced by the American Psychological Association, Washington, DC, USA.

MENTAL HEALTH TERMS

1. explode "Mental-Health"
2. explode "Psychiatry"
3. explode "Mental-Disorders"
4. mental health in ti,ab
5. mental* ill* in ti,ab
6. mental* ill-health in ti,ab
7. psychiatry in ti,ab
8. psychiatric in ti,ab
9. mental* disorder* in ti,ab

OUTCOMES TERMS

1. explode "Treatment-Outcomes"
2. explode "Psychological-Assessment"
3. explode "Quality-of-Life"
4. (outcome* or process*) near3 assessment*
5. health status indicator*
6. health status
7. health outcome* in ti,ab
8. quality of life in ti,ab
9. outcome measure* in ti,ab
10. measure* in ti,ab
11. assess* in ti,ab
12. (score* or scoring) in ti,ab
13. index in ti,ab
14. indices in ti,ab
15. scale* in ti,ab
16. monitor* in ti,ab
17. #10 or #11 or #12 or #13 or #14 or #15 or #16
18. outcome* in ti,ab
19. #18 near3 #17
20. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9
21. #19 or #20

CINAHL

Cinalhl is a commercially produced database which includes bibliographic details pertaining to nusrsing and allied care

MENTAL HEALTH TERMS

1. explode "Mental-Health"/ all topical subheadings / all age subheadings
2. explode "Psychiatry"/ all topical subheadings / all age subheadings
3. explode "Mental-Disorders"/ all topical subheadings / all age subheadings
4. mental health in ti,ab
5. mental* ill* in ti,ab
6. mental* ill-health in ti,ab
7. psychiatry in ti,ab
8. psychiatric in ti,ab
9. mental* disorder* in ti,ab

OUTCOMES TERMS

1. explode "Health-Status"/ all topical subheadings / all age subheadings
2. explode "Health-Status-Indicators"/ all topical subheadings / all age subheadings
3. explode "Outcome-Assessment"/ all topical subheadings / all age subheadings
4. "Outcomes-(Health-Care)"/ all topical subheadings / all age subheadings
5. explode "Quality-of-Life"/ all topical subheadings / all age subheadings
6. health outcome* in ti,ab
7. quality of life in ti,ab
8. outcome measure* in ti,ab
9. measure* in ti,ab
10. assess* in ti,ab
11. (score* or scoring) in ti,ab
12. index in ti,ab
13. indices in ti,ab
14. scale* in ti,ab
15. monitor* in ti,ab
16. #9 or #10 or #11 or #12 or #13 or #14 or #15
17. outcome* in ti,ab
18. #17 near3 #16
19. #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8
20. #18 or #19

BNI/RCN

MENTAL HEALTH TERMS

1. mental health
2. mental* ill*
3. mental* ill-health
4. psychiatry
5. psychiatric
6. mental* disorder*

OUTCOMES TERMS

1. health status
2. status indicator*
3. (outcome* or process*) near3 assessment*
4. health outcome*
5. quality of life
6. outcome* measure*
7. assess*
8. score* or scoring
9. index
10. indices
11. scale*
12. monitor*
13. #7 or #8 or #9 or #10 or #11 or #12
14. outcome*
15. #14 near3 #13
16. #1 or #2 or #3 or #4 or #5 or #6
17. #16 or #15

CCTR

The Cochrane controlled trials register contains bibliographic details of controlled trials identified fomr literature and hand searches of a number of electronic databases and journals

MENTAL HEALTH TERMS

1. MENTAL-HEALTH*:ME
2. PSYCHIATRY*:ME
3. MENTAL-DISORDERS*:ME
4. MENTAL:TI NEAR HEALTH:TI
5. MENTAL:AB NEAR HEALTH:AB
6. MENTAL*:TI NEAR ILLNESS:TI
7. MENTAL*:AB NEAR ILLNESS:AB
8. MENTAL*:TI NEAR ILL:TI
9. MENTAL*:AB NEAR ILL:AB
10. PSYCHIATRY:TI OR PSYCHIATRY:AB
11. MENTAL*:TI NEAR DISORDER*:TI
12. MENTAL*:AB NEAR DISORDER*:AB
13. PSYCHIATRIC:TI OR PSYCHIATRIC:AB
14. MENTAL*:TI NEAR ILL-HEALTH:TI
15. MENTAL*:AB NEAR ILL-HEALTH:AB

OUTCOMES TERMS

1. HEALTH-STATUS-INDICATORS:ME
2. OUTCOME-AND-PROCESS-ASSESSMENT-HEALTH-CARE:ME
3. OUTCOME-ASSESSMENT-HEALTH-CARE:ME
4. QUALITY-OF-LIFE:ME
5. OUTCOME:TI AND MEASURE*:TI
6. OUTCOME:AB AND MEASURE*:AB
7. HEALTH:TI AND OUTCOME*:TI
8. HEALTH:AB AND OUTCOME*:AB
9. QUALITY:TI NEAR LIFE:TI
10. QUALITY:AB NEAR LIFE:AB
11. MEASURE:TI OR MEASURE:AB
12. ASSESS*:TI OR ASSESS*:AB
13. SCORE*:TI OR SCORING:TI OR SCORE*:AB OR SCORING:AB
14. INDEX:TI OR INDEX:AB
15. INDICES:TI OR INDICES:AB
16. SCALE*:TI OR SCALE*AB
17. MONITOR*:TI OR MONITOR*:AB
18. #11 OR #13 OR #14 OR #15 OR #16 OR #17
19. OUTCOME*:TI OR OUTCOME*:AB
20. #19 AND #18
21. #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10
22. #21 OR #20

Appendix 3: Survey questionnaire

♦ National Outcomes Survey ♦

2 0 0 0

Please take a few minutes to complete this survey – the answers that you give will be treated in strictest confidence and will not be attributed

- Thank you for your help

A. Please indicate your area(s) of

	Main	*Subspeciality*
General adult psychiatry	☐	☐
Community psychiatry	☐	☐
Rehabilitation psychiatry	☐	☐
Old age psychiatry	☐	☐
Child and adolescent	☐	☐
Liaison psychiatry	☐	☐
Psychotherapy	☐	☐
Drugs and alcohol	☐	☐
Academic psychiatry	☐	☐
Other (please specify)	☐	

B. What type of hospital/trust do you work in?

	tick which applies
Teaching hospital/Trust	☐
District general hospital/non-teaching trust	☐
Other (please specify)	☐

C. How long have you worked as a consultant psychiatrist?

☐ years

D. In any one week, approximately how many patients are you responsible for?

Inpatients	☐ patients
Day hospital patients	☐ patients
Outpatients (average seen per week)	☐ patients

The following questions are about your use and experience of standardised outcomes measures.

By standardised outcomes measures, we mean questionnaires designed to do any of the following:

- elicit or measure patients' symptoms,
- identify specific problems/needs
- measure quality of life/social functioning.

1. Do you use standardised outcomes measures to check for any of the specific psychiatric problems outlined below?

- tick which apply and specify any measures which you use

	never	occasionally	routinely	which measure do you use?
Depression/anxiety	☐	☐	☐	
Schizophrenia/ Psychosis	☐	☐	☐	
Cognitive impairment	☐	☐	☐	
Drugs/alcohol problems	☐	☐	☐	
Other psychiatric problem (please				

2. Do you use standardised outcomes measures to check for deficits in social functioning, quality of life, or to assess patient needs?

- indicate for which diagnostic groups you do this and specify any measures

	never	occasionally	routinely	which measure do you use?
Depression/anxiety	☐	☐	☐	
Schizophrenia/ Psychosis	☐	☐	☐	
Cognitive impairment	☐	☐	☐	
Drugs/alcohol problems	☐	☐	☐	
Other (please specify)	☐	☐	☐	

3. Do you use outcomes measures to measure *change over time* or therapeutic response?

- please specify for which diagnostic groups of patients and which

	never	occasionally	routinely	which measure do you use?
Depression/anxiety	☐	☐	☐	
Schizophrenia/ Psychosis	☐	☐	☐	
Cognitive impairment	☐	☐	☐	
Drugs/alcohol problems	☐	☐	☐	
Other (please specify)	☐	☐	☐	

4. Do you use standardised outcomes measures as tools for clinical audit?

	never	occasionally	routinely	which measure do you use?
Depression/anxiety	☐	☐	☐	
Schizophrenia/ Psychosis	☐	☐	☐	
Cognitive impairment	☐	☐	☐	
Drugs/alcohol problems	☐	☐	☐	
Other (please specify)	☐	☐	☐	

5. Do you use any of the following outcomes measures for audit purposes?

	Yes	No	*- Indicate which, if any*
Mortality	☐	☐	
Suicide	☐	☐	
Length of stay	☐	☐	
Re-admission rates	☐	☐	
Use of the mental health act	☐	☐	
Other - please specify			

6. Are you asked to routinely collect standardised outcomes measures by your hospital/trust?

- indicate for which patients/diagnostic groups you do this

	No	Yes	Which measures are you asked to use?
All patients	☐	☐	
Depression/anxiety	☐	☐	
Schizophrenia/	☐	☐	
Drugs/alcohol	☐	☐	
Cognitive	☐	☐	
Other (please			

7. Are you required/asked by your hospital/trust to collect the following information on your patients?

	Yes	No
Health of the Nation Outcomes (HoNOS) scores	☐	☐
Needs assessments **(e.g. using Camberwell Assessment of Needs; MRC Needs for**	☐	☐

Other (please specify)

8. Do you use standardised outcomes measures as part of the care programme approach?

Yes ☐ No ☐

if yes, please indicate which measures

9. Does your hospital/trust routinely collect any of the following data/ outcomes?

	Yes	No
Length of stay	☐	☐
Readmission rates	☐	☐
Use of the Mental Health Act	☐	☐
Suicides	☐	☐
Deaths	☐	☐
Adverse/untoward events (please	☐	☐
Other (please specify)	☐	☐

10. Are any of these data ever fed back to you?

Yes ☐ No ☐

11. Have outcomes measures ever been used in planning your services or allocating specific resources to your service?

Yes ☐ No ☐

Please use this box to give examples

12. Please indicate which (if any) of the following measures you use in your clinical practice

- you may have indicated that you use one or more of these already, but please complete this section as well

	Never	Sometimes	Routinely
General Health Questionnaire (GHQ)	☐	☐	☐
Hospital Anxiety and Depression (HAD)	☐	☐	☐
Zung Depression Inventory	☐	☐	☐
Brief Psychiatric Rating Scale (BPRS)	☐	☐	☐
Short Form 36	☐	☐	☐
Anxiety and Depression Outcome Scale	☐	☐	☐
Lancashire Quality of Life Scale	☐	☐	☐
MRC Needs for Care Assessment	☐	☐	☐
HoNOS	☐	☐	☐
HoNOS 65+	☐	☐	☐
Symptom Check List (SCL) – 60	☐	☐	☐
Positive and Negative Syndrome Scale	☐	☐	☐
Abbreviated Mental Test Score (AMTS)	☐	☐	☐
Camberwell Assessment of Need (CAN)	☐	☐	☐
Hamilton Depression Rating Scale	☐	☐	☐
Beck Depression Inventory (BDI)	☐	☐	☐
Mini Mental State Examination (MMSE)	☐	☐	☐
Quality of Life Index	☐	☐	☐

Other (please indicate) ☐

13. Please use this space to tell us more about your experience in using outcomes in the routine care of your patients and in planning your clinical services - continue over the page if necessary.

Thank you for completing this questionnaire - please use the stamped addressed

this questionnaire should be returned to:

Dr. Simon Gilbody
NHS Centre for Reviews and Dissemination
University of York
YO10 5DD

Appendix 4: Covering letters for questionnaire survey

Appendix 5– Covering letter #1

♦ National Outcomes Survey ♦

2 0 0 0

Mobile Phone 07940 576699
e-mail smg5@york.ac.uk

Dr «Initials» «Surname»
«Address_line_1»
«Address_line_2»
«Adress_line_3»
«Town» «County» «Postcode»

Dear Dr «Surname»

I am writing to ask for your help in an important survey funded by the Medical Research Council.

You will no doubt be aware that psychiatrists in the United Kingdom are encouraged to record the outcomes of care for their patients. The purpose of our survey is to ask clinicians who work in the real world what (if anything) they use in the way of standardised measures. We also want to know your thoughts on their value.

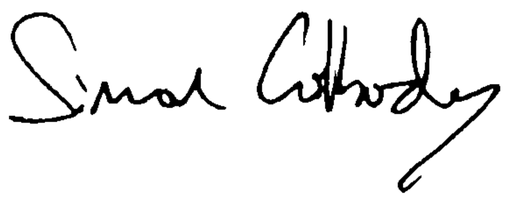
The attached brief questionnaire should not take more than a couple of minutes of your time. Your answers will help provide the first complete picture of current practice in the NHS, and will hopefully be useful in formulating future policy. We have provided a stamped addressed envelope in which to return the form.

In recognition of your contribution, respondents will be entered into a prize draw to win a copy of the new Oxford Textbook of Psychiatry. Your returned questionnaire will automatically be entered into this draw.

Please do not hesitate to contact me on the above number (or by fax or e-mail) if you require any further information or help completing this survey.

With many thanks in anticipation.

Yours sincerely



Dr Simon Gilbody *MRCPsych*
MRC Clinical Fellow in Health Services Research

Covering letter #2

♦ National Outcomes Survey ♦
2 0 0 0

Mobile Phone 07940 576699
e-mail smg5@york.ac.uk

Dr «Initials» «Surname»
«Address_line_1»
«Address_line_2»
«Adress_line_3»
«Town» «County» «Postcode»
02 January 2002

Dear Dr «Surname»

I am writing again to ask for your help with the Medical Research Council sponsored survey of the use of outcomes measures by UK Psychiatrists. A questionnaire was sent, which you might not have had time to complete.

We hope to gain the first really comprehensive picture of psychiatrists' experiences in using these measures. Your contribution to this project will help ensure that our results are valid and representative of those who provide mental health care in the real world.

Respondents will be entered into a prize draw to win a copy of the new Oxford Textbook of Psychiatry. Your returned questionnaire will automatically be entered into this draw.

I apologise if you have already returned the questionnaire, and these letters have 'passed' in the post.

With many thanks in anticipation.

Yours sincerely

Simon Gilbody

Dr Simon Gilbody *MRCPsych*
MRC Clinical Fellow in Health Services Research

Final reminder

♦ National Outcomes Survey ♦
2 0 0 0

Mobile Phone 07940 576699
e-mail smg5@york.ac.uk

Dr «Initials» «Surname»
«Address_line_1»
«Address_line_2»
«Adress_line_3»
«Town» «County» «Postcode»
02 January 2002

Dear Dr «Surname»

I am writing again to ask for your help with the Medical Research Council sponsored survey of the use of outcomes measures by UK Psychiatrists.

We hope to gain the first really comprehensive picture of psychiatrists' experiences in using these measures. Your contribution to this project will help ensure that our results are valid and representative of those who provide mental health care in the real world.

I have enclosed a further copy of the questionnaire and a pre-paid reply envelope, in case you need this. Respondents will be entered into a prize draw to win a copy of the new Oxford Textbook of Psychiatry. Your returned questionnaire will automatically be entered into this draw.

I apologise if you have already returned the questionnaire, and these letters have 'passed' in the post.

With many thanks in anticipation.

Yours sincerely

Dr Simon Gilbody *MRCPsych*
MRC Clinical Fellow in Health Services Research

Appendix 5: Quality scoring instruments for randomised trials

Jadad Scale

Randomisation

☐ yes ☐ no

Extra point ☐ yes ☐ no

Randomisation – trials that report using the following methods are to **receive one point**: reporting that the trial was a 'randomised' one. Trials that describe (and was appropriate) the method of randomisation, such as random table numbers, computer generated, receive an additional point. However, if the report described the trial as randomised, and it was appropriate, such as date of birth, hospital numbers, **a point is deducted.**

Double blinding ☐ yes ☐ no

Extra point ☐ yes ☐ no

Double blinding – trials that report using the following methods **are to receive one point**: reporting that a trial was double blind. Trials that describe (and was appropriate) the method of double blinding, such as identical placebos, **receive an additional point**. However, if the report described the trial as double blind, and it was inappropriate, a point is deducted.

Withdrawals and dropouts

☐ yes ☐ no

Withdrawals and dropouts – trials that report using the following methods are to **receive one point**: the number and reasons for dropouts and withdrawals in each group must be stated. However, if there is no statement on withdrawals, this item must be given **no points**.

Schulz components

Randomisation generation

Adequately stated? ☐ yes ☐ no

Randomisation generation – trials that report using the following methods are to receive a point: trials that report using either a random number table, computer random number, coin tossing, dice throwing or shuffling.

Allocation concealment

Adequately stated? ☐ yes ☐ no

Allocation concealment – trials that report using either central randomisation, numbered or coded bottles or containers, or a statement indicating that drugs were prepared by a pharmacy. Serially numbered, opaque, sealed envolopes is another example of adequate allocation concealment. Reports using 'sealed envelopes' without mention of opaque are not to receive a point.

Double blinding

☐ yes ☐ no

Adequately stated

Double blinding – trials that report using the following methods are to receive a point: trails purporting to be accounts of double blinding trials.

Patrick, D. L., Bush, J. W. & Chen, M. M. (1973) Methods for measuring levels of wellbeing for a health status index. *Health Sevices Research* **8**, 229-234.

Patrick, D. L. & Deyo, R. A. (1989) Generic and disease-specific measures in assessing health status and quality of life. *Med Care* **27,** (3 Suppl), S217-32.

Petticrew, M., Gilbody, S. M. & Sheldon, T. A. (1999) Relation between hostility and coronary heart diseaese. *British Medical Journal* **319**, 917-918.

Phelan, M., Slade, M., Thornicroft, G., Dunn, D., et al (1995) The Camberwell Assessment of Need (CAN): the validity and reliability of an instrument to assess the needs of people with severe mental illness. *British Journal of Psychiatry* **167**, 589-95.

Pignone, M., Gaynes, B. G., Lohr, K., Orleans, C. T., et al (2001) Systematic review is incomplete. *British Medical Journal* **323**, 167.

Pigou, A. C. (1920) *The Economics of Welfare*. London: MacMillan.

Pincus, T., Summey, J. A., Soraci, S., Wallson, K., et al (1983) Assessment of patietn satisfaction in activities of daily living using the modified Sanford Health Assessment Questionnaire. *Arthritis and Rheumatism* **26**, 1346-1353.

Pocock, S. J. (1983) *Clinical trials: a practical approach*. London: Wiley.

Pope, A. & Tarlov, A. (Eds.). (1991) *Disability in America: toward a national agenda for prevention.* Washington, DC.: National Academy Press.

Power, M. (1997) *The audit society: rituals of verification*. Oxford: Oxford University Press.

Pynsent, P., Fairbank, J. & Carr, A. (1993) *Outcome measures in orthopaedics*. Oxford: Butterworth Heinemann.

Randolph, F. L., Blasinsky, M., Leginski, W., Parker, L. B., et al (1997) Creating integrated service systems for homeless persons with mental illness: the ACCESS programme. Access to Community Care and Effective Services and Supports. *Psychiatric Services* **48**, 369-374.

Reiger, D. A. & Kaelber, C. T. (1995) The Epidemiological Catchment Area (ECA) Programme: studying the prevalence and incidence of psychopathology(eds M. T. Tsuang, M. Tohen & G. E. P. Zahner). New York: Wiley and Sons.

Reilfer, D. R., Kessler, H. S., Bernhard, E. J., Leon, A. C., et al (1996) Impact of screening for mental health concerns on health service utilisation and functional status in primary care patients. *Archives of Internal Medicine* **156**, 2593-2599.

Relman, A. S. (1988) Assessment and accountability: the third revolution in medical care [editorial]. *N Engl J Med* **319,** (18), 1220-2.

Remmington, M. & Tyrer, P. (1979) The Social Functioning Schedule: a brief semi-structured interview. *Social Psychiatry* **14**, 151-157.