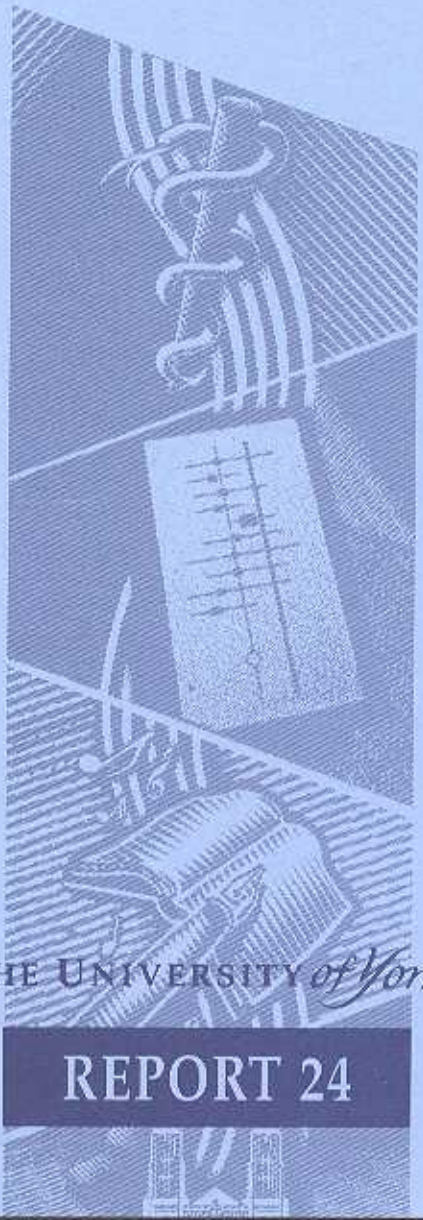
CENTRE FOR REVIEWS AND DISSEMINATION

Outcomes Measurement in Psychiatry

A critical review of outcomes measurement in psychiatric research and practice



THE UNIVERSITY of York

REPORT 24

Outcomes Measurement in Psychiatry

A critical review of outcomes measurement in psychiatric research and practice

Dr Simon M Gilbody[1]
Professor Allan O House[1]
Professor Trevor A Sheldon[2]

[1] Academic Unit of Psychiatry, University of Leeds

[2]Department of Health Sciences, University of York

July 2003

ISBN 1 900640 26 0

This report can be ordered from: The Publications Office, Centre for Reviews and Dissemination, University of York, York YO10 5DD. Telephone 01904 433648; Facsimile: 01904 433661: email: crdpub@york.ac.uk

Price £12.50

The Centre for Reviews and Dissemination is funded by the NHS Executive and the Health Departments of Wales and Northern Ireland. The views expressed in this publication are those of the authors and not necessarily those of the NHS Executive or the Health Departments of Wales or Northern Ireland.

Printed by York Publishing Services Ltd.

Centre for Reviews and Dissemination

The Centre for Reviews and Dissemination (CRD) is a facility commissioned by the Department of Health NHS Research and Development Division. Its aim is to identify and review the results of good quality health research and to disseminate actively the findings to key decision makers in the NHS and to consumers of health care services. In this way health care professionals and managers can ensure their practice reflects the best available research evidence. The reviews will cover: the effectiveness of care for particular conditions; the effectiveness of health technologies; and evidence on efficient methods of organising and delivering particular types of health care.

Further Information

General Enquiries:	01904 433634
Information Service	01904 433707
Publications:	01904 433648
Fax:	01904 433661
Email:	crd@york.ac.uk

CRD Reports

1. Which way forward for the care of critically ill children? (1995)	£7.50
2. Relationship between volume & quality of health care (1995)	£5.00
4. Undertaking systematic reviews of research on effectiveness. 2nd edition (2001)	£12.50
5. Ethnicity and health (1996)	£12.50
6. Improving access to cost-effectiveness information for health care decision making: the NHS Economic Evaluation Database. (2nd Edition 2001)	£9.50
7. A pilot study of 'Informed Choice' leaflets on positions in labour (1996)	£7.50
8. Concentration and choice in the provision of hospital services (1997)	
Summary Report	£6.00
Part I - Hospital volume and quality of health outcomes	£12.50
Part II - Volume and the scope of activity and hospital costs	£9.50
Part III - Concentration, patient accessibility and utilisation of services	£7.50
Complete set of reports	£30.00
9. Preschool vision screening: results of a systematic review (1997)	£9.50
10. Systematic review of interventions in the treatment and prevention of obesity (1997)	£12.50 £12.50
11. A systematic review of the effectiveness of interventions for managing childhood nocturnal enuresis (1997)	£12.50
13. Screening for ovarian cancer: a systematic review (1998)	£12.50
14. Women and secure psychiatric services: a literature review (1999)	£12.50
15. Systematic review of the international literature on the epidemiology of mentally disordered offenders (1999)	£12.50
16. Scoping review of literature on the health and care of mentally disordered offenders (1999)	£12.50
17. Therapeutic community effectiveness: community treatment for people with personality disorders and mentally disordered offenders (1999)	£12.50
18. A systematic review of water fluoridation (2000)	£20.00
19. The longevity of dental restorations: a systematic review (2001)	£20.00
20. Informed choice in maternity care: an evaluation of evidence based leaflets (2001)	£15.00
21. Scoping review of the effectiveness of mental health services (2001)	£12.50
22. The effectiveness of interventions used in the treatment/management of chronic fatigue syndrome and/or myalgic encephalomyelitis in adults and children (2002)	£12.50
23. Access to the online evidence base. A survey of the Northern and Yorkshire Region (2003)	£7.50

Acknowledgements

The work described in this report was conducted during the tenure of a Fellowship in Health Services Research funded by the Medical Research Council, and formed the basis of a SG's DPhil, University of York 2002.

Abbreviations

AHCPR.....Agency for Health Care Policy and Research
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
HRQoL ..... Health Related Quality of Life
NNT .....Number Needed to Treat
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

Contents

List of tables and figures

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.

Chapter 1 Outcomes measurement – an introduction

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'.[1] 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 report 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.

The rise of outcomes measurement

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'[2] 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.[3-5] What Donabedian was in fact reflecting was a resurgence of attention to the *results* of medical care. For example, Davies et al[6] 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.'

There have been important historical contributions to the measurement of health outcome and quality improvement from Florence Nightingale working in the Crimea, and Ernest Codman in Boston.[5, 7, 8]

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.[9] Particularly for chronic diseases, there has (necessarily) been a change in the way in which health and healthcare are measured and evaluated.[10] 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.[11,12] 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.[13] Where more 'patient based' measures are used to evaluate changes in health status and antecedent healthcare, then we have 'patient based measures of outcome'.[14] Some areas of medical speciality have readily incorporated or adopted patient based measures of outcome, for example oncology[15] and rheumatolgy.[16]

A major focus of the current report is to examine the measurement of *patient based outcomes*. The term cannot easily be defined,[12] 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.[17] One of the key features of patient based outcomes measures is the recognition of the fact that the *patient's perspective* is worthy of measurement in its self.[13] The patient's 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.[13] 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.[18]

A number of synonyms are used for patient based outcome: particularly quality of life, health status and health related quality of life (HRQoL). 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*.

Table 1. Components of Health Status and HRQoL Adapted from Ware[10, 22]

Physical health
Ability to perform physical activity and self care activities (e.g. eating, bathing and 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.

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.[19] These are health related, in that they are influenced by disease, injury, treatment or health policy.[20] Such items reflect states that are felt to be universally desirable.[21] 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. HRQoL 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.[20] Table 1 produces a comprehensive summary of some of the wide variety of components that have been included in operationalised measures of HRQoL.

Chapter 2 Uses of outcome measures

Outcome measures 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.[10, 20, 23-27] 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 the report 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, healthcare systems have in the past been evaluated by crude measures of activity and utilisation, rather than patient based measures of health status.[28] 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'.[126] 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.[29] 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'.[30] 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'.* [31] 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 MOS to mental healthcare evaluation in particular will be examined in detail in chapter 5.

The self completed health status questionnaires developed in the MOS eventually evolved into the SF36, which has become one of the most widely used and heavily promoted patient based outcomes measures in the 1990s.[31-37]

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.[38] 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.[39] In the UK, the Department of Health[40] 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.[41,42] 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.

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 the 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.[33] 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.[27] 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.[27]

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).[43]

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.[44] 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,[45] 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 to chose between alternative programmes for treating the same patients 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.[46]

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.[47,48] Audit has tended to use measures of process in preference to measures of outcome as the 'standards' that are measured.[49] However, the systematic measurement of outcome has been proposed as a 'standard' in audit.[50] Two scenarios whereby outcome can be usefully measured in the audit process have been identified.[51] 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.[52] 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 traditional epidemiological indices of mortality and morbidity.[53] Some authors discuss this use as an example of 'outcomes measurement',[54-56] 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 3 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.[57] 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 disorders 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[58,59] and investigations of aetiology.[60] 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,[61] 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'.[62]

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 itself to the measurement of psychopathology.[63] In a survey of over 2000 randomised trials conducted in schizophrenia, 640 scales were found 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 that 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.[64] 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[42,64] since this is a core component of the dimensions and domains considered to be integral to health related quality of life.[10,22] 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 2. These are usually (but not always) clinician or interviewer administered and rated instruments.

Table 2: Content of two common symptom based measures

Hamilton Depression rating Scale (HDRS)[61]
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[65]
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[66] which represented an attempt to rate health/sickness on a 100 point scale. Subsequent modifications include the Global Assessment Scale in 1976,[67] and the Global Assessment of Functioning scale, which forms axis V of the fourth edition of the Diagnostic and Statistical Manual - DSM-IV.[68] 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,[69] 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 3.

Table 3: An example of a global outcome measure

The Global Assessment of Functioning Scale[69]

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.[70] Since the 1960s, there has been a proliferation of instruments to measure social and role functioning.[71-72] Wiersma[70] identifies the major domains that are included in popular measures of social and role function:

- Occupational role (work, education, household, regular 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,[73] Katz Adjustment Scale,[74] Social Functioning Scale SFS,[75] and Index of activities of Daily Living.[76]

Quality of life and health related quality of life

There are a number of quality of life and health related quality of life instruments that have been developed specifically for 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,[10] such as social functioning. According to Lehman,[77] 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 Index – QOLI,[78] which is described in Table 4.

Table 4: Lehman's Quality of Life Index[78]

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 seven 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 do 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.[78]

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 and outcomes research.

Chapter 4 Outcomes measurement in clinical trials in psychiatry

Clinical trials are considered to be the most robust form of evidence in deciding what works in healthcare in general,[79] and also in mental health.[80] 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.[38, 81] 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.[79] 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.[82]

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[79] 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.

Survey aims

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

Survey methods

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.[83] 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.

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 RCTs 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 a disorder. Examples of this form of outcome measure include the Global Assessment of functioning (GAF)[68] and Global Assessment Scale (GAS).[67]

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)[84] or Sickness Impact Profile (SIP).[85]

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[10] are 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.[86]

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.

Table 5 Interventions examined in the survey of outcomes measures in clinical trials

Schizophrenia and related severe mental disorders	Depression, anxiety and related disorders
Anticholinergic medication for neuroleptic-induced tardive dyskinesia	Antidepressant drug treatment for postnatal depression
Assertive community treatment for people with severe mental disorders	Antidepressant plus benzodiazepine for major depression
Benzodiazepines for neuroleptic-induced tardive dyskinesia	Antidepressant versus placebo for depressed elderly
Beta-blocker supplementation of standard drug treatment for schizophrenia	Antidepressants for depression in people with physical illness
Carbamazepine for schizophrenia and schizoaffective psychoses	Antidepressants using active placebos
Calcium channel blockers for neuroleptic-induced tardive dyskinesia	Brief psychological interventions ("debriefing") for trauma-related symptoms and prevention of post-traumatic stress disorder
Case management for people with severe mental disorders	Cognitive behaviour therapy for adults with chronic fatigue syndrome
Chlorpromazine versus placebo for schizophrenia	Counselling for Depression in primary care
Clotiapine for acute psychotic illnesses	Drugs versus placebo for dysthymia
Cognitive behaviour therapy for schizophrenia	Lithium for maintenance treatment of mood disorders
Cognitive rehabilitation for people with schizophrenia and related conditions	Pharmacotherapy for Posttraumatic Stress Disorder
Crisis intervention for people with severe mental illnesses	Psychosocial and pharmacological treatments for deliberate self harm
Depot bromperidol decanoate for schizophrenia	SSRIs versus other antidepressants for depressive disorder
Depot fluphenazine for schizophrenia	St John's Wort for depression
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	

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,[11, 64, 87-90] 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.[91]

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.[92] 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.[93]

Survey results

In total 490 individual trials were identified. The topic area of individual reviews and their component trials is given in Table 5. 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).

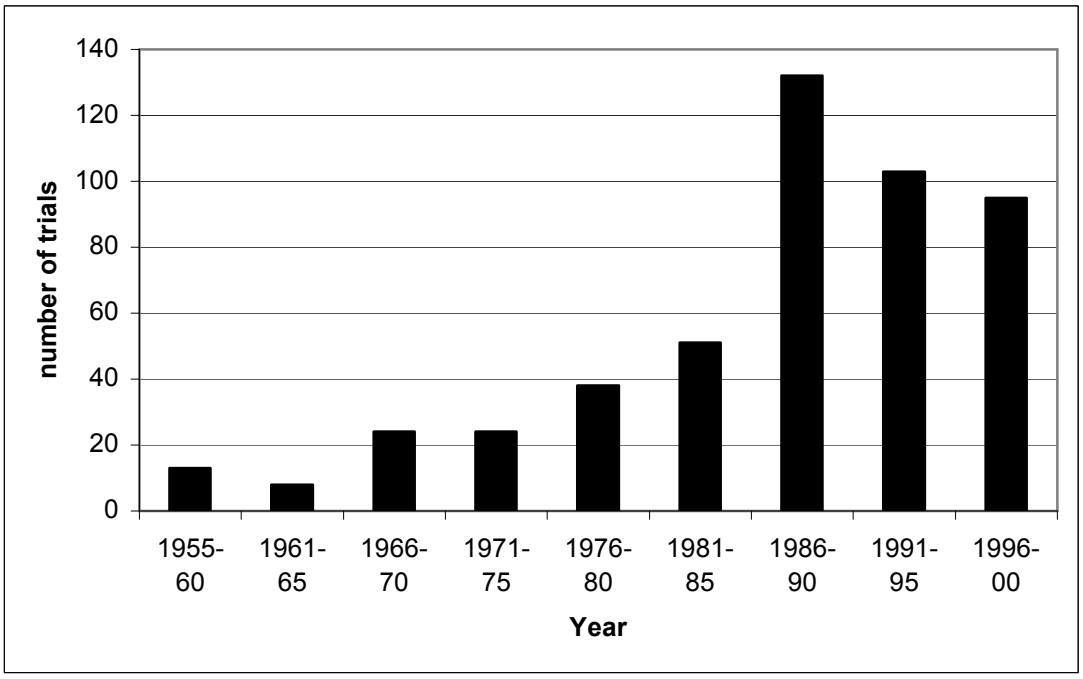


Figure 1: Publication of trials over time

General overview
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 6 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,[94] and the Brief Psychiatric Rating Scale.[65] For depression and related disorders, the most commonly used measures were: the Hamilton Depression rating scale.[61]

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,[68] and the Global Assessment Scale.[67]

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 SF36[84] – n=5, the Dartmouth COOP[95] – 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,[96] the Lehman Quality of Life Interview,[97] the Oregon Quality of Life Questionnaire – OQLQ.[98]

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,[73] Katz Adjustment Scale,[74] Social Functioning Scale SFS,[75] REHAB scale.[99]

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 7.

Table 6: 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 7: 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 [73] – n=18; Katz Adjustment Scale [74] – n=5; Social Functioning Scale SFS [75] – n=3; REHAB scale [99] – n=2.
	Role functioning	8/233	Index of activities of Daily Living [76] – n=6
	Perceptions of well-being	8/233	Rosenberg's Self Esteem Scale [100] – n=6
	Cognitive functioning	10/233	IQ and intelligence tests - various
	Satisfaction	6/233	Client Satisfaction Questionnaire [101] – n=4
Depression, anxiety and related disorders (n=257)	**Physical Health**	1/257	Pain and Disability Index [102]
	Social Functioning	9/257	SAS [73] – n=6; Katz Adjustment Scale [74] – n=3
	Role functioning	4/257	Karnofsky Perfomance index [103] – n=2
	Perceptions of well-being	nil	none
	Cognitive functioning	2/257	IQ and intelligence tests - various
	Satisfaction	1/257	Client Satisfaction Questionnaire [101] – n=1

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).

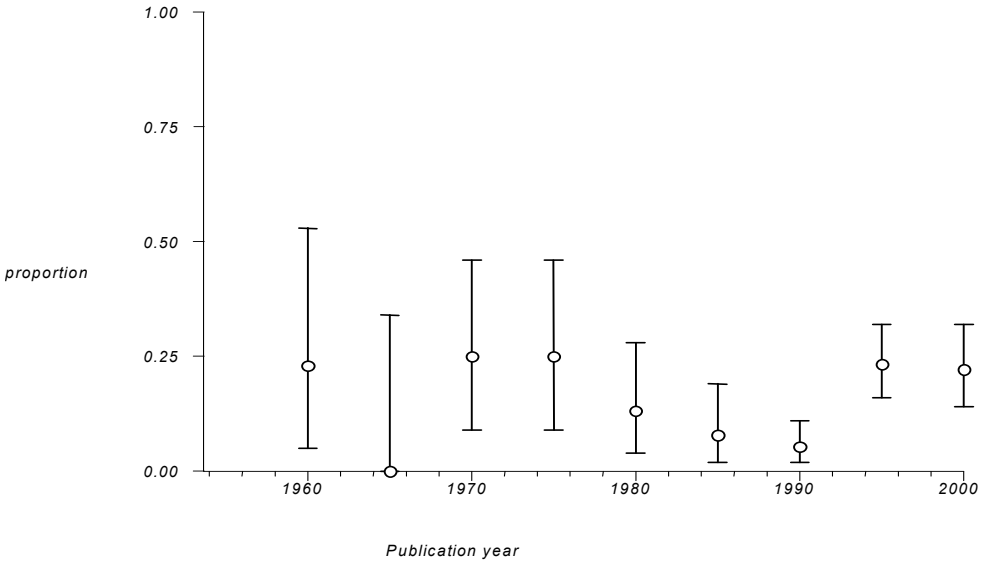


Figure 2: Proportion of trials using a patient based outcome measure, measured over time
(Five-year periods indicated, together with 95% confidence intervals)

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. [61] The use of symptom based measures in drug trials is discussed below.

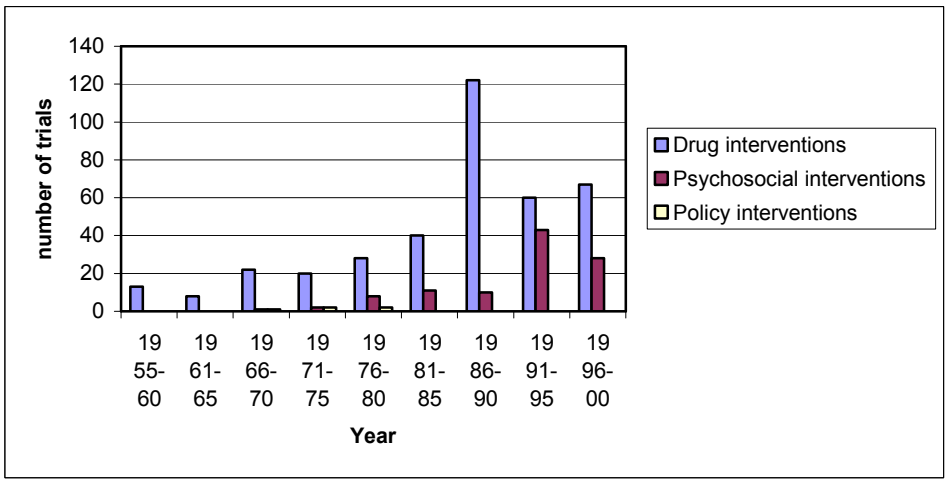


Figure 3: Type of intervention in trials measured over time

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: first, the evaluation of new drugs for schizophrenia; second, the evaluation of new drugs for depression; and third, 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.[104] 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,[105-112] 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.[113] 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 SF36, 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.[114,115] 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[78] 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.

Discussion
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 SF36 – 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[42] in a survey of trials included in the Cochrane Controlled Trials Register[116] 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[70] 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 present survey therefore lends empirical support to observations that have been made previously, for example in the sphere of schizophrenia research, Collins et al[117] have 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.[118] 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 (effectiveness), 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 a product licence.[119] 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[120] 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.[78] Generic patient based outcomes measures, such as the Nottingham health Profile or the SF36, 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[120] 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.

Chapter 5 Outcomes research in psychiatry

RCTs have generally been accepted as the 'gold standard' design when deciding what interventions work in psychiatry.[80] Most randomised studies in psychiatry have investigated the effect of drug or psychotherapy interventions in tightly controlled and largely artificial experimental conditions,[63,121] 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.[122] Important questions relating to the organisation and delivery of mental health services are also rarely addressed in randomised trials.[123]

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*.[124] Another has been to synthesise various data sources using *decision analysis*.[125] 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*.[126-128]

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.[126] 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. According 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.[129] 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'.*[130] It was decreed by Congress (via the General Accounting Office[131]) that new primary research conducted by the PORTs was not to be the traditional RCT, rather it was to be observational in design, utilising the vast amounts of data routinely collected on US patients. This health research policy produced a new breed of health researchers known as database analysts,[127, 132] with the motto *'Happiness is a humongous database'*.[133]

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.[123] Quasi-experimental studies are therefore more like randomised trials, and are considered to be clearly different in their approach and ethos to outcomes research.[134] 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 paper.

Enthusiasm for outcomes research has, in the US, led to the establishment by the American Psychiatric Association of Practice Research Networks – PRNs.[135, 136] 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.[137-140] 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,[141] 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.[142]

As in the surveys of randomised trials and economic evaluations 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.

Survey 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.

Survey methods

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 the Appendix.

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,[143] 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.[144] For example, the PRiSM Psychosis study[145] 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.[146]

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

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.

Results

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 8. In the following section important characteristics of these nine studies are reviewed.

Table 8: 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 and sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Medical Outcomes Study (MOS)[147]	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. SF36).	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[150]	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).[164]	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.
Rosenheck et al[149]	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[151]	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 8: 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 and sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Rosenheck et al[156]	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. Complete 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[163]). 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.
Melfi et al[152]	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 month 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.
Croghan et al[154]	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 months is taken to be a proxy measure of effective antidepressant therapy and good outcome.[165] 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 8: 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 and sample size	Outcomes studied	Methods used in adjusting for case mix	Results
Hylan et al[155]	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 prescription.	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[153]	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. They 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.

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.[33,147,148] 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.[148]

Three other studies researched health policy and organisation questions, such as the consequences of the withdrawal of mental health benefits from insurance plans,[149] the effectiveness services directed at homeless persons,[150] and the difference in outcome between private and publicly funded health providers.[151]

(2) The evaluation of new technologies.
Four studies,[152-155] utilised an outcomes research design to demonstrate the worth of new antidepressants and anti-psychotics in routine care settings. One further study[156] 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 SF36,[157] now the most commonly used generic measure of health related quality of life.

A further study,[156] 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,[158] supplemented with routinely collected disease-specific patient outcomes measures collected for all patients in receipt of care for PTSD.[159]

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.[152,154,155] 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.[160] Similarly, depression is consistently under-identified by clinicians,[161] and mislabelled or underreported, in part as a consequence of the stigma of mental illness.[162]

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[155] 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,[147] achieved smaller sample sizes (n=1772) than those which selected subjects retrospectively from large existing datasets[149, 154] - 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.[163] 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.[122, 148]

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[152,155] 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).

Discussion
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'.[166]

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.[167] The consequence is that it is not always possible to apply the results in clinical practice – that is, trials lack external validity.[168]

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.[126] 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.[142] 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?'".[169]

The other major problem with outcomes research, as with all observational research, is the problem of confounding and selection bias.[142,144] 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.[170]

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,[147] and Rosenheck's study of PTSD.[171]

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.[172] The most ambitious example of this work in the UK has been in intensive care.[173] 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[174] in order to generate 'practice based evidence' of effectiveness framed within routine services.[139] 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[171], who provide one of the more rigorous examples of outcomes research, outline several ingredients of a successful clinical outcomes 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.[148] 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.[175-178]

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.[171] 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 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.

Chapter 6 Measuring outcome in psychiatric practice – a survey of UK consultant psychiatrists

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'.[179]

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.[10, 21, 26, 27]

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.[180] 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,[179] and the National Health and Community Care Act[181] 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.

Methods

A questionnaire survey of consultant psychiatrists practising in the UK was conducted. Since there are approximately 4000 general adult psychiatrists practising in the UK,[182] 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 and Kalton[183] and Fowler.[184]

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.[185] This is a commercially available resource, updated and published annually, containing the details of all medical practitioners listed in the medical register held by the General Medical Council. In compiling and updating the Medical Directory, all practitioners 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.

Sampling procedure
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.[186]

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.

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 and 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[184] and Dilman,[187] and summarised in a recent systematic review by McColl et al.[188] The questionnaire was extensively piloted.

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:

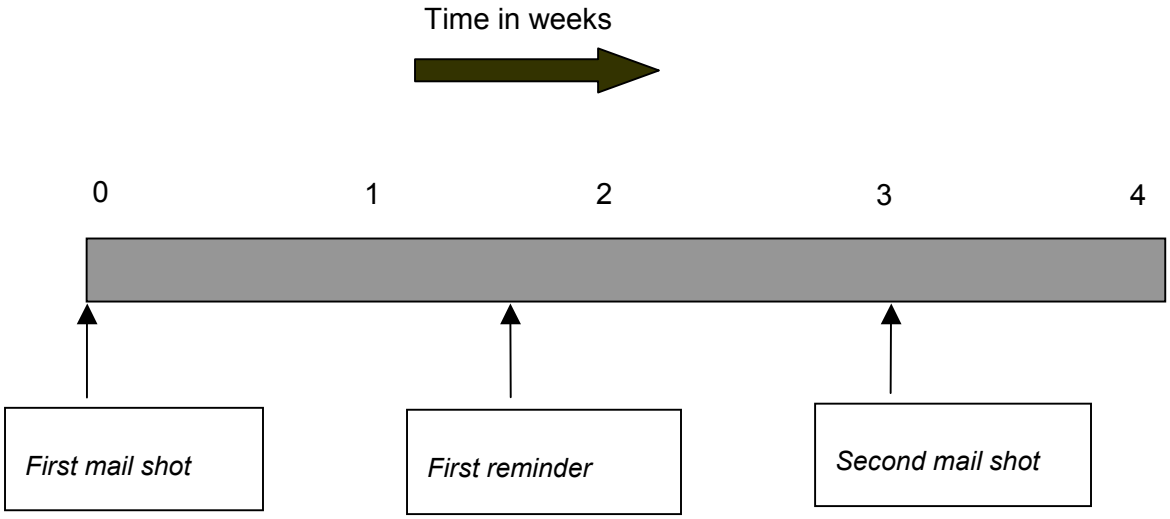


Figure 4: Time scale of the survey

All data were entered into a specifically designed Microsoft Access relational database,[91] and respondents were sequentially eliminated from the mail address database. Reminders and second mail shots were sent only to non-respondents.

Results of the survey

Questionnaire responses

In total, 369 (74%) of the 500 questionnaires were returned in a six-week period. 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).

Twenty nine of the 369 returned 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 9. 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 9: Specialities of respondents

Speciality	Main speciality	Sub specialty/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[189] (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[94] (25/340), the Health of the Nation Outcome Scale – HoNOS[190] (25/340), and the Brief Psychiatric Rating Scale[65] (17/340). For drugs and alcohol problems, the most commonly reported measure was the CAGE questionnaire[191] (10/340).

The most commonly used instruments are given in Tables 10-13.

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% 995%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[190] (20/340), the Social Adjustment Scale[73] (9/340), and the

Social Functioning Schedule[75] (5/340). For schizophrenia/psychosis, the most commonly reported measures were the PANSS (20/340), the BPRS (13/340), and the HoNOS[190] (16/340). For cognitive impairment and drugs and alcohol problems, the most commonly reported measure was the HoNOS (13/340 and 12/340 respectively).

The most commonly used instruments are given in Tables 10-13.

Table 10: The use of questionnaires for depression and 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 11: 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 12: 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%	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 13: 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 – 16%%	HoNOS – 12/37; SAS – 2/37; MRC – 1/37; GAF – 1/37
Measuring therapeutic response in drugs and alcohol problems	29/340	HoNOS – 10/29; Maudsley Addictive Profile – 2/29
Clinical audit of drugs and alcohol problems	30/340	HoNOS – 8/30; ARPQ – 2/30; Maudsley Addictive Profile – 2/30; QLI – 1/30

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[192] (49/340), HAD[193] (41/340), HDRS[61] (23/340), and the HoNOS (18/340). The most commonly used measure in the case of schizophrenia/psychosis were the PANSS[94] (20/340), the BPRS[65] (13/340), and the HoNOS[190] (16/340). The most commonly used questionnaires in the case of cognitive impairment was the MMSE[189] 60/340, and the HoNOS[190] (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[190] (10/340).

The most commonly used instruments are given in Tables 10-13.

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[192] (18/340), the HoNOS[190] (18/340), the HDRS[61] (13/340), and the HAD[193] (12/340). For those 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[190] (24/340), the PANSS[94] (6/340), and the BPRS[65] (8/340). Standardised measures were very rarely used for those with cognitive impairment or drugs or alcohol problems.

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 (Table 14).

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 14: 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). Clinicians were specifically questioned about being asked or required to collect the Health of the Nation Outcome Scale,[190] 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.

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 (adolescent) 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.'

. **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)

Forty 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 and 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 and 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 beach 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=22). 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 and 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)
Twenty-six 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)

Fifteen 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'

Discussion of the main results of the survey

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 to adopt measures such as the HoNOS and needs assessment tools.[194,195] 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'.[195] The desirability of these 'performance indicators' as the main measurement of success or failure is debatable.[196,197] Of particular concern is that these figures are the easiest to manipulate or 'improve', without conferring any overall health gain on the population or service under consideration.[198] 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'.[196, 199] 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.[79] 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,[200] with little consideration of basic epidemiological principles when judging the results of such studies.[201] Better-resourced attempts have also been made to use routinely collected data in order to judge the effectiveness of new technologies in general,[141] 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.[202, 203] 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,[204] and this instrument has specifically failed to show sensitivity to change in the underlying condition,[205] 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[180] 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',[126] 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. Slade et al, Marks[180, 206]) 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.[179] It was also developed by psychiatrists' own professional organisation – the Royal College of Psychiatrists.[190] It forms a central component of the most recent major policy document in mental health,[207] which stipulates that a minimum data set[194] 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,[208] 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.[180] 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[204] 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[209] provide a useful insight into the real value of this instrument when adopted on a service wide basis as a routine outcomes instrument. The HoNOS questionnaire was routinely administered 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.'

Chapter 7 Does routine outcome measurement improve outcome in mental illness? A systematic review

The previous chapter 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 report 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,[210] 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.[211,212] 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.[95]

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[213, 214] 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[215] 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.[216]

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.[217] Examples of such questionnaires include the General Health Questionnaire,[218] and the Beck Depression Inventory[192] and the Hospital Anxiety and Depression scale.[193] 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.[219] 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 SF36 (SF36) contain items and sub-scales which measure 'psychological well-being'.[31] 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.[34,35] Psychological well being is in fact a core component of many 'health status measures'[10, 64] 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.[220] 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,[206] 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.[147, 221] 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.[222, 223] 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.[224] Consequently, clinicians' perceptions of these problems are often poor, and they underestimate the health status or health related quality of life of patients when patient and clinician ratings are compared.[78, 97, 223, 225] The use of more comprehensive outcomes 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'.*[179] 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.[204] 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.[190] 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,[226] such that healthcare provision might be more rational, responsive and 'appropriate'.[227, 228] Examples of individual patient needs assessment tools for use in severe mental illness include the Camberwell Assessment of Need (CAN),[229] MRC Needs for Care Assessment.[230]

Possible disadvantages of routine outcome measurement

The routine measurement of outcome has not been without its critics,[49,197] 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 examines the benefits of informing clinicians of their patients' health status scores.[212] 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)[231] or modified Health Assessment Questionnaire (MHAQ).[232] 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.[220,233,234] 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.[203] If they are not sensitive to change, then their results will not be interpretable and important changes will not be detected or acted upon.[235] 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.[202, 236]

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.[27]

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.

Methods of the review

The methods employed in this systematic review follow guidelines laid down by CRD report 4 (second edition),[237] and adhere to methods outlined in the Handbook of the Cochrane Collaboration.[238] The review was conducted under the auspices of both the Depression and Anxiety Group, and Schizophrenia Group of the Cochrane Collaboration and has been published in the Electronic Cochrane Library.[239, 240] Part of this review has also been published in paper format.[241]

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.[10]

c. **An assessment tool measuring 'patient need'**. These tools measure unmet emotional, physical, social and financial needs of the individual patient,[226, 230, 242] 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,[243] rather than the 'evaluative' approach which is inherent in the perspective of 'outcome measurement'.[14, 244] 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 into 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)[190] 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-RCTs 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.[245] 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).

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 databases were searched to April 2002: 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 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 1 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[246] (see Appendix 1 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.

It will be seen from the appendix, 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.[91]

- Author and Year
- Design
- Population
- Setting
- Sample size
- Routine outcome measure used
- Intervention and control conditions
- Length of follow up and 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,[247] the criteria of Schulz[248] and Cochrane criteria.[238] 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.

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.[249]

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),[238] and confidence intervals for rate differences and ratios were calculated using Stats Direct version 1.7.[93]

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.[250]

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 was conducted[251] using STATA version 6.0.[186] This method can be used to pool both dichotomous and continuous data, and weights studies by their individual variance or sample size. 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.[237] This was supplemented by formal statistical tests such as Cochran's Q statistic.[252] 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,[253] calculated using STATA version 6.0.

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[254] following the publication of an earlier version of this review.[241] Of these, twenty-four studies fulfilled the inclusion criteria (see Table 15). The flow of studies through the review is summarised in Figure 5, according to guidelines laid down in the QUOROM statement on the reporting of systematic reviews and meta-analyses.[255]

The literature searches failed to find any trials of the routine use of outcome measures in psychiatric settings. Twenty four 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 16.

Table 15: 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

Primary care and general hospital studies

In total 24 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 and 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.[248] 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,[256] or according to alternate allocation.[257] 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.[258, 259]

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.[260] 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[259] [261-268], 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'.[269]

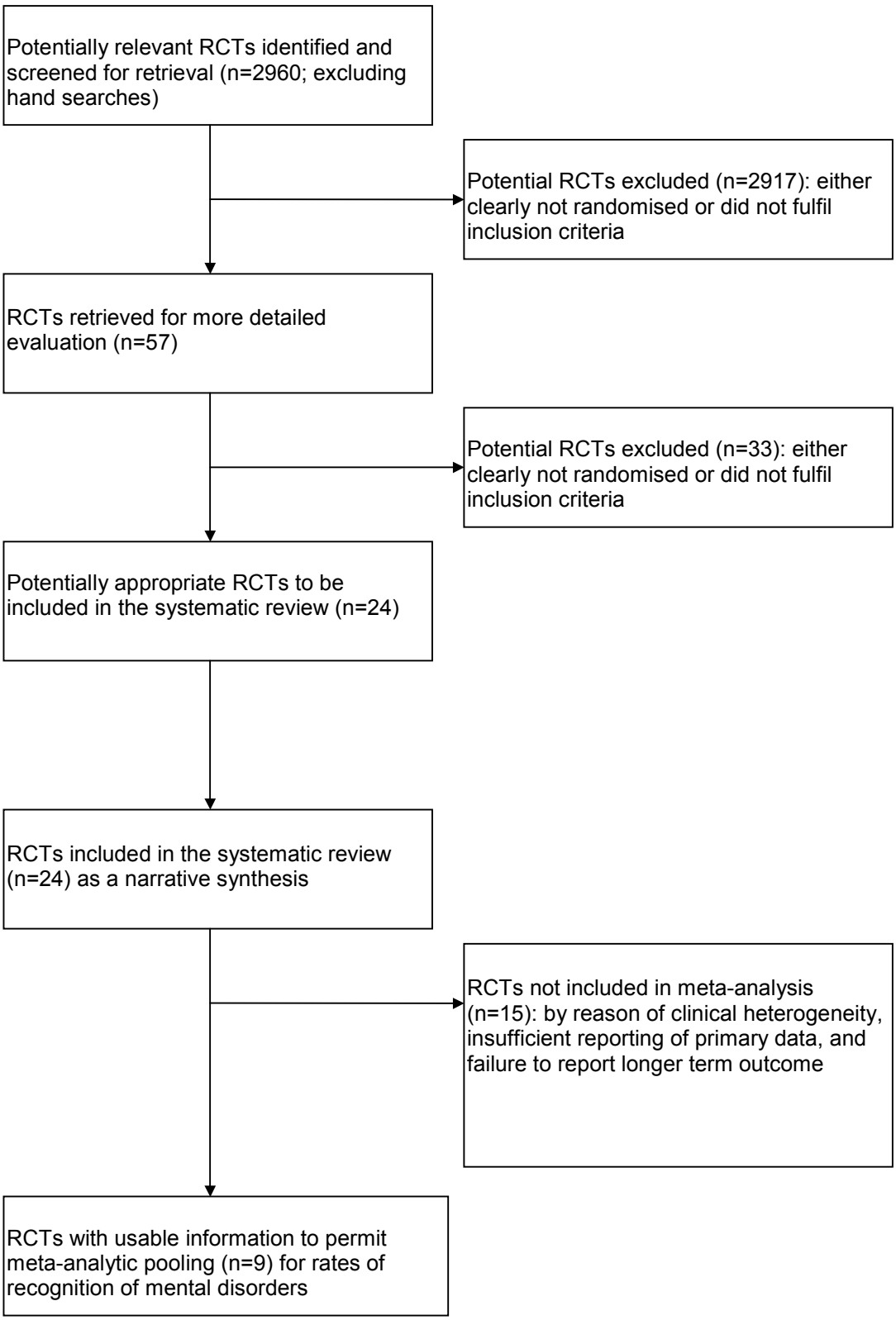


Figure 5: QUOROM Trial flow diagram[255]

Sample size varied between 52 and 2209, and three studies,[257,270,271] 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 Hooper et al,[272] 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 et al,[273] 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,[274] 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,[273] included a greater proportion of elderly patients, or were specifically targeted at elderly patients.[268]

Outcome instrument used
The most commonly used instruments were self-completed scales designed to detect depression and anxiety (Beck Depression Inventory – BDI,[192] General Health Questionnaire – GHQ,[218] Zung SDI.[274] Eight studies[212, 259, 261, 262, 264-266, 275] investigated the use of generic health status measures: the SF36[84] the functional status questionnaire FSQ[276], the Dartmouth COOP[277], and the Sickness Impact Profile SIP.[278] One study[279] combined an anxiety questionnaire, the anxiety components of the Symptom Check List 90,[280, 281], with a generic health status questionnaire (the SF36).[84] Another study[267] used a self administered diagnostic interview schedule,[282] 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,[283] 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 as 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[263] 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,[263, 266] 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,[263] 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[284] 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[256] 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 and Golberg,[256] the information was fed back following the clinical encounter, and in another Linn and Yager,[284] 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[212, 262, 263, 265] 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,[256, 267, 270, 279, 283, 285, 286] 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 and 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.
- 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 and Goldberg.[256] 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.[273] 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.[212, 256, 262, 263, 265, 267, 270, 279, 283, 285, 286] For example Johnstone and Goldberg[256] administered the GHQ to GP attenders and measured the changes in these scores at twelve months in both intervention and control

groups. Similarly, Dowrick and Buchan[270] assessed the effect of feedback of the BDI on subsequent BDI scores in both intervention and control groups. Lastly, Kazis et al[212] 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,[259] 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 and Goldberg[256] 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[272] 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 and Goldberg[256] and Hoeper et al[272] have important differences in terms of participating clinicians, mode of feedback of outcome measure and identification of psychiatric morbidity. Johnstone and Goldberg[256] 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 and Goldberg[256] 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[283] 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 study[284] 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[271] 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[273, 286, 287] 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[263] 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,[212, 259, 261, 262, 264-266, 275] four report the effect of these measures alone in improving the overall rate of recognition of emotional problems.[212, 262, 265, 266] Three of the four studies[212, 262, 265] show no differences for any subscale of the FSQ or AIMS (including mental health) at 12 months. In contrast, a later study by Rubestein et al[266] 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.[256, 258, 268, 271-273, 283, 284, 286-288]

Two studies[256, 288] 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[284] 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[271] 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[268] 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 et al[258] 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 (see Figure 6). 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,[272, 283] 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 7), where the funnel is found to be substantially asymmetrical (p=0.024 using Egger's test.[253]

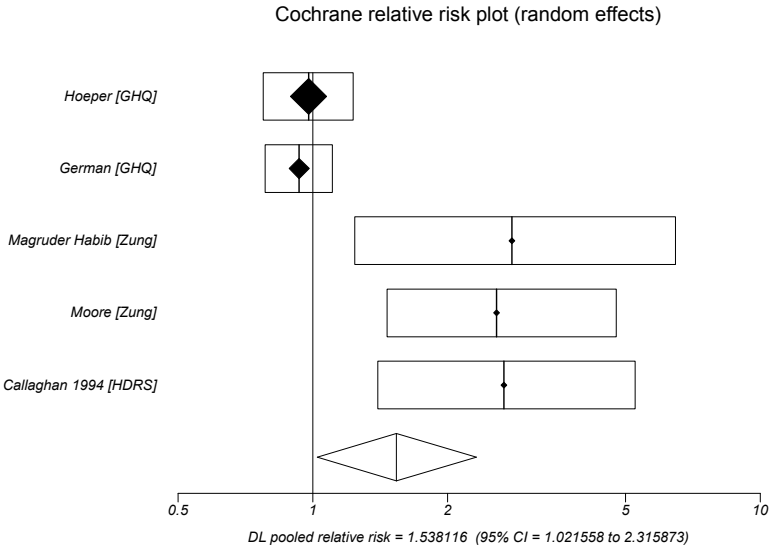


Figure 6: Forrest plot for studies examining the effect of feedback on the rate of recognition of depression

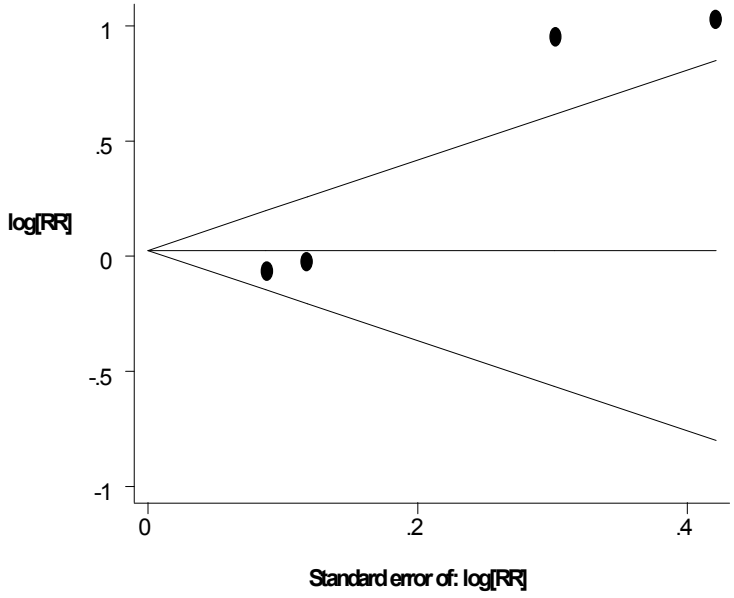


Figure 7: Funnel graph of studies examining the effect of feedback on the rate of recognition of depression

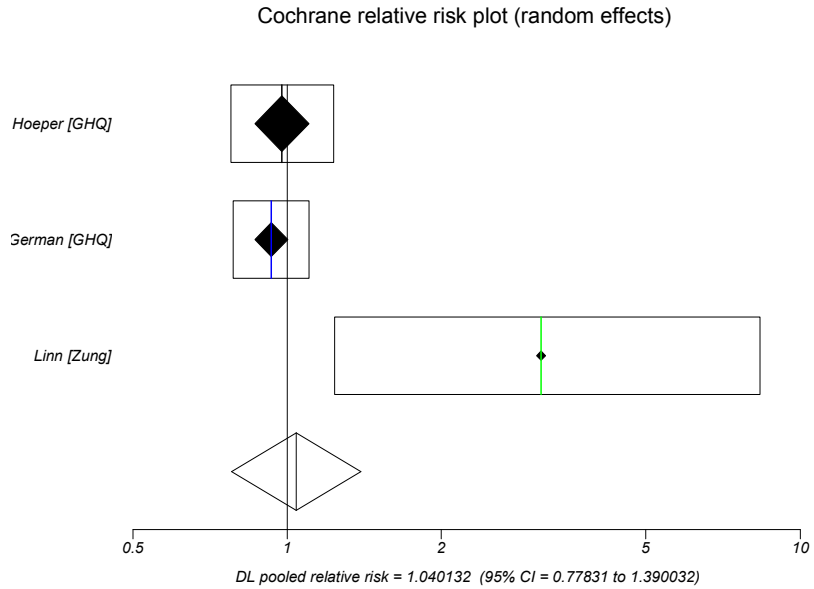
It was noted in the previous discussion that two broadly different strategies were employed in the provision of feedback in the trials included in this review: Firstly, an *unselected* form of feedback, whereby outcomes measures were administered to all patients seen in a clinical service, and their 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 6) and funnel plot (Figure 7) 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 6. 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 Hoeper et al and German et al[272, 283] 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).

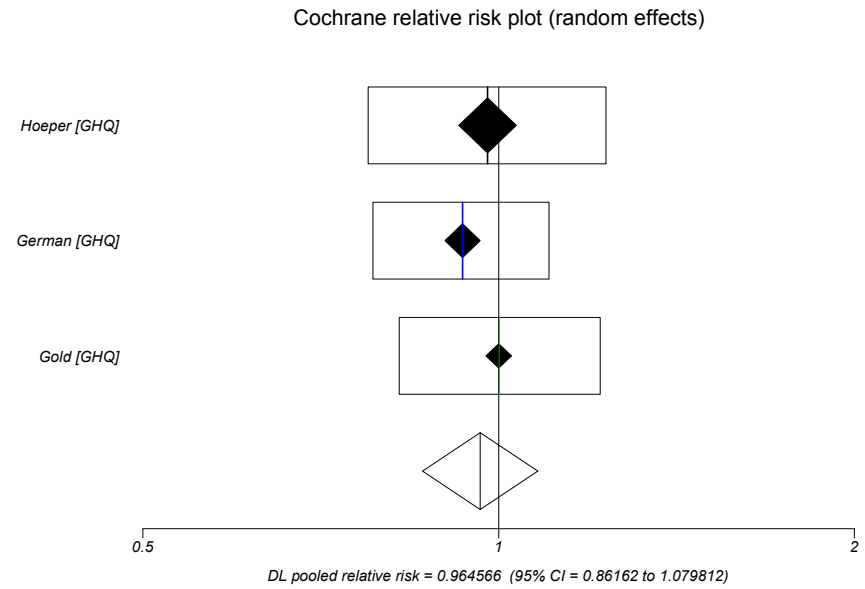
One further study[284] also used an unselected approach, but was excluded from the main analysis due to the questionable validity of pooling five separate arms, which each used a different variation of the timing and mode of feedback of questionnaire results. This study was reintroduced into the preceding meta-analysis in order to test the robustness of the overall result to the inclusion of this positive study (Figure 8). 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 et al[258] used an unselected approach. The introduction of this study did not alter the overall conclusion of the meta-analysis (Figure 9) - DerSimonian-Laird pooled relative risk = 0.97, 95% CI = 0.86 to 1.08 (Q = 0.298, df = 2, p = 0.861).



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

Figure 8: Meta-analysis of studies employing unselected feedback, with the inclusion of Linn et al, as a sensitivity analysis

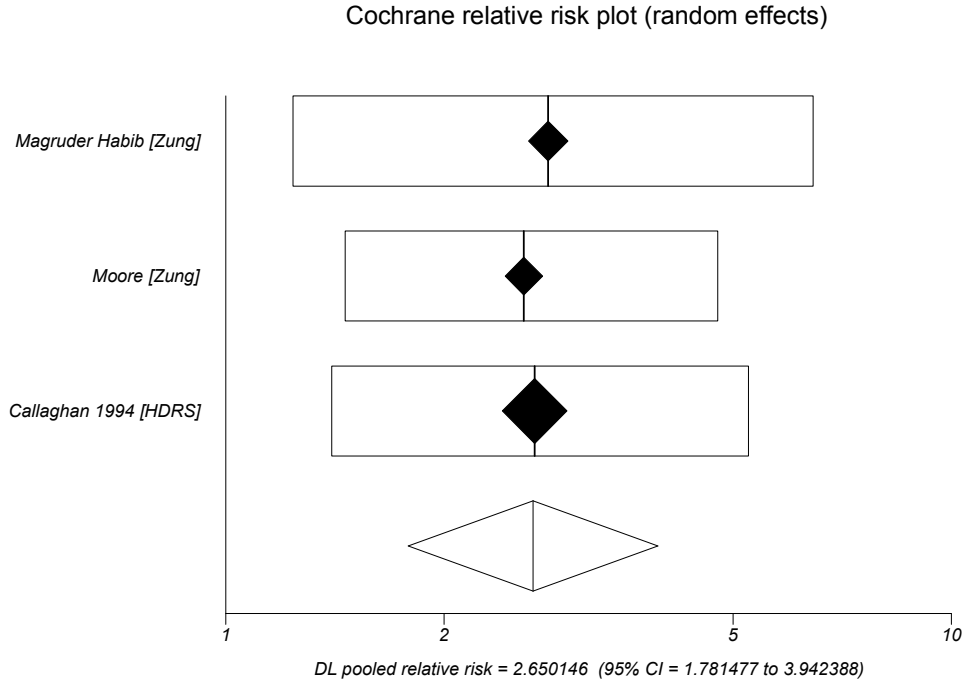


Q ("non-combinability" for relative risk) = 0.298 (df = 2) P = 0.861

Figure 9: Meta-analysis of studies employing unselected feedback, with the inclusion of Gold et al, 1989, as a sensitivity analysis

Meta-analysis of high risk feedback studies

Meta-analysis of the three studies by Moore et al[287], Magruder Habib et al[273] and Callahan et al[286] show that this high risk strategy was 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 10. 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.



Q ("non-combinability" for relative risk) = 0.020 (df = 1) P = 0.887.

Figure 10: Meta-analysis of studies employing *high-risk* feedback

Effect of routine outcome measurement on initiation of treatment for emotional problems

Ten studies investigated the effect the feedback of questionnaire results on the rate of intervention for emotional problems[257, 263, 268, 271, 273, 283, 284, 286, 289, 290] and all but two[273, 286] 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[284] showed that feedback increased the rate at which clinicians recognised depression, 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[283] 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,[283] 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[263] 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.[212, 256, 262, 263, 266-268, 270, 271, 286, 290] Results from Johnstone et al[256] 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 and Buchan,[270] who re-administered the Beck Depression Inventory at 12 months, 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[290] 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 and Goldberg[263] 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[266] 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 and Goldberg[256] 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[290] which also showed that rates of psychiatric and non-psychiatric referrals were unchanged as a result of feedback.

The study by Mazonson et al[263] 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[267] 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[259] examined the effect of the administration and feedback of the generic health status questionnaire, the SF36, on 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 Relfer et al[267] 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[271] 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[263] 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 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Callahan et al 1994[286]	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[270]	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.
German et al 1987, Shapiro et al 1987[283, 285]	RCT Individual patients randomised.	US adult and 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[258]	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%).

Table 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Hoeper et al 1984[272]	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%).
Johnstone & Goldberg 1976[256]	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 and 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% and 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 from feedback. Low scores resolve spontaneously, irrespective of feedback.
Linn and Yager 1980a[284] and Linn and Yager 1980b[291]	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 clinician following consultation, no impression of depression sought. Int 5.: no screening by SDS, but impression of depression sought. Cont.: no screeing by SDS, no physician 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 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Lewis et al 1996[290]	RCT Individual patients randomised.	UK General practice attenders at a single practice with GHQ-12 score >2.	GHQ-12 and computerised assessment of psychiatric symptomatology	Int 1: GHQ administered and 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 and clinician attribution of encounters as due to psychological or physical problems. Prescription of a psychotropic 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 psychotropic 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[273]	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.
Mathias et al 1994, Mazonson et al 1994[263, 279]	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).

Table 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Moore et al 1978[287]	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[267]	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 SF36. 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.
Weatherall 2000[257]	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).

Table 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Williams et al 1999[271]	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.[292]	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% and 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.
Whooley et al 2000[268]	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 addition, clinic attenders screened positive were offered a series of organised educational sessions. Cont: GDS administered, 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 utilisation (number of clinic visits and hospitalisations). 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.

Table 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Zung et al 1983[288]	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 and 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.
Calkins et al 1994[265]	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[261]	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.

Table 16: 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 and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Kazis et al 1990[212]	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.
Rubenstein et al 1989[262]	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[266]	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 16: Studies that evaluate the use of routine outcome measures for psychiatric disorders in primary care and general hospital settings (continued)

Author and Year	Design	Population, setting and sample size	Routine outcome measure used	Intervention and control conditions	Length of follow up and outcomes studied	Results
Street et al 1994[259]	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 and 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[275]	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[264]	RCT Individual clinicians randomised in blocks according to patient demo-graphics.	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 clinician.	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.

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 areahave produced contradictory recommendations without any clear indication as to how their authors have arrived at their conclusions.[217, 293] 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. 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 be 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[290] contains unreported data on the rates of recognition of depression by general practitioners, 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.[294]

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 methods 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.[295] 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[253] 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[296] 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.[245] 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 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 et al)[297] 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 from 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.[298] 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',[269] 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 and colleagues[297] 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.[245] 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 et al, Mantel et al,[251, 299] 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 has not been found to influence actual clinician behaviour.[218,300] 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[266]). 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 measures 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.[79] 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').[272] 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.[219]

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.[215] 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 are therefore not supported.[293]

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.[263] 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.[301]

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[283] 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 which can prevent or delay the detection of mood disorders.[302] 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.[26, 27]

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.[303, 304] 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.[271] 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.[79] 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.[270] 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.[266] 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.[305] Generic outcomes measures are essentially designed to evaluate healthcare and to identify need at a *population* level, and extrapolation of use beyond this is not justified.[10]

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) [190] are dependant upon individual clinicians collecting these data in the context of routine care.[204] 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

Chapter 8 Overall discussion of outcomes measurement in psychiatry

The report 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 report 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 report 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.[79] 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 clinicians, 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 report did not explore what outcomes clinicians would like to see collected in evaluative research. This is a 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 has not permeated clinical practice. [179, 180, 195, 206, 207] 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,[180] 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 report 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 is 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 Feinstein[306] 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.[306]

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.[190] 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'.[194]

The survey of UK consultants in this report 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.[195, 207] 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.[50] 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.[307] 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 and Dawson,[307] in their exploration of the impact of routinely collected outcomes in changing clinical practice and in improving the quality of care, highlight 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.

Distortions of the behaviour of organisations occurs when there is a pre-occupation with a small number of easy to measure outcomes indicators.[196,199] 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 17. Davies and Crombie[197] have highlighted the need for studies that examine the impact on organisations and individuals of the regular feedback of outcomes data.

Table 17: Perverse consequences of a limited focus on outcomes measures[197,199]

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 here. 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 (described below) 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 earlier, 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[311] 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 RCTs 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.[128] The present report 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 this technique 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, [122,139] 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 here. 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.[120,308] Table 18 provides the key properties and dimensions which must be satisfied by patient based outcomes measures in order that be used within clinical trials.

Table 18: Essential properties of a patient based outcome measure[120]

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 matter 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.[309, 310] 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 report. 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, which was not supportive of this strategy.

Appendix: Search strategies

The search strategies employed in this report 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.

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 nursing 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 from 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

References

1. Relman AS. Assessment and accountability: the third revolution in medical care [editorial]. *New England Journal of Medicine* 1988,319:1220-2.

2. Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q* 1966,44:Suppl:166-206.

3. Scroeder SA. Outcome assessment 70 years later: are we ready? *New England Journal of Medicine* 1987,316:160-162.

4. Lohr KN. Outcome measurement: concepts and questions. *Inquiry* 1988,25:37-50.

5. Brookes RG. Heath Status Measurement: a perspective on change. Basingstoke: McMillan, 1995.

6. Davies AE, Doyle MA, Lansky D. Outcomes assessment in clinical settings: consensus statement on principles and best practices in project management. *Journal of Quality Improvement* 1994,20:6-16.

7. Donabedian A. The end result of health-care: Ernest Codman's contribution to quality assessment and beyond. *The Millbank Quarterly* 1989;67:233-56.

8. Rosser R. The history of health realted quality of life in 10 and 1/2 paragraphs. *Journal of the Royal Society of Medicine* 1993;86:315-318.

9. Ebrahim S. Clinical and public health perspectives and applications of health-related quality of life measurement. *Social Science & Medicine* 1995;41:1383-94.

10. Ware JE. The status of health assessment in 1994. *Annual Review of Public Health* 1995;16:327-354.

11. McDowell I, Newell C. Measuring Health: A guide to rating scales and questionnaires. Oxford: Oxford University Press; 1996.

12. McDaniel RW, Bach CA. Quality of life: a concept. *Nursing Research* 1994;3:18-22.

13. Fitzpatrick R, Hinton J, Newman S, Scambler G, Thompson J, editors. The Experience of Illness. London: Tavistock Publications; 1984.

14. Jenkinson C. Measuring health and medical outcomes: an overview. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994.

15. Selby P. Measuring the quality of life of patients with cancer. In: Walker SR, Rosser RM, editors. Quality of Life Assessment: Key issues for the 1990s. Dordecht: Kluwer Academic; 1993.

16. Liang MH, Katz JN. Measurement of outcome in rheumatoid arthritis. *Baillieres Clinical Rheumatology* 1992;6:23-37.

17. Mor V, Guadagnoli E. Quality of life measures: a psychometric tower of Babel. *Journal of Clinical Epidemiology* 1988;41:1055-1058.

18. Albrecht GL. Subjective health assessment. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994.

19. Testa MA, Nackley JF. Methods for quality-of-life studies. *Annual Review of Public Health* 1994;15:535-59.

20. Patrick DL, Bergner M. Measurement of health status in the 1990s. *Annual Review of Public Health* 1990;11:165-83.

21. Faden R, Leplege A. Assessing quality of life. Moral implications for clinical practice. *Medical Care* 1992;30:Ms166-75.

22. Ware JE. Standards for validating health measures: definition and content. *Journal of Chronic Diseases* 1987;40:473-480.

23. Kane RA, Kane RL. Assessing the Elderly: A Practical Guide to Measurement. Toronto: Lexington Books; 1981.

24. Steinwachs DM. Application of health status measures in policy research. *Medical Care* 1989;27.

25. Nelson E, Berwick D. The measurement of health status in clinical practice. *Medical Care* 1989;27:S77-S90.

26. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and issues in assessment. *British Medical Journal* 1992;305:1074-7.

27. Fitzpatrick R. Applications of health status measures. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994. pp. 27-41.

28. Kindig DA. Purchasing Population Health. Ann Arbour, MI.: The University of Mitchigan Press; 1977.

29. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of wellbeing for a health status index. *Health Sevices Research* 1973;8:229-234.

30. Brook RH, Ware JE, Jr., Rogers WH, Keeler EB, Davies AR, Donald CA, et al. Does free care improve adults' health? Results from a randomized controlled trial. *New England Journal of Medicine* 1983;309:1426-34.

31. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care* 1992;30:473-83.

32. Wright L. The long and the short of it: the development of the SF-36 general health survey. In: Jenkinson C, editor. Measuring health and medical outcomes. London: UCL Press; 1994.

33. Tarlov AR, Ware JE, Jr., Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association* 1989;262:925-30.

34. McHorney CA, Ware JE, Jr. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. *Medical Care* 1995;33:15-28.

35. McHorney CA, Ware JE, Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care* 1993;31:247-63.

36. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. *Health Economics* 1993;2:217-27.

37. Brazier J. The SF-36 health survey questionnaire--a tool for economists [comment]. *Health Economics* 1993;2:213-5.

38. Guyatt GH, Sackett DL, Cook DJ. 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* 1993;270:1232-7.

39. Guyatt G, Feeny D, Patrick D. Issues in quality-of-life measurement in clinical trials. *Controlled Clinical Trials* 1991;12:81s-90s.

40. Department of Health. Assessing the Effects of Health Technologies: Principles, Practice and Proposals. Advisory Group on Health Technologies. London: HMSO; 1992.

41. Aaranson N. Quality of life assessment in clinical trials: methodologic issues. *Controlled Clinical Trials* 1989;10:195-208s.

42. Sanders C, Egger M, Donnovan J, Tallon D, Frakel S. Reporting on quality of life in randomised controlled trials: bibliographic study. *British Medical Journal* 1998;317:1191-1194.

43. Williams A, Kind P. The present state of play about QALYs. In: Hopkins A, editor. Measures of Quality of Life. London: Royal College of Physicians; 1992.

44. Spiegelhalter DJ, Gore SM, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR. Quality of life measures in health care. III: Resource allocation [see comments]. *British Medical Journal* 1992;305:1205-9.

45. Williams A. The economics of coronary artery bypass grafting. *British Medical Journal* 1985;291:326-329.

46. Smith R. Rationing: the search for sunlight. *British Medical Journal* 1991;303:1561-1562.

47. Standing Committee on Postgraduate Medical Education. Medical Audit - the educational implications. London: SCOPME; 1989.

48. Higginson I. Quality of care and evaluating services. *International Review of Psychiatry* 1994;6:15-14.

49. Crombie IK, Davies HTO. Beyond health outcomes: the advantages of measuring process. *Journal of Evaluation in Clinical Practice* 1997;4:31-38.

50. Frater A, Costain T. Any better? Outcome measures in medical audit. *British Medical Journal* 1992;304:519-520.

51. Long AL, Dixon P. Monitoring outcomes in routine practice: defining appropriate measurement criteria. *Journal of Evaluation in Clinical Practice* 1996;2:71-78.

52. Kelly H, Russell EM, Stewart S, McEwan J. Needs assessment: taking stock. *Health Bulletin* 1996;54:115-18.

53. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for epidemiologists and clinicians. *Journal of the Royal College of General Practionners* 1985;35:185-188.

54. Delamonthe T. Using outcomes research in clinical practice. *British Medical Journal* 1994;308:1583-1584.

55. Geigle R, Jones SB. Outcomes measurement: a report from the front. *Inquiry* 1990;27:7-13.

56. Frater A. Health outcomes: a challenge to the status quo. *Quality in Health Care* 1992;1:87-88.

57. Cooper JE, Kendall RE, Gurland BJ, Sharpe L, Copeland JRM, Simon R. Psychiatric Diagnosis in New York and London. London: Oxford University Press; 1972.

58. Sartorius N, Jablensky A, Korten A, Ernberg G, Anker M, Cooper JE. Early manifestations and first contact incidence of schizophrenia in different cultures. *Psychological Medicine* 1986;16:909-928.

59. Reiger DA, Kaelber CT. The Epidemiologic Catchement Area (ECA) Programme: studying the prevalence and incidence of psychopathology. In: Tsuang MT, Tohen M, Zahner GEP, editors. New York: Wiley and Sons; 1995.

60. Shepherd M, Watt D, Fallon I, Smeeton N. The Natural History of Scizophrenia: A five year outcome and prediction in a representative sample of schizophrenics. Cambridge: Cambridge University Press; 1989.

61. Hamilton M. Development of a rating scale for primary depressive illness. *British Journal of Social & Clinical Psychology* 1967;6:278-296.

62. Hamilton M. Rating scales in depression. In: Kielholtz P, editor. Depressive Illness. Berlin: Huber; 1972.

63. Thornley B, Adams CE. Content and quality of 2000 controlled trials in schizophrenia over 50 years. *British Medical Journal* 1998;317:1181-1184.

64. Bowling A. Measuring Health: A review of quality of life measurement scales. Buckingham: Open University Press.; 1997.

65. Overall JE, Gorham DR. The brief psychiatric rating scale. *Psychological Reports* 1962;10:799-812.

66. Luborsky L. Clinicians judgements of mental health. *Archives of General Psychiatry* 1962;7:407-417.

67. Enndicot J, Spitzer RL, Fleis JL, Cohen J. The global assessment scale: a procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry* 1976;33:766-771.

68. American Psychiatric Association. Diagnostic and Statstical Manual - 4th Edition. Washington DC: American Psychiatric Association; 1994.

69. Spitzer RL, Gibbon M, Williams JBW, Endicott J. Global Assessment of Functioning (GAF) Scale. In: Sederer LI, Dickey B, editors. Outcomes Assessment in Clinical Practice. Baltimore: Williams and Wilkins; 1996.

70. Wiersma D. Measuring social disabilities in health. In: Thornicroft G, Tansella M, editors. Mental Health Outcome Measures. Berlin: Springer Verlag; 1996.

71. Weissman MM. The assessment of social adjustment. A review of techniques. *Archives of General Psychiatry* 1975;32:357-365.

72. Katching H. Methods for measuring social adjustment. In: Helgason T, editor. Methods in Evaluation of Psychiatric Treatment. Cambridge: Cambridge University Press; 1983. pp. 205-208.

73. Weissman MM, Bothwell S. The assessment of social adjustment by self report. *Archives of General Psychiatry* 1976;33:1111-1115.

74. Katz MM, Lyerly SB. Methods for measuring adjustment and social behaviour in the community. 1. Rationale, discriminative validity and scale development. *Psychological Reports* 1963;13:503-535.

75. Remmington M, Tyrer P. The Social Functioning Schedule: a brief semi-structured interview. *Social Psychiatry* 1979;14:151-157.

76. Katz S, Ford AB, Moskovitz RW. Studies of illness in the aged: a standardised measure of biological and social function. *Journal of the the American Medical Association* 1963;185:914-919.

77. Lehman AF. Measures of quality of life for people with severe mental disorders. In: Tansella M, Thornicroft G, editors. Mental Health Outcome Measures. Second ed. London: Gaskell; 2001.

78. Lehman AF. The well being of chronic mental patients: assessing their quality of life. *Archive of General Pscyhiatry* 1983;40:369-373.

79. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A basic science for clinical medicine. Boston, MA.: Little, Brown and Company; 1991.

80. WHO. Evaluation of Methods for the Treatment of Mental Disorders. Geneva: WHO; 1991.

81. Pocock SJ. Clinical trials: a practical approach. London: Wiley; 1983.

82. Chalmers I, Altman DG, editors. Systematic Reviews. London: BMJ; 1995.

83. The Cochrane Database of Systematic Reviews. In: The Cochrane Library, Issue 4. Oxford: Update Software; 2000.

84. Ware JE, Snoww KK, Kosinski M, Gandek B. SF-36 Health Survey: Manual and Interpretation Guide. Boston, MA.: The Health Institute, New England Medical Centre.; 1993.

85. Bergner M, Bobbitt RA, Kressell S. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. *International Journal of Health Services* 1976;6:393-415.

86. Ruggeri M. Measuring satisfaction with psychiatric services: towards a multidimensional measurement of outcome. In: Tansella M, Thornicroft G, editors. Mental Health Outcome Measures. Second ed. London: Gaskell; 2001.

87. Bech P, Malt UF, Denker SJ, Ahlfors UG, Elgen K, Lewander T, et al. Scales for the assessmetn of diagnosis and severity of mental disorders. *Acta Psychiatrica Scandinavica* 1993;87:Supplementum.

88. Thompson C. The Instruments of Psychiatric Research. Chichester: John Willey; 1989.

89. Sederer LI, Dickey B, editors. Outcomes Assessment in Clinical Practice. Baltimore: Williams and Wilkins; 1996.

90. Bowling A. Measuring Disease. Buckingham: Open University Press.; 1995.

91. Microsoft Corporation. Microsoft Access 97. ver. Version 3. Microsoft, 1998.

92. Microsoft Corporation. Microsoft Excel. ver. Version. Microsoft, 1997.

93. Buchan I. StatsDirect. ver. 1.6. 2000.

94. Kay SR. Positive and negative syndromes in schizophrenia. New York: Brunner-Mazel; 1991.

95. Nelson EC, Landgraf JM, Hays RD, Wasson JH, Kirk JW. The functional status of patients: how can it be measured in physician's offices? *Medical Care* 1990;28:1111-1126.

96. Heinrichs DW, Hanlon ET, Carpenter WT. The quality of life scale: an instrument for rating the schizophrenic defecit syndrome. *Schizophrenia Bulletin* 1984;10:388-98.

97. Lehman AF. The effect of psychiatric symptoms on quality of life assessments among the chronically menatally ill. *Evaluative Programme Planning* 1983;6:143-151.

98. Bigelow DA, Gareau MJ, Young DJ. A quality of life interview. *Psychosocial Rehabilitation Journal* 1982;14:94-98.

99. Baker R, Hall JN. REHAB: a new assessment instrument for chronic psychiatric patients. *Schizophrenia Bulletin* 1988;14:95-113.

100. Rosenberg M. Conceiving the Self. New York: Plenum Press; 1979.

101. Larsen D, Attkinson CC, Hargreaves WA. Assessment of clinet/patient satisfaction: development of a general scale. *Evaluation and Programme Planning* 1979;2:197-207.

102. Tait RC, Pollard CA, Margolis RB. The pain disability index: psychometric and validity data. *Archives of Pysical and Medical Rehabilitation* 1987;68:438-441.

103. Karnofsky DA, Ablemann WH, Craver LF. The use of the nitrogen mustrds in the paliative treatement of carcinoma. *Cancer* 1948;1:634-656.

104. Geddes JR, Freemantle N, Mason J, Eccles MP, Boynton J. Selective serotonin reuptake inhibitors (SSRIs) for depression (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

105. Lewis R, Bagnall A-M, Leitner ML. Ziprasidone for schizophrenia and severe mental illness (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

106. Lewis R, Bagnall A-M, Leitner ML. Sertindole for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

107. Kennedy E, Song F, Gilbody S. Risperidone versus typical antipsychotic medication for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

108. Gilbody SM, Bagnall AM, Duggan L, Tuunainen A. Risperidone versus other atypical antipsychotic medication for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

109. Srisurapanont M, Disayavanish C, Taimkaew K. Quetiapine for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

110. Duggan L, Fenton M, Dardennes RM, El-Dosoky A, Indran S. Olanzapine for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

111. Tuunainen A, Gilbody SM. Newer atypical antipsychotic medication versus clozapine for schizophrenia (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

112. Bagnall A-M, Fenton M, Lewis R, Leitner ML, Kleijnen J. Molindone for schizophrenia and severe mental illness (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

113. Bagnall AM, Lewis R, Gilbody SM, Kleijnen J. New atypical drugs for schizophrenia. *Health Technology Assessment* 2001;In Press.

114. Marshall M, Gray A, Lockwood A, Green R. Case management for people with severe mental disorders (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

115. Marshall M, Lockwood A. Assertive community treatment for people with severe mental disorders (Cochrane Review). In: The Cochrane Library, Issue 2. Oxford: Update Software; 2001.

116. The Cochrane Controlled Trials Register. In: The Cochrane Library, Issue 4. Oxford: Update Software; 2000.

117. Collins EJ, Hogan TP, Himansu D. Measurement of therapeutic response in schizophrenia. *Schizophrenia Research* 1991;5:249-253.

118. Tollefson GD, Beasley CM, Tran PV, Street JS, Krueger JA, Tamura RN, et al. Olanzapine versus haloperidol in the treatment of schizophrenia

and schizoaffective and schizophreniform disorders: results of an international collaborative trial. *American Journal of Psychiatry* 1997;154:457-465.

119. FDA. The Food and Drug Administration Modernisation Act. Rockville: FDA; 1997.

120. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient based outcome measures for use in clinical trials. *Health Technology Assessment* 1998;2.

121. Hotopf M, Lewis G, Normand C. Putting trials on trial--the costs and consequences of small trials in depression: a systematic review of methodology. *Journal of Epidemiology & Community Health* 1997;51:354-8.

122. Wells KB. Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry* 1999;156:5-10.

123. Gilbody SM, Whitty PA. Improving the delivery and organisation of mental health services: beyond the conventional RCT. *British Journal of Psychiatry* 2002;180:13-18.

124. Hotopf M, Churchill R, Lewis G. Pragmatic randomised trials in psychiatry. *British Journal of Psychiatry* 1999;175:217-223.

125. Lilford R, Royston G. Decision analysis in the selection, design and application of clinical and health services research. *Journal of Health Services Research & Policy* 1998;3:159-166.

126. Ellwood PM. Shattuck lecture - outcomes management. A technology of patient experience. *New England Journal of Medicine* 1988;318:1549-56.

127. Anonymous. Databases for healthcare outcomes. *Lancet* 1989;396:195-196.

128. Wennberg J. What is outcomes research? In: Ginzberg E, editor. Health Services Research: Key to health policy. Cambridge, Massachusetts.: Harvard University Press.; 1991. pp. 33-46.

129. Wennberg JE, Barry MJ, Fowler FJ, Mulley A. Outcomes research, PORTs, and health care reform. *Annals of the New York Academy of Sciences* 1993;703:52-62.

130. Anderson C. Measuring what works in health care. *Science* 1994;263:1080-1082.

131. General Accounting Office. Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC.: GAO; 1992.

132. Anonymous. Cross design synthesis: a new strategy for studying medical outcomes. *Lancet* 1992;340:944-946.

133. Smith DM. Database research: Is happiness a homongous database? *Annals of Internal Medicine* 1997;127:725-756.

134. Aday LA, Begley CE, Lairson DR, Slater CH. Evaluating Healthcare System: Effectiveness, efficiency and equity. Second ed. Chicago: AHSR; 1998.

135. Zarin DA, Pincus HA, West JC, McIntyre JSSO. Practice-based research in psychiatry. *American Journal of Psychiatry* 1997;154:1199-208.

136. Zarin DA, West JC, Pincus HA, McIntyre JS. The American Psychiatric Association Practice Research Network (PRN). In: Sederer LI, Dickey B, editors. Outcomes Assessment in Clinical Practice. Baltimore: Williams and Wilkins; 1996.

137. Barkham M, Evans C, Marginson F, McGrath G, Mellor-Clark J, Milne D, et al. The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Mental Health* 1998;7:35-47.

138. Mellor-Clarke J, Barkham M, Connell J, Evans C. 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* 1999;3:357-374.

139. Marginson FR, McGrath G, Barkham M, Mellor Clark J, Audin K, Connell J, et al. Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *British Journal of Psychiatry* 2000;177:123-130.

140. Guthrie E. Psychotherapy for patients with complex disorders and chronic symptoms: The need for a new research paradigm. *British Journal of Psychiatry* 2000;177:131-137.

141. Sheldon TA. Please bypass the PORT. *British Medical Journal* 1994;309:142-143.

142. Iezzoni LI. Assessing quality using administrative data. *Annals of Internal Medicine* 1997;127:666-674.

143. Donoghue J, Tylee A, Wildgust H. Cross sectional database analysis of antidepressant prescribing in general practice in the United Kingdom, 1993-5. *British Medical Journal* 1996;313:861-2.

144. Cook TD, Campell DT. Quasi-experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin; 1979.

145. Thornicroft G, Strathdeee G, Phelan M, Holloway F, Wykes T, Dunn G, et al. Rationale and design: PRiSM psychosis study I. *British Journal of Psychiatry* 1998;173:363-370.

146. Rosenheck R, Leda C, Frisman L, Gallup P. Homeless mentally ill veterans: race, service use, and treatment outcomes. *American Journal of Orthopsychiatry* 1997;67:632-638.

147. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *Journal of the American Medical Association* 1989;262:914-9.

148. Wells KB, Sturm R, Sherbourne CD, Meredith LS. Caring for Depression. Massachusetts: Harvard University Press; 1996.

149. Rosenheck RA, Druss B, Stolar M, Leslie D, Sledge W. Effect of declining mental health service use on employees of a large corporation. *Health Affairs* 1999;18:193-203.

150. Lam JA, Rosenheck R. Street outreach for homeless persons with serious mental illness: is it effective? *Medical Care* 1999;37:894-907.

151. Leslie DL, Rosenheck RA. Comparing quality of mental health care for public sector and privately insured populations. *Psychiatric Services* 2000;51:650-655.

152. Melfi C, Chawla A, Croghan T, Hanna M, Kennedy S, Sredl K. The effects of adherence to antidepressant treatment guidelines on relapse and recurrence of depression. *Archives of General Psychiatry* 1998;55:1128-32.

153. Hong WW, Rak IW, Ciuryla VT, Wilson AM, Kylstra JW, Meltzer HY, et al. Medical-claims databases in the design of a health-outcomes comparison of quetiapine ('Seroquel') and usual-care antipsychotic medication. *Schizophrenia Research* 1998;32:51-8.

154. Croghan T, Melfi C, Dobrez D, Kniesner T. Effect of mental health specialty care on antidepressant length of therapy. *Medical Care* 1999;37:AS20-3.

155. Hylan T, Crown W, Meneades L, Heiligenstein J, Melfi C, Croghan T, et al. SSRI antidepressant drug use patterns in the naturalistic setting: a multivariate analysis. *Medical Care* 1999;37:AS36-44.

156. Rosenheck R, Stolar M, Fontana A. Outcomes monitoring and the testing of new psychiatric treatments: work therapy in the treatment of chronic post-traumatic stress disorder. *Health Services Research* 2000;35:133-152.

157. Stewert AL, Ware JE, editors. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Durham, N. C.: Duke University Press; 1992.

158. National Committee on Quality Assurance. National Committee on Quality Assurance report card pilot project. Washington DC: NCQA; 1995.

159. Rosenheck RA. Department of Veterans Affairs national mental health programme performance monitoring system: fiscal 1995 report. West Haven, CT.: Northeast Programme Evaluation Centre; 1996.

160. Streator SE, Moss JT. Non approved usage identification of off label anti-depressant use and cost in a network model HMO. *Drug Benefit Trends* 1997;9:42-47.

161. Jencks SF. The recognition of mental distress and diagnosis of mental disorder in primary care. *Journal of the American Medical Association* 1985;253:1903-1907.

162. Rost K, Smith GR, Matthews DB, Guse B. The deliberate misdiagnosis of major depression in primary care. *Archives of Family Medicine* 1994;3:333-342.

163. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997;127:757-763.

164. Randolph FL, Blasinsky M, Leginski W, Parker LB, Goldman HH. Creating integrated service systems for homeless persons with mental illness: the ACCESS programme. Access to Community Care and Effective Services and Supports. *Psychiatric Services* 1997;48:369-374.

165. Agency for Health Care Policy Research. Depression in primary care. Washington DC: US Department of Health and Human Services; 1993.

166. Office of Health Technology Assessment - US Congress. Identifying Health Technologies that Work: Searching for Evidence OTA-H-608. Washington DC: US Government Printing Office; 1994.

167. Anonymous. From research to practice. *Lancet* 1994;344:417-418.

168. Naylor CD. Grey zones of clinical practice: some limits to evidence based medicine. *Lancet* 1995;345:840-842.

169. Blumberg MS. Potentials and limitations of database research illustrated by the QMMP AMI Medicare mortality study. *Statistics in Medicine* 1991;10:637-646.

170. Green SB, Byar DP. Using observational data from registries to compare treatments. *Statistics in Medicine* 1984;3:351-370.

171. Rosenheck RA, Fontanna A, Stolar M. Assessing quality of care: administrative indicators and clinical outcomes in post traumatic stress disorder. *Medical Care* 1999;37:180-188.

172. Black NA. High Quality Clinical Databases: breaking down barriers. *Lancet* 1999;353:1205-1206.

173. Rowan KM. Intensive Care National Audit and Research Centre: past present and future. *Care of the Critically Ill* 1994;10:148-149.

174. Clifford P. M is for Outcome: the CORE outcomes initiative. *Journal of Mental Health* 1998;317:1167-1168.

175. Al-Shahi R, Warlow C. Using patient identifiable data for observational research and audit: overprotection could damage the public interest. *British Medical Journal* 2000;321:1031-2.

176. Medical Research Council. Personal information in medical research. London: Medical Research Council; 2000.

177. Kmietowicz Z. Registries will have to apply for right to collect patients' data without consent. *British Medical Journal* 2001;322:1199.

178. Anderson R. Undermining data privacy in health information: new powers to control patient information contribute nothing to health. *British Medical Journal* 2001;322:442-3.

179. Department of Health. The Health of the Nation: a strategy for England. London: HMSO; 1991.

180. Slade M, Thornicroft G, Glover GSO. The feasibility of routine outcome measures in mental health. *Social Psychiatry & Psychiatric Epidemiology* 1999;34:243-249.

181. House of Commons. The National Health Service and Community Care Act. London: HMSO; 1990.

182. Department of Health. Medical and Dental Workforce - detailed statistics 2000. 2000. Available from: URL: http://www.doh.gov.uk/public/stats3.htm#workforce

183. Moser K, Kalton G. Survey methods of social investigation. London: Gower and Aldershot; 1971.

184. Fowler FJ. Survey Research Methods. Second Edition ed. California: SAGE; 1993.

185. Financial Times Healthcare. The Medical Directory on CD-ROM. London: FT Pharmaceuticals; 2000.

186. STATA corporation. STATA. ver. version 6. STATA corporation, 1999.

187. Dilman D. The design and administration of mail surveys. *Annual Review of Sociology* 1991;17:225-249.

188. McColl E, Jacoby A, Tomas L, Souter J, Bamford C, Steen N, et al. The conduct and design of questionnaire surveys in healthcare research. In: Stevens A, Abrams K, Brazier J, Fitzpatrick R, Lilford R, editors. The Advanced Handbook of Methods in Evidence Based Healthcare. London: Sage; 2001.

189. Folstein MF, Folstein SE, McHugh PR. 'Mini Mental State': a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975;12:189-98.

190. Wing J. Measuring mental health outcomes: a perspective from the Royal College of Psychiatrists. In: Delamonthe T, editor. Outcomes into clinical practice. London: BMJ Publishing; 1994. pp. 147-152.

191. Mayfield D, Millard G, Hall P. The CAGE questionnaire. *American Journal of Psychiatry* 1974;131:1121-3.

192. Beck AT, Ward CH. An inventory for measuring depression. *Archives of General Psychiatry* 1961;4:561-571.

193. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression scale. *Acta Psychiatrica Scandinavica* 1983;67:361-70.

194. Glover G, Knight S, Melzer D, Pearce L. The development of a new minimum data set for specialist mental health care. *Health Trends* 1997;29:48-51.

195. Secretary of State for Health. National Service Framework - Mental Health. London: HMSO; 1999.

196. Davies HTO, Lampel J. Trust in performance indicators. *Quality in Health Care* 1998;7:159-162.

197. Davies HTO, Crombie IK. Interpreting Health Outcomes. *Journal of Evaluation in Clinical Practice* 1997;3:187-199.

198. Smith P, editor. Measuring Outcome in the Public Sector. London.: Taylor and Francis Limited.; 1996.

199. Smith P. A framework for analysing the measurement of outcome. In: Smith P, editor. Measuring Outcome in the Public Sector. London.: Taylor and Francis Limited.; 1996.

200. Aitchison KJ, Kerwin RW. Cost effectiveness of clozapine: A UK clinic based study. *British Journal of Psychiatry* 1997;171:125-130.

201. Adams CE. Establishing cost-effectiveness of antipsychotic drugs. *British Journal of Psychiatry* 1997;171:486.

202. Streiner D, Norman G. Health Measurement Scales: A practical guide to thier development and use. Oxford, UK.: Oxford University Press.; 1995.

203. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Annals of Internal Medicine* 1993;118:622-9.

204. Stein GS. Usefulness of the Health of the Nation Outcome Scales. *British Journal of Psychiatry* 1999;174:375-377.

205. Bebbington P, Brugha T, Hill T, Marsden L, Window S. Validation of the Health of the Nation Outcome Scales. *British Journal of Psychiatry* 1999;174:389-394.

206. Marks ISO. Overcoming obstacles to routine outcome measurement. The nuts and bolts of implementing clinical audit. *British Journal of Psychiatry* 1998; 173:281-286.

207. Department of Health. Modernising Mental Health Services: safe, sound and supportive. London: HMSO; 1998.

208. Curtis R, Beevor A. Health of the Nation Outcome Scales (HoNOS). In: Wing J, editor. Measurement for Mental Health: contributions from the College Research Unit. London: Royal College of Psychiatrists; 1995. pp. 33-46.

209. Sharma VK, Wilkinson G, Fear S. Health of the Nation Outcome Scales: a case study in general psychiatry. *British Journal of Psychiatry* 1999;174:395-398.

210. Sprangers MA, Aaranson NK. 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* 1992;45:743-760.

211. Young JB, Chamberlain, M. A. The contribution of the Stanford Health Assessmetn questionnaire in rheumatology clinics. *Clinical Rehabilitation* 1987;1:97-100.

212. Kazis LE, Callahan LF, Meenan RF, Pincus TSO. Health status reports in the care of patients with rheumatoid arthritis. *Journal of Clinical Epidemiology* 1990;43:1243-53.

213. Goldberg D, Eastwood MR, Kedwood HB. A standardised pscyhiatric interview for use in community surveys. *British Journal of Preventative and Social Medicine* 1970;24:18-23.

214. Goldberg D, Huxley P. Mental Illness in the Community. London: Tavistock; 1980.

215. Feldman E, Mayou R, Hawton K, Ardern M, Smith EB. Psychiatric disorders in medical in-patients. *Quarterly Journal of Medicine* 1987;63:405-412.

216. van Hemert AM, Hengeveld MW, Bolk JH, Rooijmans HG, Vandenbroucke JPSO. Psychiatric disorders in relation to medical illness among patients of a general medical out-patient clinic. *Psychological Medicine* 1993;23:167-173.

217. Meakin CJ. Screening for depression in the medically ill. *British Journal of Psychiatry* 1992;160:212-216.

218. Goldberg D. The Detection of Psychiatric Illness by Questionnaire. Oxford: Oxford University Press; 1972.

219. Goldberg D. The use of the general health questionnaire in clinical work. *British Medical Journal* 1986;293:1188-1189.

220. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. *Medical Care* 1992;30:Ms23-41.

221. Orley J, Saxena S, Herrman H. Quality of life and mental illness. *British Journal of Psychiatry* 1998;172:291-293.

222. Anthony W, Rogers S. Relationship between psychiatric symptomotology, work skills, and future vocational performance. *Psychiatric Services* 1995;46:353-358.

223. Becker M, Diamond R, Sainfort F. A new patient focussed index for measuring quality of life in persons with severe and persistent mental illness. *Quality of Life Research* 1993;2:239-251.

224. Revicki DA, Murray M. Assessing health related quality of life outcomes of drug treatements for psychiatric disorders. *CNS Drugs* 1994;1:465-476.

225. Sainfort F, Becker M, Diamond R. Judgements of quality of life of individuals with severe mental disorders: Patient self report versus provider perspectives. *American Journal of Psychiatry* 1996;153:497-502.

226. Thornicroft G, Brewin C, Wing J. Measuring Mental Health Needs. London: Royal College of Psychiatrists; 1992.

227. Stevens A, Gillam S. Needs assessment: from theory to practice. *British Medical Journal* 1998;316:1448-1452.

228. Wright J, Williams R, Wilkinson JR. Development and importance of health needs assessment. *British Medical Journal* 1998;316:1310-1313.

229. Phelan M, Slade M, Thornicroft G, Dunn D, Holloway F, Wykes T, et al. 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* 1995;167:589-95.

230. Brewin CR, Wing JK. The MRC Needs for Care Assessment: progress and controversies [editorial]. *Psychological Medicine* 1993;23:837-41.

231. Meenan RF. The AIMS approach to health status measurement: Conceptual background and measurement properties. *Journal of Rheumatology* 1982;9:785-788.

232. Pincus T, Summey JA, Soraci S, Wallson K, Hummon NP. Assessment of patietn satisfaction in activities of daily living using the modified Sanford Health Assessment Questionnaire. *Arthritis and Rheumatism* 1983;26:1346-1353.

233. Deyo RA, Patrick DL. Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care* 1989;27:S254-68.

234. Deyo RA, Carter WB. Strategies for improving and expanding the application of health status measures in clinical settings. A researcher-developer viewpoint. *Medical Care* 1992;30:Ms176-86.

235. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. *Quality in Health Care* 1992;1:89-93.

236. Dunn G. Statistical methods for measuring outcomes. In: Thornicroft G, Tansella M, editors. Mental Health Outcome Measures. Berlin: Springer Verlag; 1996. pp. 3-15.

237. NHS Centre for Reviews and Dissemination. Undertaking Systematic Reviews of Research on Effectiveness: CRD report 4 (second edition). York: University of York; 2001.

238. Mulrow CD, Oxman AD. Cochrane Collaboration Handbook [updated June 1999]. In: Collaboration TC, editor. The Cochrane Library [database on disk and CDROM]. Oxford: Update Software; 1999.

239. Gilbody SM, House AO, Sheldon TA. Outcome and needs assessment for schizophrenia and related disorders (Cochrane Review). In: The Cochrane Library - Issue 1. Oxford: Update Software; 2001.

240. Gilbody SM, House AO, Sheldon TA. 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; 2001.

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

242. National Institute of Mental Health. Toward a model plan for a comprehensive community based mental health system. Washington, DC.: National INstitute of Mental Health; 1987.

243. Thornicroft G. Needs Assessment. In: Knunsden HC, Thornicroft G, editors. Mental Health Service Evaluation. Cambridge: Cambridge University Press; 1996.

244. van den Bos GAM, Triemstra AHM. Quality of life and as an instrument for need assessment and outcome assessment of health care in chronic patients. *Quality in Health Care* 1999;8:247-252.

245. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG, Donner A. Methods in health service research. Evaluation of health interventions at area and organisation level. *British Medical Journal* 1999;319:376-9. Available from: URL: http://www.biomednet.com/db/medline/99365199

246. Brettle AJ, Long AF, Grant MJ, Greenhalgh J. Searching for information on outcomes: do you need to be comprehensive? *Quality in Health Care* 1998;7:163-167.

247. Jadad AR, Moore RA, Carroll D. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996;17:1-12.

248. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 1995;273:408 - 412.

249. Ukoumunne OC, Gulliford MC, Chinn S, Sterne AC, Burney PGJ. Methods for evaluating area-wide and organisation based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999;3.

250. Meinert CL. Clinical Trials: Design, conduct and analysis. Oxford: Oxford University Press; 1986.

251. DerSimonian R, Laird N. Meta-analysis in Clinical Trials. *Controlled Clinical Trials* 1986;7:177-188.

252. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10:154-173.

253. Egger M, Davey-Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997;315:629-634.

254. Pignone M, Gaynes BG, Lohr K, Orleans CT, Mulrow CD. Systematic review is incomplete. *British Medical Journal* 2001;323:167.

255. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999;354:1896-8.

256. Johnstone A, Goldberg D. Psychiatric screening in General Practice. *Lancet* 1976;1:605-612.

257. Weatherall M. A randomized controlled trial of the Geriatric Depression Scale in an inpatient ward for older adults. *Clinical Rehabilitation* 2000;14:186-91.

258. Gold I, Baraff LJ. Psychiatric screening in the emergency department: its effect on physician behaviour. *Annals of Emergency Medicine* 1989;18:875-880.

259. Street RL, Jr., Gold WR, McDowell T. Using health status surveys in medical consultations. *Medical Care* 1994;32:732-44.

260. Roethlisberger FJ, Dickinson WJ. Management and the Worker. Cambridge, MA: Harvard University Press; 1939.

261. Goldsmith G, Brodwick M. Assessing the functional status of older patients with chronic illness. *Family Medicine* 1989;21:38-41.

262. Rubenstein LV, Calkins DR, Young RT. Improving patient functioning: a randomised trial of functional disability screening. *Annals of Internal Medicine* 1989;111:836-842.

263. Mazonson PD, Mathias SD, Fifer SK, Beusching DP, Patrick DL. The mental health patient profile: does it change primary care physicians practice patterns? *Journal of the American Board of Family Practice* 1994;9:336-345.

264. Wasson J, Hays R, Rubenstein L, Nelson E, Leaning J, Johnson D, et al. The short-term effect of patient health status assessment in a health maintenance organization. *Quality of Life Research* 1992;1:99-106.

265. Calkins DR, Rubenstein LV, Cleary PD. Functional disability screening of ambulatory patients: a randomised controlled trial in a hopital based group practice. *Journal of General Internal Medicine* 1994;9:590-592.

266. Rubenstein LV, McCoy JM, Cope DW, Barrett PA, Hirsch SH, Messer KS. Improving patient quality of life with feedback to physicians about functional status. *Journal of General Internal Medicine* 1995;10:707-614.

267. Reilfer DR, Kessler HS, Bernhard EJ, Leon AC, Martin G. Imapct of screening for mental health concerns on health service ustilisation and functional status in primary care patients. *Archives of Internal Medicine* 1996;156:2593-2599.

268. Whooley MA, Stone B, Soghikian K. Randomized trial of case- finding for depression in elderly primary care patients. *Journal of General Internal Medicine* 2000;15:293-300.

269. Divine GW, Brown JT, Frazer LM. The unit of analysis error in studies about physicians' patient care behavior. *Journal of General Internal Medicine* 1992;7:623-629.

270. Dowrick C, Buchan I. Twelve month outcome of depression in general practice: does detection or disclosure make a difference? *British Medical Journal* 1995;311:1274-1276.

271. Williams JWJ, Mulrow CD, Kroenke K. Case-finding for depression in primary care: a randomized trial. *American Journal of Medicine* 1999;106:36-43.

272. Hoeper EW, Nycz GR, Kessler JD, Pierce WE. The usefulness of screening for mental illness. *Lancet* 1984;1:33-35.

273. Magruder Habib K, Zung WW, Feussner JR. Improving physicians' recognition and treatment of depression in general medical care. Results from a randomized clinical trial. *Medical Care* 1990;28:239-250.

274. Zung WWK. A self rating depression rating scale. *Archives of General Psychiatry* 1965;12:63-70.

275. Wagner AK, Ehrenberg BL, Tran TA, Bungay KM, Cynn DJ, Rodgers WH. Patient based health status measurement in clinical practice: a study of its impact in epileps patients. *Quality of Life Research* 1997;6:329-341.

276. Jette AM, Davies AR, Calkins DR. The Functional Status Questionnaire: its reliability and validity when used in primary care. *Journal of General Internal Medicine* 1986;1:143-149.

277. Wasson J, Keller A, Rubenstein L, Hays R, Nelson E, Johnson D. Benefits and obstacles of health status assessment in ambulatory settings. The clinician's point of view. The Dartmouth Primary Care COOP Project. *Medical Care* 1992;30:Ms42-9.

278. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care* 1981;19:787-805.

279. Mathias SD, Fifer SK, Mazonson PD, Lubeck DP, Beusching DP, Patrick, D. P. Necessary but not sufficient: the effect of screening and feedback on outcomes of primary care patients with untreated anxiety. *Journal of Internal Medicine* 1994;9:606-615.

280. Derogatis LR. Symptom Checklist-90-R (SCL-90-R) administration, scoring and procedures manual. 3rd Edition ed. Minneapolis: National Computer Systems; 1994.

281. Fifer SK, Mathias SD, Patrick DL, Mazonson PD, Lubeck DP, Beusching DP. Untreated anxiety among adult primary care patients in a health maintainance organisation. *Archives of General Psychiatry* 1994;51:740-750.

282. Broadhead WE, Leon AC, Weissman MM. Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Archives of Family Medicine* 1995;4:211-219.

283. German PS, Shapiro S, Skinner EA. Detection and management of mental health problems of older patients by primary care providers. *Journal of the American Medical Association* 1987;257:489-496.

284. Linn LS, Yager J. The effect of screening, sensitisation and feedback on notation of depression. *Journal of Medical Education* 1980;20:942-953.

285. Shapiro S, German PS, Skinner EA, VonKorf M, Turner RW, Klein LE, et al. An experiment to change the detection and management of mental morbidity in primary care. *Medical Care* 1987;25:327-339.

286. Callahan CM, Hendrie HC, Dittus RS, Brater DC, Hui SL, Tierney WMl. Improving treatment of late life depression in primary care: a randomized clinical trial. *Journal of the American Geriatrics Society* 1994;42:839-46.

287. Moore JT, Silimperi DR, Bobula JA. Recognition of depression by family medicine residents: the impact of screening. *Journal of Family Practice* 1978;7:509-513.

288. Zung WW, Magill M, Moore JT, George DT. Recognition and treatment of depression in a family medicine practice. *Journal of Clinical Psychiatry* 1983;44:3-6.

289. Dorwick C. Does testing for depression influence diagnosis or management by general practioners? *Family Practice* 1995;12:461-465.

290. Lewis G, Sharp D, Bartholomew J, Pelosi AJ. Computerized assessment of common mental disorders in primary care: effect on clinical outcome. *Family Practice* 1996;13:120-6.

291. Linn LS, Yager J. Screening for depression in relationship to subsequent patient and physician behaviour. *Medical Care* 1980;20:1233-1245.

292. Depression Guideline Panel. Depression in primary care: Quick reference guide for clincians. Clincinal practice Guideline Number 5. Rockville, MD: AHCPR; 1993.

293. Wright A. Should general practitioners be testing for depression? *British Journal of General Practice* 1994;44:132-135.

294. Thompson S. Why sources of heterogeneity in meta-analysis should be invetsigated. In: Chalmers I, Altman DG, editors. Systematic Reviews. London: BMJ; 1995.

295. Gilbody SM, Song F. Publication bias and the integrity of psychiatry research. *Psychological Medicine* 2000;30:253-258.

296. Petticrew M, Gilbody SM, Sheldon TA. Relation between hostility and coronary heart diseaese. *British Medical Journal* 1999;319:917-918.

297. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice. A systematic review of rigorous evaluations. *Lancet* 1993;342:1317-22.

298. Rice N, Leyland A. Multi-level models: applications to health data. *Journal of Health Services Research & Policy* 1996;3:154-64.

299. Mantel N, Haenszel W. Statistical aspects of the analysis of data from reterospective studies of disease. *Journal of the National Cancer Institute* 1959;22:719-730.

300. Goldberg DP, Williams P. The user's guide to the General Health Questionnaire. Windsor: NFER-Nelson; 1988.

301. Thompson C, Kinmonth J, Stevens L, Peveler RC, Stevens A, Ostler KJ, et al. 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* 2000;355:50-57.

302. Goldberg D, Bridges K. Somatic presentation of psychiatric illness in primary care setting. *Journal of Psychosomatic Research* 1988;32:137-144.

303. Cochrane AL, Holland WW. Validation of screening procedures. *British Medical Bulletin* 1971;27:3-8.

304. Mant D, Fowler G. Mass screening: theory and ethics. *British Medical Journal* 1990;300:916-18.

305. McHorney CA, Tarlov AR. The use of health status measures for individual patient level applications: problems and prospects. *Quality of Life Research* 1994;3:43-44.

306. Feinstein A. Clinical Judgement. Baltimore, MD: Williams and Wilkins; 1967.

307. Mannion R, Goddard M. Impact of published clinical outcomes data: case study of NHS hospital trusts. *British Medical Journal* 2001;323:260-264.

308. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technology Assessment* 1999;3.

309. Tugwell P, Boers M. Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. *Journal of Rheumatology* 1993;20:555-556.

310. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis and Rheumatism* 1993;36:729-740.

311. Power M. The audit society: rituals of verification. Oxford: Oxford University Press, 1997.