THE UNIVERSITY *of York*

CHE
Centre For Health Economics

**Investigating Patient Outcome
Measures in Mental Health**

CHE Research Paper 48

Investigating Patient Outcome Measures in Mental Health

Rowena Jacobs

Centre for Health Economics, University of York, York YO10 5DD, UK

May 2009

Background

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The new CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Acknowledgements

Disclaimer

Papers published in the CHE Research Paper (RP) series are intended as a contribution to current research. Work and ideas reported in RPs may not always represent the final position and as such may sometimes need to be treated as work in progress. The material and views expressed in RPs are solely those of the authors and should not be interpreted as representing the collective views of CHE research staff or their research funders.

Further copies

Table of contents

List of tables

List of figures

Executive summary

1. This report examines the feasibility of incorporating patient outcomes in mental health into a productivity measure. It examines which outcome measures are most commonly used in mental health, the practical issues about collecting these outcome measures, whether they can be converted into a generic measure, whether there is a time series of data available, and whether the data exists to examine changes in the mix of treatments over time. The criteria that were assumed to be important for an outcome measure to be included in a productivity index, were that it should have wide coverage, should be routinely collected, could readily be linked to activity data, could potentially be converted to a generic outcome measure, and would be available as a time-series. The report focuses predominantly on mental health outcomes within the working age population. Literature searches on outcome measurement in mental health covered numerous databases and retrieved over 1500 records. Around 170 full papers were obtained.

2. Two measures emerged as the most likely contenders for inclusion in a productivity index in mental health, namely the Health of the Nation Outcome Scales (HoNOS) and Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). HoNOS is used for patients with severe and enduring mental illness, while CORE-OM covers patients most often treated in the community setting or involved in psychological therapies. HoNOS is a clinician-rated measure, while CORE-OM is a self-report measure.

3. Most rating scales in psychiatry are completed by clinicians. The patient voice has tended to be ignored in the development of instruments. Potential problems with clinician ratings include the cost, training, potential for gaming, reliability of the assessment and inter-rater reliability. Potential concerns with self-reports from patients include concerns around social desirability, idiosyncratic completion, and use with patients who are acutely unwell. There is usually not much agreement in ratings between clinicians and patients on different scales or even when using the same instrument.

4. International approaches to routine outcome measurement have primarily been driven to inform programs and systems, rather than helping to inform individual patient-level treatment outcomes. Purchaser-driven pressures in the US have lead to the mandating of routine outcome assessment in a number of different settings, for example the Veterans Administration and the Ohio Department of Mental Health. Outcome assessment is linked to performance measures on recording compliance and agency certification. Australia has the most coherently developed approach to treatment-level routine outcome assessment. They have mandated the use of HoNOS as a standard outcome measure for all patients, as well as a self-report instrument. Perceptions of the value of the outcome measurement system seem to be mixed. Equal numbers seem to find HoNOS to be of value and not of value, although more positive observations have been made about the user-rated outcome measures.

5. In the UK, an Outcomes Reference Group was established by the Department of Health in 2002 to advise on best practice guidance which culminated in the Fonagy report. One of the key conclusions was that it was essential for local Trusts to develop systems to use local outcomes data effectively to inform service delivery, before the data could be used for higher level purposes (such as productivity measurement). The HoNOS was recommended for use for all patients on enhanced CPA care, with at least one measurement taken per year. If other measures are adopted, they should be able to generate an equivalent HoNOS score. These recommendations were subsequently issued as Technical Guidance for the minimum requirements to complete the mandatory Mental Health Minimum Data Set (MHMDS) for England since April 2003. The Fonagy recommendations therefore become a mandate by the DoH.

6. Alongside the Outcomes Reference Group, pilot work in four mental health Trusts tested the practicalities of using different outcome measures, including HoNOS. Key conclusions from the pilots were that response rates for the clinician-rated measures were much higher than for user-rated measures, outcome measurement relies critically on the availability of additional data such as diagnosis in order to inform treatments, and most notably Trusts lack effective systems to promote feedback to clinicians so that the benefits of outcome measurement can be appreciated.

7. While there is consensus that outcomes should be routinely measured, there is in fact limited evidence that routine outcome measurement can deliver improvements in local service delivery and patient-level care. The evidence from RCTs suggests that one-off or infrequent outcome

measurements have very little effect on improving quality of life or other subjective secondary outcome measures. But outcome measurement that is done longitudinally and more regularly (with more than one or two administrations) can significantly improve patient's quality of life or reduce psychiatric admissions.

8.   HoNOS was developed by the Royal College of Psychiatrists' Research Unit (CRU) between 1993 and 1995 and is recommended by the National Service Framework (NSF). HoNOS instruments are available for Children and Adolescent services, older people, forensic services, learning disabilities and acquired brain injury. There are 12 items on HoNOS each scored from 0 (no problem) to 4 (severe problem) yielding a total score in the range of 0 (best) to 48 (worst). The HoNOS takes on average between 5 and 15 minutes to complete, depending on the experience of the rater and the complexity of the patient's problems. The rating period covers the previous two weeks. CRU provides training for raters and training for trainers, however the cost is not trivial. HoNOS has undergone a number of independent studies to examine its psychometric properties and has been found to have good validity, and adequate reliability, sensitivity to change and acceptability. Comparing completion rates for HoNOS in the MHMDS in 2004/05 and 2005/06, the coverage of HoNOS dropped from 44 to 37 providers (out of 84), although the overall completion rates for those who submitted HoNOS returns remained around 9.5%. Due to the drop in the number of providers completing HoNOS, the overall coverage dropped from around 5% to 4%. Time-series activity and outcome data exists from 2003 onwards in the MHMDS for use in a productivity index, though data quality suggests only the latest 2 years should be used.

9. CORE-OM was developed between 1995 and 1998 through a multi-disciplinary team of practitioners representing the major psychological therapy professions. CORE Information Management Systems (IMS), a not-for-profit organisation, provides support to the CORE Network of users who can voluntarily anonymously donate aggregate to the National Research Database. CORE-OM is a free tool and comprises 34 items each with 5 levels from 0 (not at all) to 4 (all the time), giving an average score of between 0 and 4. It takes approximately 5 minutes for the patient to complete. Normative tables exist with cut-offs (severity bands) for clinical and non-clinical samples of the UK population. There are also two software versions, namely CORE-PC and CORE-NET. Studies to examine the psychometric properties of the CORE-OM have found it to have reasonable reliability, validity, sensitivity to change and acceptability. CORE-OM is believed to be the most widely used outcome measure in psychological therapy and counselling services, used in around 250 services in the NHS. The database which CORE IMS holds, covers around 100,000 patients per annum. Around 30 Mental Health Trusts and around 75 PCTs use the software version in psychotherapy and counselling services. Time-series activity and outcome data exists from 1999 onwards in the CORE IMS Database, but access would need to be explored for use in a productivity index. Data quality in early years would need to be tested.

10. Generic outcome measures such as the QALY are often not applicable for mental health patients and are not typically used in routine care. There would be considerable challenges in both trying to value or map either the HoNOS or CORE-OM to a generic measure like a QALY for use in a productivity index. The complexity of the valuation problem is enormous and while mapping could in principle be done, it would require the collection of both sets of data, the disease-specific and generic measure on the same set of patients. Mapping depends on the degree of overlap in content between the instruments being mapped and this will be questionable in these circumstances.

11. In addition to the literature search, interviews were held with 28 policymakers, academics, and NHS staff involved in outcome measurement, including managers, clinicians and commissioners. Ethics approval was obtained. The themes emerging from the interviews included the main barriers to outcome measurement, the incentives that drove the process, the choice of the instrument, and the policy context. Probably the most crucial barrier to the introduction of outcome measures is that clinicians are unable to see the clinical benefits, partly because they have not been given a clear rationale for their use, partly because they are simply told to complete scores, but primarily because they never receive any feedback on them. Many see it as a paper-filling exercise for managers. Unreliable and out-of-date IT systems and a lack of IT skills was another major barrier. For clinicians to get valuable feedback, often required them to design their own IT system. Another one of the main barriers, was the lack of an outcome-oriented clinician culture, one which is open and learning rather than fearful. Peer pressure was seen as potentially useful. Time was not seen as a major barrier and nor was training.

12. In terms of incentives that might drive the implementation, clinical champions enthusiastic about outcome measurement, were seen as imperative. Management support was seen as essential, though a top-down push from management was seen as a disincentive. A key theme which emerged is that routine outcome measurement is unlikely to be viewed as a mandatory activity, until it becomes a biting target in the Healthcare Commission's performance management regime. The Healthcare Commission could incentivise uptake of HoNOS by including coverage in the MHMDS as a performance measure. This is currently the case for coverage on the ethnicity completion within the MHMDS, though such a performance target would drive managers rather than clinicians. Application for Foundation Trust status on the back of the Healthcare Commission annual health check was seen as another key driver. Trusts felt they had a strong incentive to show effectiveness. Commissioners were seen as weak and not a strong driver for the uptake of routine outcome measurement. There were mixed views about whether Payment by Results (PbR) would be a potential incentive to drive the introduction of routine outcomes in the future. Some felt that it would incentivise only managers, but not clinicians, though PbR was felt to still be a long way off for mental health. Interestingly, a modified version of HoNOS called HoNOS Plus has been piloted as an assessment tool for determining 13 activity categories for use in PbR.

13. Current policy developments on outcome measurement fall under the National Outcomes Measures project of the Care Services Improvement Partnership (CSIP) who are trying to ensure the implementation of some of the recommendations from the Fonagy report. They are also responsible for the development of a Compendium of outcome measures which will outline the available measures, their psychometric properties, their uses in different circumstances, and the copyright issues. Outcome measurement does not however receive high policy priority, for example in the NSF.

14. There is a large literature on calculating price indices for mental health care in the US which specifically explores the change in the composition of treatments over time. Rising expenditure on mental health has generated considerable interest in constructing price indices, for example for major depression, schizophrenia and bipolar disorder. This literature shows it is important to focus on the direct medical costs of treating an episode of illness, rather than changes in the prices of the inputs used in treatment. For all three disorders, studies suggest that the price of treating an episode have declined in recent years. This is contrary to many of the officially reported figures because the conception of output allows for a substitution among inputs as a result of technological change. These studies cannot readily be reproduced with existing UK data. While output indices could in principle be generated from the MHMDS and there may even be the possibility of tracking compositional changes in treatment over time, there is no way of producing price indices equivalent to those generated from the medical claims data in the US. No costing data is at present readily or routinely available.

15. It is probably premature to incorporate mental health outcomes into a productivity index. Outcome data collection needs to be improved first and recommendations for this include improving policy guidance and IT systems, and ensuring feedback mechanisms, management support and an outcome-oriented culture.

1.   Introduction

This report examines the feasibility of incorporating patient outcomes in mental health in the NHS into a productivity measure. Measuring health outcomes is a crucial element of assessing productivity. Productivity measurement ideally involves measuring the value of health outputs in relation to the value of health inputs. Both sides of this equation are extremely complex to measure in terms of real value. However not taking account of the health benefits achieved for patients in a productivity measure, can seriously undermine its accuracy.

The Atkinson Review [1] recommended a number of improvements in the measurement of productivity as previously undertaken by the Office for National Statistics, notably improvements in the measurement of output and better measurement of aspects of quality of health care. Research has been carried out to include 'quality' improvements in estimates of NHS productivity [2]. This included adjusting the productivity index for survival rates, waiting times and patient experience. This research did not however attempt to make any 'quality' adjustments for outputs in either primary care or mental health, largely due to data constraints.

This report explores the feasibility of using health outcome measures in mental health for productivity measurement. The objectives of this report are to answer the following questions:

1.   What outcome measures are most commonly used in mental health?
2.   What are the practical issues about collecting these outcome measures routinely?
3.   Can these disease specific outcome measures be converted into a QALY or some other generic measure?
4.   Is there a time series of these outcome measures available?
5.   Is it feasible to examine changes in the mix of treatments in mental health over time?

In considering the key features of an outcome instrument that would be suitable in a productivity index, the following criteria have been assumed about the measure:

- it should have wide (national) coverage,
- it should be applicable in a number of care settings,
- it should be routinely collected in clinical practice,
- it can be readily linked to activity data,
- it can potentially be converted to a generic outcome measure, and
- is available as a time-series

Based on these criteria and in discussion with various experts, two measures have emerged as the most likely contenders for inclusion in a productivity index, namely the Health of the Nation Outcome Scales (HoNOS) and Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). This report has therefore focused efforts on these two measures.

With the assistance of the Centre for Reviews and Dissemination at the University of York, a literature search has been undertaken of a number of bibliographic databases. The search covered the areas of outcome measurement, health status indicators, productivity and performance measurement and output indices for mental health. Over 1500 records were retrieved after duplicates were removed. These were then reduced after inspection of abstracts to around 170 papers. Full papers for these were obtained. The search strategy is outlined in Appendix 1.

Specific attention was paid in the literature to the instrument Health of the Nation Outcome Scales (HoNOS) since it forms a part of the mandatory Mental Health Minimum Data Set (MHMDS) for England. Another reason for focusing on HoNOS, is that the Outcomes Reference Group established by Prof Louis Appleby, National Director for Mental Health and chaired by Prof. Peter Fonagy, concluded that HoNOS '*could provide a developmental anchor*' for all outcome instruments [3], pg. 6. They concluded that while other outcome measures should be explored, they need to be empirically anchored against HoNOS.

Given the policy priority given to HoNOS, the literature has therefore specifically focussed on HoNOS, its psychometric properties, evidence on its reliability and validity, and its use in the NHS.

The literature also covers some international experiences with the mandatory introduction of HoNOS in mental health, notably Australia, where HoNOS is collected alongside one or two other instruments, including a patient self-report instrument. The practical issues around this system, the usefulness and the enthusiasm with which outcome measurement has been taken up, and specific experiences with HoNOS are reviewed.

In addition, in discussions with various experts, a second instrument has emerged as being important in the NHS, namely Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). While HoNOS is used for patients with severe and enduring mental illness, CORE-OM covers patients most often treated in the community setting or involved in psychological therapies - patients with depression and anxiety disorders. And while HoNOS is completed by the clinician or a member of staff, CORE-OM is a self-report measure.

The literature is therefore also focusing on the psychometric properties of CORE-OM and evidence on its reliability and validity and use in the NHS.

While other measures were also explored, such as FACE (Functional Analysis of Care Environments) [4], CUES (Carers and Users Expectations of Services) [5], and MANSA (Manchester Short Assessment of Quality of Life), since these were used in various pilot studies, these instruments do not seem to have as broad a coverage of use in routine settings as the above two measures, and the focus of this report has therefore been primarily on HoNOS and CORE-OM.

Since the literature comes from a number of different perspectives, it tends to use different terminology, namely patient, client, user and consumer. While we mostly use the term patient, occasionally different terms are used interchangeably.

In addition to a comprehensive literature review, this report also covers an extensive number of interviews undertaken with academic and policy experts on outcome measurement, as well as commissioners, managers and clinicians who have tried to grapple with the practicalities of implementing routine outcome measures. In total 28 interviews were undertaken. The interviews focused primarily on the incentives and barriers to routine outcome measurement, issues around the specifics of the instrument choice, and policy considerations. The key messages which emerged from these interviews are described in section 11.

This report starts with a discussion of outcome measurement in mental health and some of the instruments used in psychiatric research (section 2.), the advantages and disadvantages of clinician-rated versus patient-rated instruments (section 3.), and international drives towards routine outcome measurement (section 4.). Section 5. describes the policy background within the UK more specifically with regard to routine outcome measurement and how HoNOS has become an important instrument from a political point of view. Section 6. describes how one might go about implementing a routine outcome measurement system within an organisation and some of the key considerations, including the existing evidence on the benefits of routine outcome measurement in mental health, while section 7. considers the criteria for the choice of a routine outcome measure, including psychometric properties such as reliability, validity and acceptability.

Section 8. then describes in some detail the HoNOS measure, its psychometric properties, its coverage in the service and a description of the Australian case study where HoNOS is mandated, including perceptions of its value as an instrument. Section 9. covers the CORE-OM measure, its psychometric properties, the different types of measures available, and its coverage. The valuation and mapping of HoNOS and CORE-OM to a generic measure such as a QALY and some of the challenges in doing this are discussed in section 10. Section 11. discusses the key themes emerging from the interviews, while section 12. discusses the literature, predominantly from the US which examines changes in the mix of treatment over time and the construction of output indices in mental health, also examining whether the existing data in the UK would enable similar studies to be undertaken. Some concluding thoughts follow in section 13.

This report focuses predominantly on mental health services within the working age population, and doesn't specifically cover Child and Adolescent mental health services (CAMHS) or services for the elderly, nor indeed specialist services such as forensics. Since most of the UK policy efforts have been focused on the working age population group, this has been the key area of development. However, where there are good examples of routine outcome measurement in other areas, for example an initiative called CORC which covers CAMH Services, this is described (see section 11.5).

2. Outcome measurement in mental health

The standard definition of an 'outcome' in mental health care is:

'*the effect on a patient's health status attributable to an intervention by a health professional or health service*'. [6], pg. 4.

There are however a number of concerns with this definition. Firstly, outcomes could improve as a result of self-help rather than professional help, there is not a straightforward link between interventions and outcomes, and outcomes may not always be positive. Maintaining a patient's health status may in some circumstances be considered a positive outcome. Outcomes may also vary according to different perspectives (for example the clinician, the patient or the carer). Mental health interventions may also occur at different levels, for example specific treatments, combinations of treatments, or population-wide treatments. The outcomes may vary at different levels. Because of this complexity, outcome measurement has historically been the domain of researchers, using them in randomised trials [7], rather than clinicians using them in routine clinical settings.

Outcome measures in mental health measure patient's reports of internal psychic phenomena which cannot be externally observed or verified. Classification systems such as the International Classification of Diseases (ICD) diagnose illness according to the presence or absence of mental symptoms that are 'subjective' in their nature. There has been significant work in producing standardised instruments with which to diagnose psychiatric disorders in a reliable manner and quantify the degree of severity of a disorder. Standardised symptom based measures are therefore common in psychiatry. These measures either tend to measure the frequency and intensity of specific psychiatric symptoms (psychopathological rating scales), or they tend to be instruments which judge the impact of the disorder on the individual (measures of social functioning and global measures of outcome, or quality of life assessment).

Slade [8] outlined seven key domains which have been used for outcome assessment in mental health. These include wellbeing (or quality of life), cognitive or emotional functioning, behavioural activities, physical health (and disability), interpersonal and social functioning, society (the burden of caring and public safety) and satisfaction with services.

Some of the more common outcome measures include:

1. Hamilton Depression Rating Scale (HDRS) (used in depression)
2. Beck Depression Inventory (BDI) (used in depression)
3. Hospital Anxiety and Depression Scale (HADS) (used in depression)
4. Brief Psychiatric Ratings Scale (BPRS) (used in schizophrenia)
5. Positive and Negative Syndrome Scale (used in schizophrenia)
6. Mini-Mental State Examination (used in dementia)
7. Health Sickness Rating Scale (global outcome measure)
8. Global Assessment Scale (global outcome measure)
9. Global Assessment of Functioning Scale (GAF) (global outcome measure)
10. Social Adjustment Scale (social functioning measure)
11. Katz Adjustment Scale (social functioning measure)
12. Social Functioning Scale (SFS) (social functioning measure)
13. Index of activities of Daily Living (social functioning measure)
14. REHAB Scale (social functioning measure)
15. Functional Analysis of Care Environments (FACE) (social functioning measure)
16. Lehman's Quality of Life Index (QOLI) (quality of life measure)
17. Heinrichs Quality of Life Scale (QLS) (quality of life measure)
18. Manchester Short Assessment of Quality of Life (MANSA) (quality of life measure)
19. Carers and Users Expectations and Services (CUES) (satisfaction measure)

Gilbody et al [9] did a comprehensive literature review covering the years 1955 to 2000, of outcome measures used in clinical trials and outcomes research in psychiatry. They found that the dominant method of outcomes measurement in randomised trials in psychiatry was the symptom based psychopathology scales listed above.

3. Clinician versus patient-rated instruments

Routine outcome assessment involves either clinician or patient monitoring and rating of changes in health status, and indicators of social functioning (including quality of life). An important distinction here is between patient/user ratings and clinician ratings. Most rating scales in psychiatry are completed by clinicians. The patient voice has tended to be ignored in the development of various instruments to rate health outcomes.

Ford [10] argues there are potential advantages and disadvantages to both the clinician-rated and user-rated approach. Potential problems with clinician ratings include the cost, training, potential for gaming, reliability of the assessment and inter-rater reliability. Potential concerns with self-reports from patients include concerns around social desirability, idiosyncratic completion, and use with patients who are acutely unwell.

Traditionally, the patient's role has been seen as secondary to that of the clinician. However, recent government policy has emphasised the importance of the patient voice and patient choice in decision-making and ideas of 'partnership' and 'shared decision-making' are becoming key in service delivery [11].

Callaly and Hallebone [12] also cite that it is important to collaborate with patients in the choice and development of appropriate outcome measures. Mental health patients want a voice in their treatment plans yet there is little consultation with them either on the choice of instrument, or the practical aspects of their introduction. It is not simply a question of professionals keeping users informed of new developments, but actively seeking and taking into account the user perspective [13]. The issue of confidentiality and the use of the outcomes data, is also of concern to users.

In the Ohio State Outcomes Initiative (described in section 4.), some provider organisations wrote a 'Cheat Sheet' for clients that had difficulty understanding some of the questions in the user instruments. The clinicians therefore tried to re-interpret the questions for clients. This practice was however forbidden in 2007 and consumers were encouraged to answer the survey questions using their own understanding. Manual guidance was written on 'Providing Assistance to the Respondent'. There is therefore clearly an issue in practice about client familiarity with such tools and potentially training users to complete instruments. There is also evidence to suggest that feedback of routine outcome measures may improve outcomes for patients with higher pre-morbid IQs. These patients may therefore be better able to benefit from specific feedback [14].

Dickey [15] recommended a multidimensional approach to rating which could incorporate the patient, clinician and family reports. This would include the subjective experience of the patient, the patient's disease and its effect on the family. This suggests that patients, clinicians and family members may be best placed to provide particular information that may be relevant. Outcomes may therefore be best measured from different perspectives.

However, there has not tended to be much agreement in ratings between clinicians and patients on different scales or even when using the same instrument [16, 17] and this is discussed further in section 8.3.1. The pragmatism of this suggestion is therefore potentially questionable, although there are several examples where clinician, family and user rated instruments are used alongside one another (see for example section 11.5).

4. International drives towards routine outcome assessment

Routine outcome assessment has been undertaken in a range of different ways internationally. Most often the approaches used in various countries have been to inform programmes and systems, rather than helping to inform individual patient-level treatment outcomes.

In the US, the focus on outcome measurement as a measure of success has largely been driven by cost containment efforts. Difficulties in finding appropriate Diagnostic Related Groups (DRGs) for mental health, and the growth in the proportion of health expenditure devoted to mental health (from 4% in the early 1980s to 25% in the early 1990s), has lead to a growing emphasis on outcome measures [8]. Outcomes measurement is increasingly being implemented in public and private programs. Purchaser-driven pressures have driven activity in routine outcome assessment here more than anywhere else. However the mandate for outcomes measures from payers is variable and outcomes measures are used more widely in specialist rather than generic managed care organisations. User-rated instruments are being used in the commercial public sector, in a Medicaid carve-out, by some private psychiatric hospitals and within some public mental health systems. Clinician ratings are used for some state hospitals, and across the Veterans Administration (VA) mental health system [10].

The VA mandated in 1997 that clinicians use the Global Assessment of Functioning (GAF) tool to assess all mental health inpatients and outpatients at least every 90 days. GAF was chosen because clinicians within the VA were familiar with it and it had been used routinely for inpatient discharges since 1991, so training needs would be limited. To incentivise implementation, a national performance measure on GAF recording compliance was introduced, with monthly published monitoring. Implementation was supported by national training initiatives.

The Ohio Department of Mental Health runs the Consumer Outcomes Program[1], one of the state's quality programs, which includes the use of a balanced scorecard, evidence-based practices, continuous quality improvement and licensing and regulation.

The Department set up a Consumer Outcomes Task Force in 1996, engaging consumers and other stakeholders to develop an outcomes measurement system. The task force developed its approach from first principles, defining what mattered to consumers and families. The Consumer Outcomes Program uses three instruments for adults and three for children and families. The adult instruments contain two consumer instruments. The youth scales contain one consumer instrument and one parental/guardian instrument. The pilot projects found that consumers liked being asked about their lives and seeing their outcomes instruments being used in discussion with staff around their treatment plans.

The state introduced a rule in 2003 with formal requirements on provider agencies to implement the consumer outcomes system, linked to agency certification. Implementation has been supported by training, technical support and subsidies. At March 2005, reports were being generated from 49 Boards and 277 provider agencies with records for 211,000 patients [10].

Clinicians in the USA are as resistant to routine outcome measures as their European counterparts for similar reasons – the driving force is seen to be the need for aggregate data for management and accountability purposes rather than direct clinical utility.

Routine datasets that provide activity data for Medicaid billing are extremely well-kept, up to date and almost entirely accurate. When a capitation scheme was introduced in Colorado State it was feared that data quality would decline. In order to keep outcomes and performance at the top of the agenda, the State offered mental health providers a cash incentive for the best outcomes [18].

Report cards and other means of evaluating services are now routinely used in many parts of the US, which has had a substantial impact on the types of care available and the lengths of treatment.

In the UK, there has been the drive towards setting national standards for mental health care in the form of the National Service Framework [19], an emphasis on clinical governance and practice

---

[1] For details see http://www.mh.state.oh.us/oper/outcomes/outcomes.index.html

guidelines, a political emphasis on patient experience, and the development of a high-profile outcome measure for routine clinical use (HoNOS). The UK policy background to the introduction of routine outcome measurement is described further in section 5. Again, the drive has been more to inform programmes and systems, rather than helping to inform individual patient-level treatment outcomes.

Australia has the most coherently developed approach to treatment-level routine outcome assessment. A systematic review of patient outcomes was undertaken as part of the first national mental health strategy [6], which resulted in proposals for the specific instruments to use routinely. These instruments were then independently field-tested for their utility and the resulting recommendations informed Australian practice in routine outcome assessment [20]. The Australian case study is described further in section 8.6.

5. The UK policy background

In December 2002, a national programme was set up to lead the introduction of routine outcome measurement in mental health in the UK. Following pilot work in four mental health Trusts and work in three other exemplar sites, the Outcomes Reference Group was established by Prof Louis Appleby, National Director for Mental Health and chaired by Peter Fonagy, University College London [3]. This reference group was set up to advise on best practice guidance on mental health outcomes measurement by addressing some of the main barriers to its implementation in practice.

One of the key conclusions by this reference group was that it was essential that local Trusts develop the expertise and systems to use local outcomes data effectively so that it can inform local service delivery. Until these core benefits have been established, higher level uses of the outcomes data, such as benchmarking (and productivity measurement) cannot be considered. The notion was to set basic national minimum standards and requirements for Trusts and encourage them to develop beyond these minimum standards.

The key recommendations from the report were as follows:

1.   Outcome measurement should begin with the most tried and tested measures, namely the clinician-rated morbidity measures. HoNOS '*could provide a developmental anchor setting the minimum domains to be covered for the relevant population*' (pg.6). Trusts should consider implementing HoNOS for all service users on enhanced CPA[2] (Care Programme Approach) care, ensuring at least one measurement is taken per year and involving service users in the completion of the measure. If other measures of morbidity are adopted, they should be able to generate an equivalent HoNOS score. In addition, Trusts should also explore the use of other clinician rated and self report measures, including quality of life measures, so as to ensure that all relevant domains can be measured across populations.

2.   Mental health providers need to develop an implementation plan for routine outcome measurement which should take into account integration with existing IT systems, a phased introduction and adequate resourcing.

3.   There needs to be a clear statement from the National Director for Mental Health regarding the uses to which the outcomes data will be put.

4.   Outcomes measurement needs to be integrated within the national IT investment in electronic record systems.

The key recommendation here therefore is that HoNOS is to be the main outcome measure in mental health and should at the minimum be completed for all patients on enhanced CPA once a year. While other outcome measures should be explored, they need to be empirically anchored against HoNOS. The challenges of this anchoring or mapping process are discussed further in section 10.2.2. Since no single measure can capture the whole range of outcomes across mental health [21], other measures could also be explored over and above these minimum requirements.

While this document was not a mandate, these recommendations were subsequently issued as Technical Guidance for the minimum requirements to complete the Mental Health Minimum Data Set (MHMDS). In this regard, the recommendations did become a mandate by the Department of Health (DoH).

It was felt that implementation of routine outcome measurement cannot await the 'perfect' instrument and there would never be agreement on that instrument. Thus Trusts that had already implemented a measure could continue to use it, provided a HoNOS equivalent score could be delivered from the alternative measure, to ensure its inclusion within the MHMDS. For example, South Essex Partnership NHS Trust adopted the FACE measure for local use and derived a HoNOS score for national reporting within the MHMDS.

---

[2] CPA arose out of concern about follow-up care for people leaving psychiatric hospitals. It ensure that support gets properly planned and co-ordinated. It involves drawing up a care plan which identifies the health and social care required from different agencies, the appointment of a care co-ordinator, and regular review. Enhanced CPA is for people with complex or more severe mental health needs (e.g. schizophrenia or manic depression) who need the input of both health and social services.

The report argued that there were good international examples of state-wide outcomes initiatives in Victoria (Australia) and Ohio (USA), which illustrated the success of routine outcome measurement. In fact, both these international examples went down the route of mandating outcome measurement and though the report pointed to the importance of incentives, it did not suggest a similar route per se of mandating outcome measurement, though in effect the recommendations have turned out that way.

The purpose of the pilot sites was to better understand the challenges and difficulties of introducing routine outcome measurement in the NHS. Four mental health Trusts were selected to provide a diverse cross-section, in terms of prior experience of using outcome measurement, and the degree of sophistication of their IT infrastructure (Leeds, Nottingham, South London and Maudsley and West Hampshire). Four different outcome measures were used: Functional Analysis of Care Environments (FACE)[3], HoNOS, Carers' and Users' Expectations of Services (CUES)[4] and Manchester Short Assessment of Quality of Life (MANSA), the latter two being self-report measures. The pilots ran in each site for approximately 6 months.

The pilot sites encountered many difficulties in the course of the implementation and some of the points that emerged from these pilot sites were the following:

1. There was a big discrepancy between response rates for the clinician-rated measures and service user-rated measures, with clinician-rated measures being much higher. Trusts seemed to have enough problems getting a clinician-rated tool like HoNOS off the ground, and the self-rated tools seemed to perform quite poorly.

2. It was found that outcome measurement relies critically on the availability of adequate additional data, such as diagnosis and the interventions provided in order for it to inform the course of treatment.

3. There was a lack of measurement tool which could fully represent the experience of people from Black and Minority Ethnic (BME) groups.

4. There was a lack of consensus around the service user and carer measures of quality of life and satisfaction.

5.   Probably the most important point was that Trusts lack effective systems to promote feedback to clinicians (and patients), and to interpret and understand the outcome measures. Without these feedback mechanisms, any national or local programme is unlikely to be effective. Therefore Trusts need to develop mechanisms to analyse and aggregate the data, so that outcome measurement can help improve the care provided.

While one of the key recommendations from the report was that the National Director for Mental Health should outline the benefits from routine outcome measurement, and this was to be actioned by dissemination of the report through the media, accessible lay guides and articles in key journals, the reality was very different. In fact, the report was made available only to Chief Executives at Foundation Trusts and was therefore not widely disseminated within the service at all.

While this report and the pilots were in many respects a milestone in the evolution of routine outcome measurement in the NHS, the lack of success from the pilots left somewhat of a hiatus in outcome measurement in their aftermath. The Fonagy report was however mentioned often in the interviews and the key points that emerged from participants' comments on it are summarised in section 11.4.1.

---

[3] For more information see www.facecode.com.
[4] For more information see www.rcpsych.ac.uk/crtu/theuseofcues.aspx.

6. How to introduce routine outcome measurement

Slade [8] outlines a four-step process to introducing patient-level routine outcome measurement.

Step 1: examine the key resources required

Consider the psychometric properties of any outcome measure – locally developed tools are unlikely to be appropriate. All collected data must inform the treatment of patients or the evaluation of services. Mechanisms for analysing the data and producing feedback are needed.

Resources required will include leadership, expertise, support staff, IT and clinical time. If these are not available, it should not be undertaken. Lack of these resources will result in haphazard collection of poor quality data which is not used, until the exercise is eventually abandoned or covertly sabotaged (with 100% non-response). This will cause clinicians to believe outcome measurement is not informative and a clinical burden.

Routine outcome measurement can add approximately 10% to the time spent by the clinician per patient [22]. Clinicians still seem unconvinced that this extra time (and thus the reduction in the number of patients they can see) is worthwhile. Since seeing patients is seen as valuable clinical time and filling in forms is not seen as 'work', there are strong disincentives to completing outcome measures [23].

The ease with which outcome measures are implemented as part of routine clinical practice and the extent to which they are taken seriously, will depend on the degree to which clinician's value their clinical usefulness and consider them to be relevant. Staff need to feel a sense of ownership of outcomes data to be committed to their use and appreciate the benefits of the measures. Clinicians have traditionally seen their responsibility as being to the individual patient, whereas the use of outcome measures often requires more of a group or population perspective [12].

A change in culture is therefore needed where the focus shifts from process measures (throughput) to outcome measures. Therefore limiting caseload sizes (to ensure a defined level of quality of care is possible) might be necessary. The introduction of payment incentives for clinicians, the monitoring of outcome data during training and other ways of changing the culture could also be explored. Some of these options are discussed further in section 11.2.

Creative approaches will need to be found to minimise clinical staff time in collecting and analysing data. Automated data entry and the use of non-clinical staff could be explored. For example, the use of computers to provide instant feedback (e.g. using previously entered data to chart progress over time) would be one way of minimising the burden and maximising the potential benefits for clinical care.

The first step is therefore a careful stock-take of all the resources that would be required to implement a system, ensuring these are all in place to facilitate an easy process of data collection and feedback so that the benefits can readily be appreciated and a change in culture can take place.

Step 2: choose the outcome domains

Identify what outcome domains are appropriate to measure. These will vary between the two extremes of what the patient defines as important domains and what the clinician defines as important. A service operating more towards the patient-defined end of the spectrum, will be more interested in domains related to how the service is experienced. A service operating more towards the clinician-defined end of the spectrum, will be more interested in domains covering symptoms and functioning. In reality, most clinical practice will take place somewhere between the two extremes on the continuum, but the exact point will determine the outcome domains selected for assessment [8].

Step 3: choose how the outcome domains will be measured

Choosing how the outcome domains will be measured involves considering issues such as:

1. What constitutes a good outcome for a patient who is not expected to recover?

2.   Is the goal to show that treatment caused improvement, or just that improvement occurred without reference to treatment?
3.   Is the focus on outcome for the patient only, or also relatives or carers?
4.   Is the focus on individual change or will it be important to aggregate the data to examine changes in groups?
5.   Are patient measures or staff measures preferred?
6.   Are generic measures or specific measures preferred?
7.   Are individualised or standardised measures preferred?
8.   Should assessment occur at 'important' times in the care pathway or at fixed time intervals?

Answers to these questions will determine the key principles of the outcome measurement framework.

Step 4: identify the outcome measures

One can then identify the outcome measures that most meet the requirements that have been identified in the first 3 steps. Measures for routine use also need to be feasible, in as much as they are brief, simple, acceptable, available, relevant and valuable [23].

These steps may help services to consider carefully how to implement a well thought-out and well-resourced approach to routinely collecting and using outcome data, rather than something that is an administrative burden added onto the 'real' work of clinicians.

Too often it seems that services start with step 4 which is choosing a measure(s) and then find they fall down at step 1 which is that they haven't got all the resources in place to make routine measurement work. This back-to-front approach to implementing routine outcome measurement may be because too often routine outcome measurement is mandated from the top-down with instruments that don't necessarily measure the domains which are considered important for the service and without having considered all the resource implications.

6.1   What are the applications of routine outcome measurement?

The purpose of the outcome measurement system, and the uses to which the data will be put, is clearly an important consideration for the success of any routine outcome measurement system. The above four-step approach describes a patient-level system aimed primarily at informing clinical decision-making and improving quality of care at a patient-level, the key application being to:

1.   Inform clinical decision-making and help improve quality.

There are, of course a host of potential other applications for outcome measurement data, beyond this level, where more aggregated data might be used. These include:

2.   Evaluate new technologies (in research),
3.   Help to better organise and deliver services,
4.   Help inform the commissioning process,
5.   Drive patient choice,
6.   Manage performance (including consultant appraisal), and
7.   Productivity measurement.

One of the key conclusions from the Fonagy report [3] was indeed that outcomes data needed to, first and foremost, inform local service delivery, before it could be put to higher level uses such as benchmarking (and productivity measurement). Until the core benefits of routine outcome measurement to improve patient-level care could be established, other aspects could not be considered. This clearly involves getting the above four-step process for implementing routine outcome measures right.

6.2   Is there any evidence for outcome measurement?

But is there in fact any evidence that routine outcome measurement can deliver improvements in local service delivery and patient-level care? While there is consensus that outcomes should be routinely measured, is there any evidence of it actually being effective in improving services in any way?

The overall evidence from various reviews seems to be very little [24] or mixed [25]. In a systematic review of nine studies on the addition of routine outcome measurement to care in both psychiatric and non-psychiatric settings, the results showed that the routine feedback of instruments had little impact on the recognition of mental disorders or on longer term psychosocial functioning. While clinicians welcomed the information imparted from the instruments, they rarely incorporated the results into routine clinical decision-making. Routine outcome measurement can be a costly exercise and the authors conclude that there is no robust evidence to suggest that it is of benefit in improving psychosocial outcomes in non-psychiatric settings [26].

In a study by Ashaye et al [27] they examined whether the Care Programme Approach (CPA) in which multidisciplinary team members, patients and their families meet, could provide a suitable forum for needs to be discussed of older people in a psychiatric day hospital. Over a period of one year, 112 day hospital patients were randomly allocated to an experimental group in which the Camberwell Assessment of Need for the Elderly (CANE) was used. Assessment reports were fed back to day hospital staff but not to the control group. Follow-up assessments were performed after three months in the day hospital. At follow-up, both the experimental and control groups had over 60% of their initial unmet needs being adequately met. In both groups, the HoNOS instrument for the elderly patient group (HoNOS 65+) scores and number of unmet needs were reduced indicating an improvement for both groups. However, there were no significant differences in outcome at follow-up between the experimental and control groups.

The evidence suggests that one-off outcome measurements do very little to shift clinical practice or change clinician behaviour.

A Randomised Controlled Trial (RCT) by Slade et al [28] to evaluate the effectiveness of standardised outcome measurement, showed that monthly outcome monitoring reduced psychiatric admissions markedly. 160 adult mental health patients and paired staff were randomly assigned to an intervention group which completed monthly postal questionnaires assessing needs, quality of life, mental health problem severity and therapeutic alliance, and received three-monthly feedback. Both patients and staff received identical feedback. The control group received treatment as usual.

Routine outcome assessment was not shown to be effective, since the intervention did not improve primary outcomes of patient-rated unmet need and quality of life. Other subjective secondary outcome measures were also not improved.

There was however a high staff turnover and a progressive drop in staff assessment return rates which may have indicated a growing lack of enthusiasm because the feedback was not perceived to be useful. The three-monthly feedback may also have been too infrequent to be useful.

However, because the study was longitudinal in nature and had more regular outcome measurement for patients (with month on month assessment) it showed that this can prompt earlier intervention by clinicians to avert relapse which would otherwise end up in hospital and thus reduce admissions. The intervention therefore reduced psychiatric inpatient days and hence service use costs and the intervention proved cost-effective. The intervention cost about £400 per person and for a PCT with a caseload of around 3500 patients, this would equate to £1.4 million. These costs would be more than offset by savings from reduced admissions and service use.

Thus while routine outcome measures did not improve subjective outcomes, they were associated with reduced psychiatric inpatient admissions.

A study by Priebe et al [29] (the MECCA study) examined how patients' views could feed into treatment decisions, because regular outcome measurement in routine settings would more likely happen if it provides a direct benefit to clinicians and patients. MECCA was a cluster randomised controlled trial which followed the same protocol in community mental health teams in six European

countries. In the experimental group, patients' subjective quality of life, treatment satisfaction and wishes for help were assessed in key worker-patient meetings every two months with the intention of informing therapeutic treatment decisions. The trial tested the hypothesis that the intervention would lead to better outcomes in terms of quality of life (and other criteria in patients with psychotic disorders) over a one-year period. A better outcome was assumed to be mediated through more appropriate joint decisions or a more positive therapeutic relationship.

The study was therefore similar to that by Slade et al [28], except that they used hand-held computers to complete the assessment tool. The results showed that while the intervention added time to the clinical appointment, it lead to a significant improvement in patient's quality of life.

The key message from these studies is therefore that one-off outcome measurement (or infrequent outcome measurement) seems to have equivocal results in terms of actually improving subjective outcomes, but outcome measurement that is done longitudinally and more regularly (with more than one or two administrations), can significantly improve patient's quality of life, or reduce psychiatric admissions.

7.   Criteria for routine outcome measures

If one has then undertaken the four-step process to introducing routine outcome measurement in a service, as outlined in section 6. , the final step in the process would have been the choice of an instrument. There are a number of important criteria which need to be considered when choosing an instrument for use in a routine services, and these are described in this section.

A review which nominated six possible outcome measures for routine use in Australian clinical practice, concluded that such measures are:

"*likely to be brief, low cost, multidimensional measures which require minimal training in their administration, scoring and interpretation, but which are sufficiently reliable, valid and sensitive to change to indicate the outcome of the therapeutic intervention*" [6] pg. 33.

Other authors have suggested a host of criteria against which outcome measures could be judged, suggesting they should [30]:

1.   Be relevant and appropriate,
2.   Be simple to use,
3.   Have clear and objective meanings,
4.   Reflect the perspectives of all participants,
5.   Be capable of identifying positive treatment effects,
6.   Be psychometrically sound,
7.   Be inexpensive,
8.   Be understandable to lay audiences,
9.   Be amenable to quick and easy feedback,
10. Be clinically useful, and
11. Be theoretically broad-based.

This is a tall order for any rating tool and while in practice the inability to meet many of these criteria might create barriers for introducing a particular tool (for example concerns over clinical usefulness are discussed at length in the interviews in section 11.1.1), the literature has for the most part focused on the psychometric properties of the instruments. These psychometric criteria, that the instruments need to be valid, reliable, sensitive to change and acceptable, are explained in more detail below.

7.1   Validity

This means that a scale is measuring what it is supposed to measure.

- Content validity refers to the instrument's comprehensiveness (how adequately the sampling of items reflects its aims). Content validity can be gauged by reviewing the content of the instrument.

- Construct validity involves conceptually defining the construct to be measured by the instrument and then assessing the internal structure of the instrument's components and the relationship between the item and subscale scores [31].

- Validity is often indexed by the Cronbach α which indicates the proportion of variance that is covariant between items. Very high values indicate that there are too many similar items and they do not add enough new information to each other. Low values indicate the items do not tap a nomothetic dimension [32].

- Concurrent validity tests the instrument against more established 'gold standard' instruments. Concurrent validity may be measured by examining the correlation of the scale with other appropriate specific measures [32]. Correlations should be high against conceptually close measures.

- Predictive validity assesses the instrument's ability to predict future outcomes such as resource use or treatment response.

The issue of validity raises the question of how a 'good outcome' is defined in mental health.

One of the first issues here is which perspective is relevant, i.e. the patient, or the clinician, or significant others. Clearly patient self-report gives an important subjective judgement of the intervention. But patients may not have information on the expected benefit of treatment or the response of other patients [11].

A second issue is the content of the measure which has traditionally been based on symptoms. Where this captures the severity of a symptom this may be appropriate, but it may be that more than a single symptom score is needed to capture the complexity of many cases [11]. Beyond symptoms, there is also an interest in other aspects of outcome, such as social functioning, satisfaction and recovery.

7.2 Reliability

Low levels of reliability mean that scores made on different occasions will vary.

- Test-retest reliability shows stability of scores over time. It tests the degree of agreement when the same instrument is used on the same patient by the same rater at two different points in time. Large score changes over a short period of time will be indicative of problems [31].

- Inter-rater reliability tests the degree of agreement when the same instrument is used on the same patient by different raters at the same point in time. Intra-class correlation coefficients are typically used to assess this agreement.

7.3 Sensitivity to change

Sensitivity to change is the degree to which an instrument detects change over time as measured against more established 'gold standard' instruments. The measure needs to be sensitive to picking up reliable and clinically significant change in health outcomes over time. Reliable change is the change found in only 5% of cases if change were simply the result of unreliability of measurement. Clinically significant change is the change in an individual's score which might move them from a clinical population to a non-clinical population equivalent score [32].

7.4 Acceptability

Traditionally the main criteria for the assessment of an outcome measure, focus on the psychometric properties. But for it to be clinically useful in a routine setting, it needs to be acceptable both to the patient and the practitioner. The trade-off exists between trying to make measures able to capture complexity, but also brief and informative in everyday clinical practice. The ideal outcome measure has been described as '*the clinical equivalent of the Swiss army knife - small and easily taken into the field, with enough blades and attachments to fit any number of circumstances that may arise*' [11] pg. 24.

Measures which lead to high numbers of missing responses or non-completion rates suggest lower levels of acceptability.

Based on this criteria Bower et al [11] argue that any routine outcome measure needs to be suitable both for primary and secondary care, and one which can reflect severity and complexity in both settings. This is however seldom accomplished in practice and most instruments considered acceptable in one setting (for example HoNOS in secondary care services), are considered inappropriate in another (HoNOS in community care settings).

7.5 Trade-offs between the criteria

There is inevitably a trade-off between precision of an instrument and its ease of use (and hence utility) [33]. The trade-offs that exist between the different criteria are a key challenge for the use of any routine outcome measure. While measures may perform well on the key psychometric properties, they may not be considered acceptable or clinically useful, and other measures which have high utility for use in routine practice, may in fact be built on poor psychometric properties.

8.   Health of the Nation Outcome Scale (HoNOS)

8.1   What is HoNOS?

HoNOS was developed partly in response to The Health of the Nation [34] policy document and the National Health and Community Care Act [35]. In 1993 the UK Department of Health commissioned the Royal College of Psychiatrists' Research Unit (CRU) to develop measurement scales of the health and social functioning of people with severe mental illness. The aim was to provide a means of recording progress towards the Health of the Nation target 'to improve health and social functioning of mentally ill people.' This resulted in an instrument with 12 items which combines the measurement of psychiatric symptoms, hospital service use and 'patient based' aspects such as social functioning [36].

HoNOS is recommended by the National Service Framework (NSF) for Mental Health and by the working group to the Department of Health on outcome indicators. HoNOS also forms part of the Mental Health Minimum Data Set (MHMDS) for England which is a mandatory data collection for all mental health provider organisations in the NHS since April 2003 [37]. The data set covers care received by service users during a spell, and includes details of clinical problems, treatments given, aspects of social care and outcomes (HoNOS scores).

Comparable instruments have also been developed for Children and Adolescent services - CAMHS (HONOSCA) [38] [39] and older people (HONOS65+) [40]. There are also scales for forensic services (HoNOS-secure), learning disabilities (HoNOS-LD) and acquired brain injury (HoNOS-ABI).

HoNOS has been translated into a number of different languages. It is being widely used in Australia and New Zealand (described in section 8.6) and is also being tested in other parts of the world to a greater or lesser degree, including Canada, Denmark, France, Italy, Germany and Norway [41-45].

The 12 HoNOS items are each scored from 0 (no problem) to 4 (severe problem) yielding a total score in the range of 0 (best) to 48 (worst). Ratings are carried out either by an individual psychiatrist, nurse, psychologist, or social worker, or by using input from the clinical team (a consensus rating based on a team discussion). Outcome is measured by comparing a patient's scores at two points in time using individual items scores, the subscale scores and the total score. Scores at time T2 can typically be more challenging to obtain if there are issues of staff changes or patient access.

Background information can also usefully be collected on the profession of the rater and the setting, whether the rating was done by an individual or a team, and background information on the patient, such as diagnosis, date of birth, marital and employment status and Mental Health Act status.

While a number of studies report mean HoNOS scores from different reference samples, e.g. by diagnostic group, such as alcohol-drug-related disorders, borderline personality disorder, severe mental illness and out-patient community samples, no normative data exists from non-clinical or healthy populations with which to make comparisons or consider clinical cut-offs [46], although HoNOS is geared towards patients with more severe and enduring mental illness.

The subscale items relate to social disability and community functioning and the rating period should cover the previous two weeks. The HoNOS takes on average between 5 and 15 minutes to complete, depending on the experience of the rater and the complexity of the patient's problems [47].

Table 1 describes the items of the HoNOS scale, the subscale structure and the scoring. Occasionally a 10-item score can also be obtained instead of the 12-item score, which excludes the last two items which are often not completed for inpatients (on living conditions and occupation) [48]. The instructions on these last two items also differ, asking for patients' 'typical' situation rather than their actual situation in the last two weeks.

Table 1: Items, structure and scoring of the HoNOS

Item	Subscales / sections	Scoring
1. Overactive, aggressive, disruptive or agitated behaviour	Behaviour (1-3)	Each item rated on a 5-point scale:
2. Non-accidental self-injury		
3. Problem-drinking or drug-taking		0. no problem
4. Cognitive problems	Function / Impairment (4-5)	1. minor problem requiring no action
5. Physical illness or disability problems		
6. Problems associated with hallucinations and delusions	Symptoms (6-8)	2. mild problem but definitely present
7. Problems with depressed mood		3. moderately severe problem
8. Other mental and behavioural problems		4. severe to very severe problem
9. Problems with relationships	Social (9-12)	
10. Problems with activities of daily living		Scoring yields individual item scores, subscale scores and a total score.
11. Problems with living conditions		
12. Problems with occupation and activities		

8.2   Training

Training is known to be important both for the correct initial use of the scale and to prevent rater 'drift' [49-52].

Training on the use of HoNOS is run by the Royal College of Psychiatrists Research Unit (CRU). CRU provides 'on site' training and implementation consultancy to Trusts.

Two types of training are available:

1.   basic training for raters, and
2.   training for trainers.

'Basic training for raters' is aimed at groups of mental health professionals within organisations who wish to be trained in the use of HoNOS so that they can use the scales in their routine clinical practice. The workshops include instruction on the rationale and usefulness of outcome measurement. The training includes presentations about the scales as well as group work to enable participants to gain practical experience in using the scales. CRU will send trainers to the organisation and it will cost the organisation £1,500 plus CRU travelling expenses plus VAT for four hours.

'Training for trainers' is targeted at Trusts that wish to implement HoNOS in their organisation by training key personnel to train raters locally. It will also enable the organisation to formulate a strategy for local implementation. It covers the same materials as the 'basic training for raters', but also covers how to run a HoNOS training course for raters and how HoNOS can be implemented in routine clinical practice. It also covers mechanisms for ensuring data quality. The training involves giving participants examples of the sorts of feedback that is provided in some Trusts and then getting them to think about what sort of feedback would be useful to them to receive, so that participants can try to ensure this is generated locally. CRU will send trainers to the organisation and it will cost them upward of £2,500 plus VAT for one day with an optional follow-up day to examine data quality at an additional cost of £1,500 plus travelling expenses plus VAT.

The cost of training staff in the use of HoNOS is therefore not trivial for organisations. Views on training emerged in the interviews and are discussed in section 11.1.5.

8.3   Psychometric properties of HoNOS

During the development of HoNOS, field trials concentrated on adult mental health services for patients in both inpatient and community settings. The developers stated that it was reliable, clinically useful, acceptable, sensitive to change and useful for administration and planning in these settings [53].

However, subsequent independent studies have explored these assertions more thoroughly.

8.3.1 Validity

For the most part, respondents in studies investigating the validity of HoNOS have suggested that it is appropriate, well designed and thorough. However concerns have been raised about the restrictions in rating Item 8 [47, 54], the ability of Item 6 to accurately describe the symptoms of schizophrenia [55], and the difficulty in rating the social Items (10-12) and the amount of information needed, or the lack of available data to make a rating on these items. Overall Item 8 seemed to cause a lot of concern because it encompasses a large number of symptoms, including people with eating, stress and sleep disorders, and neurosis/anxiety disorders, giving it less discriminatory power [47, 56]. They conclude that HoNOS does not provide a good assessment of symptoms.

Concerns have also been raised about the failure of HoNOS to take account of factors such as culture, poverty, abuse, risk and safety, bereavement and medication compliance [31].

Cronbach's alpha has ranged in studies from 0.59 to 0.76 suggesting moderately high internal consistency and low item redundancy. However Trauer in various studies [57, 58], has questioned its sub-scale structure and argued that it does not measure a single construct of mental health status.

In statistical terms one would expect the 12 items to contribute equally to the overall scale, approximately one-twelfth or roughly 8%. McClelland et al [47] found that Items 7-9 make the greatest contribution to the overall score, 15%, 19% and 14% respectively. Again the high loading on Item 8 underscores its lack of discriminatory power and some of the dissatisfaction with it. In contrast, Items 11 and 12 contributed only 3% each, indicating that patients are generally scored low on these items.

In terms of concurrent validity, HoNOS has been shown to perform well against other clinician-rated instruments [31], with the exception of low correlations against the Brief Psychiatric Rating Scale and the Beck Depression Inventory.

Significant associations have been found between HoNOS and other clinician-rated measures such as the Social Behaviour Scale, the Location of Community Support Scale, the Brief Psychiatric Rating Scale, the Global Assessment Scale [21], the Role Functioning Scale, the Life Skills Profile, the Mini-Mental State Examination, the Social Adjustment Scale, the Hamilton Rating Scale for Depression, the Positive and Negative Symptoms Scale, the Broad Rating Scale, and the Global Assessment of Functioning Scale [59].

HoNOS has generally shown poor performance against user-rated instruments [31, 33, 48], including a self-rating version of the HoNOS with a similar question structure [60]. While the correlations tend to vary across different domains of the user-rated instruments, they tend to be lower than those between the HoNOS and clinician-rated instruments. There are a number of reasons why clinician-rated instruments may not match patient assessments. The closer the relationship between the patient and the clinician, the more agreement there is on the assessment. And the more specific and behavioural the questions are on the instrument, the closer the agreement [48]. The HoNOS does not have specific behavioural questions and is often completed by a rater with a limited knowledge of the patient.

Only 3 of the 12 HoNOS scales have been shown to discriminate for outcomes in patients in psychological therapy outpatient settings. There was a floor effect for scores on the other HoNOS scales which suggests that the level of functioning in the psychotherapy and counselling population is too high to be detected by many of the HoNOS scales which are designed for psychiatric populations with severe and enduring mental illness [46]. The HoNOS does not provide enough coverage for the range of problems presented by psychotherapy patients.

Studies have found that high scores on certain items of HoNOS are associated with particular diagnoses. HoNOS can therefore discriminate between different types of patient groups. HoNOS can also discriminate between patients with differing levels of need or disability. HoNOS seems to perform best as a measure of social functioning [56].

HoNOS has been found to have reasonably good predictive validity, being able to explain a significant proportion of the variance in resource use (for example measured by bed days, length of stay and costs) and treatment outcome (for example measured by readmission rates, treatment response and death [49, 61, 62]. It has been shown to have strong associations with service utilisation and therefore

may be able to play a role in casemix systems. Other studies have found a poorer correspondence between HoNOS total scores and resource use [63, 64] and conclude that HoNOS cannot reliably be used in a general 'gatekeeping' role.

8.3.2 Reliability

Test-retest reliability on HoNOS has generally been found to be fair to moderate [54, 55]. Particularly low reliability scores have been reported for Items 1,3,7, and 10. Inter-rater reliability between pairs of raters has been found to be fair to moderate [48, 54, 56, 59], although agreement is poor on Items 4,7,8,9,11 and 12.

Slightly more ratings tend to be done by individuals rather than on the basis of a team consensus (roughly 60%-40% split) [47]. Variation in the number of raters can contribute to low reliability [65].

McClelland et al [47] found a drop of around 50% between time T1 and T2 ratings.

The reliability of ratings are also affected by the variability in the sources of information used when making a rating [65]. In particular, difficulty has been noted in rating the social Items (10-12) when there is a lack of available information, particularly for patients who are hospitalised (Item 11: living conditions, Item 12: occupation and activity) [49], hence the use of the 10-item scale for inpatients. Raters tend to use as their main sources of information: direct observation, interview with the patient, medical records, other staff, and information from family/carers, and to a lesser extent information from GPs, police, ambulance, neighbours, and other sources. There is considerable variability within clinician groups regarding the extent to which they use these different information sources [65]. It is argued that to obtain a valid reflection of patient functioning, a wide range of these information sources should be accessed, though this will have implications for completion time and resource use.

8.3.3 Sensitivity to change

Studies have examined whether HoNOS is capable of detecting change by comparing changes in HoNOS scores to judgements made by clinicians or patients as to whether they had improved, remained stable or deteriorated.

Numerous studies [66-68] found correlations between total HoNOS scores and clinical judgements suggesting that ratings corresponded with patients' and clinicians' views on change.

Other studies have examined change in HoNOS over time in particular settings but have tended to find that there may be an interaction between the care setting, diagnosis, severity and certain items on the HoNOS scale.

Other studies have tested HoNOS' ability to detect change against other instruments. While HoNOS has performed well against some instruments, it has done less well against others [56]. Correlations tend to be better for patients with extreme changes (improvement or deterioration). Some studies have concluded that HoNOS is sensitive to change [33], while others [46] have found that only 3 of the 12 HoNOS scales appeared to be useful for identifying change.

HoNOS may therefore be limited in its ability to detect clinically meaningful change in circumstances where there are less dramatic changes [49].

8.3.4 Acceptability

There has been considerable debate about the utility of the HoNOS [31]. Its suitability for use in routine practice in busy psychiatric services has been questioned [56, 69-72] with suggestions that it has limited value in informing care planning. More enthusiastic views suggest it is a comprehensive, user-friendly tool [41, 47, 53, 68, 73] that could make a valuable contribution in informing clinical judgements.

Audits of the coverage of HoNOS have generally lent support to the more positive view about its utility. In a trial in New Zealand, Eager et al [74] found that 95% of episodes of care had at least one HoNOS rating completed (and the majority had few missing items). On the other hand, only 58% had completed a HoNOS rating at the beginning and end of the episode.

Glover et al [75] found that 60% of mental health provider Trusts in Britain had implemented routine outcome measurement (with the majority using HoNOS. Another study [76] found that 77% of patients in a UK district service had HoNOS scores recorded in their care plans. Many raters are however still considered 'unwilling conscripts' in the process of generating these HoNOS scores [63]. More up to date estimates of coverage from the MHMDS are provided in section 8.4.

Reports of clinicians' experiences with it have been mixed. While James and Kehoe [76] found that UK clinicians were relatively positive about HoNOS, viewing it as potentially useful, many insist its use will depend on adequate resourcing, infrastructure, feedback and training [56]. Some of the positive comments about HoNOS include the fact that it can be completed by non-clinicians, that it acts as a useful aide memoire or 'process measure' in clinical decision-making, and that it acts as a useful focus of discussion in multi-disciplinary teams. Other positive comments were that it allowed one to quantify clinical observations [63], it highlighted patient problems quickly, it was easy to use [77], and indicated the level of risk and improvement [47].

On the other hand, Gilbody et al [9] finds there is still some resistance to its use. Some of the critical comments on HoNOS relate to: the time it takes to complete, inadequate psychometric properties, lack of additional information that it adds to the routine clinical assessment, and lack of enthusiasm among staff. It has been seen as a blunt instrument that is open to misinterpretation [47]. Gilbody et al [9] concluded that for clinicians to be willing to collect such data in the context of routine care, would require them seeing value in terms of the instrument improving the management of the individual patient, and they could find no evidence to support its implementation.

Clinician and manager views from the interviews delve extensively into the perceived utility of the HoNOS in section 11.

8.3.5 Overall considerations

Overall, the HoNOS seems to have adequate or good validity, reliability, sensitivity to change and acceptability. However there may be scope for additional research on aspects such as inter-rater reliability and sensitivity to change, given that HoNOS is being used in a routine context and a number of different raters may be involved in the rating.

In the routine context, it is quite often the case that minor modifications are made to the way the instruments are used, for example, in Australia, in the acute inpatient setting, the rating period is changed from the previous two weeks, to the previous three days to reflect the brevity of such admissions.

McClelland et al [47] also found deviations from the instructions of the instrument, driven by heavy demands placed on staff from their day-to-day responsibilities and the lack of time allocated for HoNOS ratings.

Any kind of amendments to the instrument or the process of rating will have implications for the validity, reliability and sensitivity to change. The above studies have for the most part, considered the psychometric properties of the original instrument used with the standard instructions. In the routine setting there may be many subtle variants on this practice.

A summary of the psychometric properties for the HoNOS, the HoNOSCA and the HoNOS65+ are shown in Table 2 [31].

Table 2: Psychometric properties of the family of HoNOS measures

		HoNOS	HoNOSCA	HoNOS65+
Validity	**Content**	Good	Insufficient evidence	Insufficient evidence
	Construct	Good	Adequate	Insufficient evidence
	Concurrent	Good	Good	Good
	Predictive	Good	Adequate	Insufficient evidence
Reliability	**Test-retest**	Adequate	Adequate	Insufficient evidence
	Inter-rater	Adequate	Adequate	Good
Sensitivity to change		Adequate	Good	Adequate
Acceptability		Adequate	Adequate	Adequate

Source: [31]

8.4 Coverage of HoNOS

HoNOS is the most widely used routine clinical outcome measure used by English mental health services [78]. Given its political backing and coverage, HoNOS would be the most likely contender for use as an outcome measure in a productivity index.

The Royal College of Psychiatrists' Research Unit (CRU) conducted a survey of NHS mental health services in England in 2000 and found 34 services had implemented the use of HoNOS in one or more settings. A further survey by the Department of Health in 2002 found that 61 Trusts were using HoNOS in their services with 5 having implemented the use of the scales across the whole service.

A survey of 340 psychiatrists by Gilbody et al [9] seemed to paint a less rosy picture than the surveys above. The majority of clinicians do not use outcome measures at all in their day-to-day practice, despite political pressures to adopt measures such as HoNOS. However, they do concede that HoNOS seems to have found a place in outcomes measurement in mental health, albeit a small one. Very few clinicians reported being required to collect outcomes measures by their Trust. When asked specifically about HoNOS, 88/340 (26%) reported being required to collect HoNOS for certain patients [79].

The MHMDS has been a mandatory collection (and HoNOS as part of it) since 2003. The following tables show some descriptive statistics from the 2004/05 and 2005/06 MHMDS on the number of Trusts who have HoNOS data integrated into their records, or integrated into their CPA records only, and the proportion of records in each case which are covered. Records in which the HoNOS rating appears, have properly integrated data about the care, the contact and other activity data. The correlation between these HoNOS integrated records and records where HoNOS is integrated into the CPA only, is around 67%. Trusts that complete HoNOS at all, tend to do so as a minimum for their CPA records (as per the minimum requirements from the DoH).

In 2004/05, just over half of the 84 Mental Health providers attached some HoNOS ratings to their records, although there is tremendous variability across these providers. The average completion of HoNOS in patient records out of the 44 providers was 9.6% while for those in CPA only it is less than 3%. The overall completion rate across all 84 providers though, was 5% (Table 4). The highest completion rate for a single provider was nearly 56% with the same provider completing 17.5% of records in CPA. The lowest completion rates were for a provider who had only 1 record with a HoNOS rating in both the overall records and CPA.

Table 3: Completion of HoNOS in MHMDS for 84 Mental Health providers in 2004/05

		No HoNOS	Some HoNOS
Number of providers		40	44
Total records		641,672	716,020
Average number of records per provider		16,042	16,273
HoNOS in properly integrated records	Average completion	-	9.6%
	Maximum completion	-	55.9%
	Minimum completion	-	0.0%
	Std. Deviation	-	13.1%
HoNOS in CPA only	Average completion	-	2.6%
	Maximum completion	-	17.5%
	Minimum completion	-	0.0%
	Std. Deviation	-	4.9%

Comparing 2004/05 and 2005/06, in fact the coverage of HoNOS dropped from 44 providers to 37 providers, although the overall completion rates for those who submitted HoNOS returns remained around 9.5%. Due to the drop in the number of providers completing HoNOS though, the overall coverage dropped from around 5% to 4%.

Table 4: Completion of HoNOS in MHMDS for 2004/05 and 2005/06

		n	Mean	Std. Deviation	Minimum	Maximum
2004/05	All providers	84	5.0%	10.6%	0%	55.9%
	Those who complete some HoNOS	44	9.6%	13.1%	0%	55.9%
2005/06	All providers	84	4.2%	8.8%	0%	45.6%
	Those who complete some HoNOS	37	9.4%	11.3%	0%	45.6%

Initial compliance therefore seems to be waning. It may however also be the case that the MHMDS underestimates actual completion rates since a lot of HoNOS is still completed on paper and may never get entered electronically. In fact, from the interviews it was evident that some HoNOS activity was going on at some Trusts which showed a nil return on the MHMDS. This is discussed in more detail in section 11. In any event, there is still a long way to go to expand HoNOS completion rates in the MHMDS.

While HoNOS is the main method by which outcome is measured for some disorders, it is still used by a minority of clinicians. Following the Health of the Nation policy document aspirations that HoNOS would be collected on a service wide basis, for use in both individual level care and in assessing service provision at a population level, this vision has not been realised in practice yet.

8.5   Is there a time series of HoNOS data?

There should be three years of HoNOS scores available in the MHMDS (2003/04, 2004/05 and 2005/06). However, the 2003/04 data is considered of poor quality and has not been used in any analyses by the Healthcare Commission. There is therefore just two years of useable HoNOS data available at present. As seen from the previous section, the coverage is still a cause for concern in these two years. The data can however readily be linked to activity data within the MHMDS and could in principle form the baseline for the calculation of productivity indices in the future if coverage on HoNOS improves.

8.6   The Australian case study

The Australian case study is instructive because they have gone the furthest down the road of mandatory patient-level routine outcome data collection. In particular, they have mandated the use of HoNOS as a standard outcome measure for all patients receiving in- or outpatient mental health care. All patients should have the HoNOS completed at least twice [48]. This is a substantial commitment of resources by mental health providers and produces a large national dataset on outcomes and casemix.

The clinician-rated outcome measures that have been mandated for use in Australia for public mental health services, are listed in Table 5. Two clinician-rated outcome measures have been mandated for

use in adult mental health services. Three clinician-rated outcome measures are mandated for use in child and adolescent services (CAMHS), and three clinician-rated outcome measures are mandated for use in elderly services [80].

In addition, all adult and elderly mental health services are also required to offer a user-rated (self-report) instrument to consumers. However each state within Australia can choose which one of three user-rated instruments to offer. Victoria, Tasmania and the Australian Capital Territory have chosen to use the BASIS-32, New South Wales, South Australia, the Northern Territory, and Western Australia have chosen the K-10+, while Queensland has chosen the MHI-38. All CAMHS are required to use the same self-report measure, the SDQ [80].

Table 5: Mandated outcome measures in Australia

	Adult services	Child and adolescent services (CAMHS)	Elderly services
Clinician-rated	• HoNOS • Abbreviated Life Skills Profile (LSP)	• HoNOSCA • Children's Global Assessment Scale (CGAS) • Factors Influencing Health Status (FIHS)	• HoNOS65+ • Abbreviated Life Skills Profile (LSP) • Resource Utilisation Groups – Activities of Daily Living (RUG-ADL)
User-rated	• Behaviour and Symptom Identification Scale 32 (BASIS-32) • Kessler-10 Plus (K-10+) • Mental Health Inventory (MHI-38)	• Strengths and Difficulties Questionnaire (SDQ)	• Behaviour and Symptom Identification Scale 32 (BASIS-32) • Kessler-10 Plus (K-10+) • Mental Health Inventory (MHI-38)

In addition the clinician completes a measure called the Focus of Care (FOC) which selects one of four options which attempt to capture the main focus of the care package (e.g. acute care, intensive care) and this is to be used both for adult and elderly care service [80].

A clear protocol has been drawn up to ensure the measures are completed regularly, but in general they are completed at initial assessment, every three months while the patient is in the service, on admission and on discharge from a bed-based service and on discharge in the community. All Australian states have signed agreements to submit routinely collected outcomes and casemix data to the Australian government on a regular basis [80].

8.6.1 Perceptions of the system

Perceptions of the value of the outcome measurement system seem to be mixed [81].

In an early study in rural Western Australia where they investigated clinicians' attitudes towards the use of routine outcome measures 12 months after their introduction, they found that 44% thought they were a waste of time, 44% thought they were useful to track patients' progress while 78% were in favour of their use. They also found that staff needed to be provided with incentives for incorporating outcome measurement into routine practice, and that management needed to support staff with provision of training and support in time and resources [82].

In another early study Trauer [83] found that clinicians at one site were extremely positive about the HoNOS, whereas those at four other sites were more ambivalent, arguing it only minimally improved their treatment.

A study using a questionnaire to evaluate clinician-perceived value of all measures (HoNOS, LSP and BASIS-32), found that perceptions decline over time. Following the introduction of outcome measures there may be an initial period of enthusiasm for their use, but this may be followed by a decline in their perceived clinical value [84].

Callaly et al [80] examined clinicians' and consumers' attitudes towards the process of collecting outcomes data. They cited concerns about the time taken to complete measures, their ability to detect change over time, and an unwillingness to engage in the process of using measures, even if they could be shown to improve services. They also highlighted concerns over their use in community services. They argued that Australian clinicians are concerned that the use of measures will impact on their professionalism [12], since their introduction might suggest that clinicians are not already discussing issues with patients and including them in planning.

Callaly et al [80] also explored clinician attitudes to the integration of outcome measures in everyday practice. Two years after their introduction in Victoria, clinicians were interviewed and results found that equal numbers found them of value and not of value. Their potential clinical value was seen to be to show the broader picture of a patient's progress, improving dialogue between patients and clinicians, using the score as a rule of thumb to gauge patient severity and giving feedback to patients which demonstrates progress. A common concern was the time taken to complete them:

"*I think time is the critical issue and that we are being asked to spend more and more time on collecting information and filling out forms*" [80] pg. 167.

Equal numbers found HoNOS (and LSP) to be of value and not of value. Some clinicians expressed concern that HoNOS gave undue attention to non-symptom areas. There were no positive observations about FOC and it was thought to be of no clinical value.

More positive observations were made about the value of user-rated outcome measures, that they helped patients identify their own needs, that they were useful in service planning, that they helped generate better dialogue and helped the clinician see the world from the patient's point of view. They could also show patients their change over time and bring up discussion of difficult issues [80].

However, the greater the patient severity, the less likely they were to be offered the self-report measure and even if offered it, the more likely they would be to refuse it. This could prohibit service planning for the severely ill group and also skew national figures by the exclusion of data from this group.

In another study, patients who completed a BASIS-32 and then had a chance to discuss their responses with their doctor, reported that they were treated with more respect and dignity by staff than those who had not completed the measure [85]. Patients and carers expressed hope that the use of outcome measures such as BASIS-32 could lead to a fundamental change in the relationship between clinicians, patients and carers [12].

The majority of clinicians believed that the Australian government's primary objective for introducing the measures was for financial management rather than to ensure quality of services. However, a few did support the idea that national data collection could support the ability to compare services and treatment types and thus lead to more efficient and effective services. Others were more sceptical:

"*… when people use it as a management tool and not a clinical tool… it loses the benefits of the tool*" [80] pg. 169.

For clinicians to be engaged in the process of collecting outcome measures, they need to be convinced that the measures can lead to improved quality of care at the patient level and that they will enhance the therapeutic relationship and the clinician-patient dialogue.

In Victoria, computer software was introduced to enable clinicians to complete outcome measures and enter ratings on screen and receive immediate feedback with simple graphs comparing recent ratings with previous ratings for an individual or specific diagnostic group. This was deemed crucial [80]. The most frequent observation was that to make outcomes more useful to clinical practice frequent refresher training is needed. More support to assist clinicians in understanding the meaning and use of outcome ratings is needed. Callaly et al [80] conclude that with appropriate support and training, the increasing use of outcome measures may lead to an increasing sense of clinical value.

9.   Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM)

9.1   What is CORE-OM?

HoNOS is not considered suitable for common mental health problems in primary care and as a result CORE-OM[5] is a widely used measure in these circumstances.

The Psychological Therapies Research Centre at the University of Leeds co-ordinated the development of CORE from 1995-1998 through a multi-disciplinary team of researchers and practitioners representing the major psychological therapy professions.

The Mental Health Foundation (a charity) and the Counselling in Primary Care Trust and the Leeds Mental Health Teaching Trust provided the development funding for the various independent elements of the CORE System from 1995 until the CORE initiative became self-financing in 1998.

CORE Information Management Systems (IMS) provide the on-going support and development for the initiative. There is a network of people who use the instrument. The CORE Network can voluntarily anonymously donate aggregate data to the CORE IMS National Research Database annually for them to build up data on a host of performance assessment aspects of service [86]. CORE IMS is a not-for-profit organisation, it is run independently, and there is no registration for using the CORE-OM tool. Its status as a free tool – copyrighted, but free to photocopy (but not to modify or make profit from), has also made it popular for routine use in practice settings.

CORE-OM is a client-completed outcome measure that has been developed for the adult population and is currently widely in use in psychological therapies in the UK, both in primary and secondary care settings. It has also been verified with older adults [87]. Its purpose is to provide an inexpensive, user-friendly outcome measure sensitive to both low- and high-intensity psychological distress and pathological symptoms for use in research and practice settings [88]. It was designed in consultation with service providers and purchasers who placed a high priority on the measurement of symptoms at intake, and the reduction in symptoms as a result of therapy or counselling [89].

The full CORE-OM comprises 34 items each with 5 levels from 0 (not at all) to 4 (all the time). The CORE-OM has 4 domains covering: *subjective well-being* (4 items), *specific symptoms or problems* (12 items), *functioning* (12 items) and *risk of harm to self* (4 items) *or others* (2 items) [90]. It fits on two sides of A4 and can be hand-scored or scanned by computer [32]. An example of the tool is shown in Appendix 2. It is generic in its theory base. It takes approximately 5 minutes for the patient to complete.

There are several methods for scoring the CORE-OM as shown in Table 6. The score is added to give a total score and then divided by 34 to give an average ranging from 0 to 4 (e.g. 1.83). Because it is easier to work with whole numbers this mean score is multiplied by 10 to give a clinical score [87]. Alternatively, a look-up table can convert the total score to the clinical score as shown in Appendix 2.

The therapist can compare the patient's CORE-OM score with clinical cut-offs for clinical and non-clinical samples of the UK population by using the look-up tables. Severity bands identify boundaries from mild to severe with 10 being the lower bound for 'mild' and 25 or over marking 'severe' in the clinical populations, as shown in Appendix 2.

Table 6: Methods for scoring the CORE-OM

To obtain mean item score:	To obtain clinical score:
1. Add the total score 2. Divide by number of completed items (e.g. 34 if none missing) 3. Result is mean item score from 0 to 4	1. Calculate mean score 2. Multiply by 10 3. Result is clinical score from 0 to 40
Example: 58 ÷ 34 = 1.71	**Example:** 58 ÷ 34 = 1.71 × 10 = 17.1
	1. Add the total score 2. Refer to look-up table to convert to clinical score (see Appendix 2) **Example:** 58 = 17.1

There is also a CORE Assessment (CORE-A) which is completed by a practitioner at initial contact which helps profile the patient, their presenting problems and their pathway into therapy [91]. This allows collection of referral data, demographics, problem severity and duration, and waiting times [92]. An End of Therapy Form can also be completed to contextualise the therapy outcomes and help inform service development. These include things like session attendance rates and therapy ending (planned and unplanned) [93].

Two equivalent short-forms (CORE-SFs A and B) have been constructed from sub-sets of CORE-OM items which can be used for repeat assessment. There is also a 'no risk' CORE-NR version which is based on the full version without the 6 risk items (giving 28 items). There is a GP-CORE version comprising 14 items for use in the general (non-clinical) population. And a version is under development for young people (YP-CORE) [87].

9.2 Psychometric properties of CORE-OM

Its psychometric properties have been tested on clinical and non-clinical samples and it has been shown to have reasonable test-retest stability in a student sample. The internal consistency has been reported as α=0.94 and the 1-week test-retest reliability as Spearman's ρ=0.90 [32].

In a study by Connell et al [94], CORE-OM was administered to a general population sample to assess normative values and find appropriate cut-offs relative to a clinical sample to determine clinical significance. Internal consistency for the general population was α=0.91. A cut-off score of 10 was recommended on the 0-40 scale of the CORE-OM.

It has been found to be acceptable both to therapists and clients and has been used in both primary and secondary care settings where counselling or psychological therapy services are used [91]. It was found to be able to discriminate between patients in secondary and primary care with those in secondary care scoring higher on risk and being above the severe threshold.

It has also shown to be sensitive to picking up change [32]. The CORE-OM has also been tested on an older adult clinical and non-clinical sample and found to be reliable, although there were larger numbers of missing items in older adult samples, potentially questioning its acceptability [90].

In order to test convergent validity CORE-OM has been compared to two older and widely used measures for depression, the Beck Depression Inventory II (BDI-II) and the Hamilton Depression Rating Scale (HDRS), as well as the Beck Hopelessness Scale (BHS) and the Inventory of Interpersonal Problems – Avoidant (IIP-Av). The BDI-II was built on the BDI which has been used for 35 years in the US for assessing severity of depression. Where the BDI-II and HRSD have been designed specifically to assess severity of depression, the CORE-OM was designed as a general-purpose measure of psychological disturbance. The CORE-OM and BDI-II were strongly correlated with each other and showed similar patterns of correlation with the HDRS, BHS and IIP-Av. The CORE-SF also showed convergent validity with the BDI-II on repeated assessments [88]. CORE-OM and BDI-II both use self-report whereas the HDRS is clinician-administered.

Indeed, the high degree of correlation between the original version of the Beck Depression Inventory (BDI) and the CORE-OM has lead to the development of a translation table between the two measures [95]. An equivalent look-up table can also be used for the CORE-OM to BDI-II, given the high correlation between BDI-I and II.

The CORE-OM has been compared to the Clinical Interview Schedule – Revised (CIS-R) and found support for convergent validity [94]. CORE-OM has also been compared to various other scales in clinical samples, including the General Health Questionnaire (GHQ), the original version of the Beck Depression Inventory (BDI), the Beck Anxiety Inventory (BAI), the Brief Symptom Inventory (BSI), the Symptom Checklist-90-R (revised version) and the 32-item version of the Inventory of Interpersonal Problems (IIP-32) [32]. Again, CORE-OM was found to have strong discriminant validity, but there was high covariance between the domains, suggesting that the risk items (6 items) should be separated from the remaining 28 items focusing on psychological distress. These 28 non-risk items scored as one scale provide the most satisfactory scale [89].

A comparison between CORE-OM and HoNOS found a reasonably weak correlation (r = 0.50) between the two which is not surprising given that they are typically used in different settings [21] and

are advocated for use at different ends of the severity spectrum. The 6-item risk sub-scale in the CORE-OM yielded a better association with the overall HoNOS than the CORE-OM total. The authors conclude that where staff trained in the use of HoNOS are not available, and where self-report is appropriate, CORE-OM can provide useful information on risk.

9.3   The software versions

In the past 6 years, there have been some important software developments (notably CORE-PC and CORE-NET).

CORE-PC is a standalone computer-based system which can be run in practices. It was launched in June 2001 and within 12 months 60 services a year were purchasing a license [86]. CORE-PC software provides data analysis and reporting using the CORE-OM tool. Once a year, these practices can anonymously donate data to CORE IMS.

Some potential downsides were found with CORE-PC, namely that this in-house resource had the potential to create 'service performance anxiety', due to the increased transparency of scores across a service, with increased levels of resistance to its use and low completion rates. There were also some practical challenges with implementing the computer system at first. While most services worked through these challenges, a number of services reverted to the original CORE-OM, and abandoned CORE-PC and in some isolated cases abandoned CORE altogether. However the CORE Network is based on the premise of transparency and accountability in the delivery of publicly funded services and tries to encourage a more open and sharing culture. Training has also evolved to address some of the clinical, philosophical and practical barriers to CORE-PC.

There is a reporting function in CORE-PC which also provides a drop-down menu of over 50 reports which have included the performance assessment requirements from the DoH (2004) and the Commission for Health Improvement (2003).

In some services CORE-PC has been developed into an appraisal tool which provides managers and clinicians with an overall profile of the practitioner's contribution to service delivery and quality [96]. An example of an individual anonymised appraisal template from CORE-PC is provided in Appendix 2. This synopsis of performance can easily be printed off as a record for the clinician. This sort of information can then in principle be used as a basis for discussing with individual clinicians, their caseload and associated outcomes data as part of a performance management or consultant appraisal scheme. It has been found that if the performance management is done in such a way that clinician's do not feel they stand to lose from such data, then data gaming is far less likely.

CORE-NET[6], which is currently being piloted, uses a short version of the CORE-OM. It is supported by IT systems and can be completed at every session by the patient. This is an internet based tool which can be used to implement CORE-OM. It provides an instant visual mapping of scores and benchmarking to provide quick feedback in the clinical setting.

It can be used as an assessment and screening tool for depression and takes 2 minutes to complete. It is based on 10 questions from the CORE-OM and covers measures of subjective well-being, functioning, risk and symptoms. It can be completed by the patient on the computer prior to the GP appointment, or together with the clinician during the appointment, or it can be completed on paper and then entered by the receptionist or other staff prior to the appointment. Rating is based on the patient's perceptions of the prior week. Figure 1 shows a shot of the computer-based programme which patient's complete. As with the CORE-OM, items are rated from 0 'not at all' to 4 'all the time'.

---

[6] A useful video presentation of the CORE Net system can be viewed from the Ashton, Leigh and Wigan PCT demonstration site at: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

Figure 1: Screenshot of CORE-NET screening tool, the CORE-10 measure
Source: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

At follow-up, patients complete a short-version of the CORE-OM on the computer, a CORE-5 questionnaire.

Figure 2 shows a progress chart which provides instant feedback with the patient's average score displayed against the clinical cut-off lines, determined by normative data from samples of clinical and non-clinical populations. The blue line tracks the overall CORE-OM score while the red dot shows/tracks the risk separately. The chart also displays a risk cut-off line. Progress is tracked over time with this CORE-5 measure and the version of the instrument used at each consultation is displayed at the bottom of the chart. The interventions used, for example, the use of therapies, or medications, or combinations of them, are all noted at the bottom of the chart for each consultation (in abbreviations). At discharge, the 10-item questionnaire might be used again. The original 34-item questionnaire and an 18-item version are also available for more in-depth assessments.
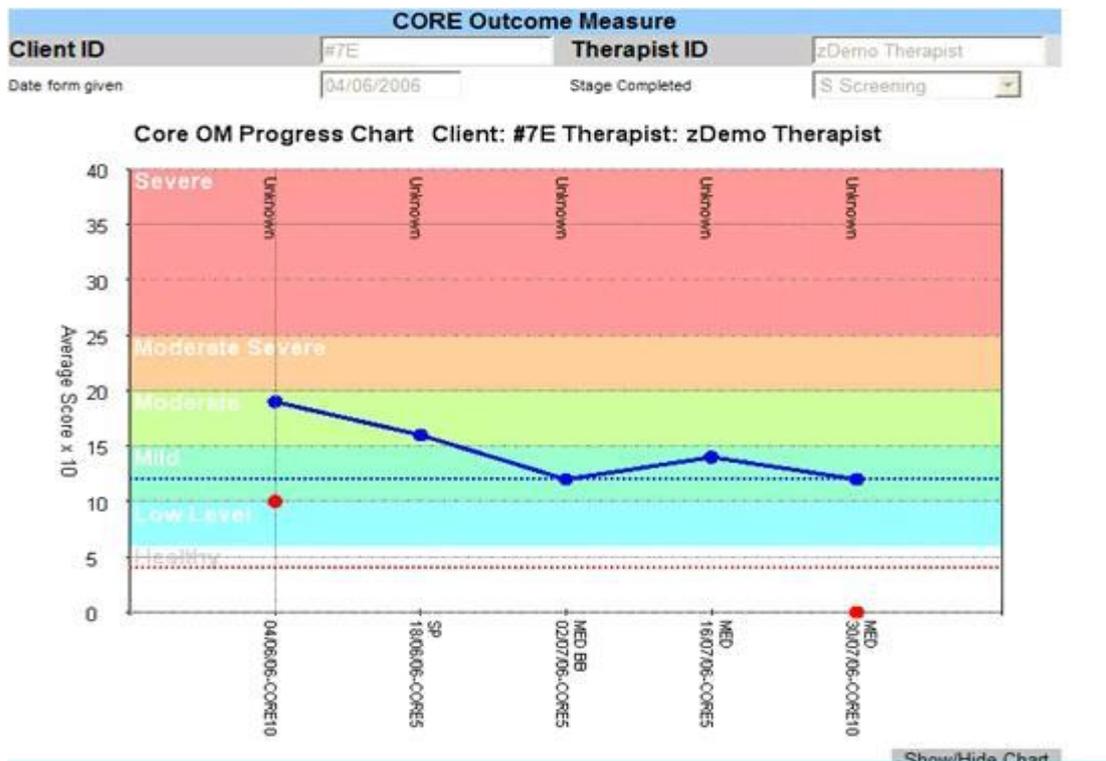
Figure 2: Screenshot of progress chart on CORE-10 and CORE-5 measures
Source: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

Patients and clinicians can track progress on screen at each consultation and a review chart highlights whether the change is clinically significant or not as shown in Figure 3.

| | | I have someone to turn to for support when needed |
| able to cope when things go wrong |
| people has felt too much for me |
| panic or terror |
| ans to end my life |
| difficulty getting to sleep or staying asleep |
| despairing or hopeless |
| unhappy |
| d images or memories have been distressing me |

Show/Hide Review

ost Score	Cut Off	Score +/-	Clinical Change	Reliable Change
2	1.19	-0.7	No Change	Improvement
33	1.36	-0.67	Improvement	Improvement
67	1.29	-0.33	No Change	No Change
17	1.44	-0.83	Improvement	Improvement
	0.43	-1	Improvement	Improvement
	1.37			

Figure 3: Review chart of significant changes in outcomes
Source: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

The visual feedback allows patients to see progress in a tangible way over multiple consultations and can provide them with positive feedback and encourage recovery. The immediate visual display is considered very fruitful in consultations as it gives feedback of the patient's experience (as shown in Figure 4).

Figure 4: Patient receiving feedback on CORE-OM
Source: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

There is also a colour-coding of the particular items on the questionnaire which have contributed most towards the overall score. These can provide a quick way of assessing the key factors to focus on in the consultation as shown in Figure 5.

Figure 5: Items with high weighting in the CORE overall score
Source: http://s140798917.websitehome.co.uk/corenetgp/corenetgp.html

Groups of patients can also be flagged or tracked, for example those who are most at risk, or those who have deteriorated over time, or those who require discharge from treatment because they haven't been seen for some time.

The CORE-NET tool can be used for Quality and Outcomes Framework (QOF) claims to screen patients for depression and it gives the severity for the morbidity component of the QOF entry as well. CORE-OM can provide nine of the eleven primary symptoms for assessing depression according to NICE guidelines [97]. The programme also includes the levels of stepped care for the treatment of mild to moderate depression as recommended in the NICE guidelines [11]. Within a stepped care system, if a first-line treatment is unsuccessful, patients are 'stepped-up' to a more intensive treatment. Routine outcome measures are seen as one way in which decisions can be made about whether patients have had a good outcome at one step and whether they need 'stepping-up'.

9.4   CORE-OM coverage

CORE-OM is believed to be the most widely used tool in routine evaluation in psychological therapy and counselling services [98].

It has extensive coverage in around 250 services in the NHS. Coverage in some services is as high as 80-90% of all patients. The database which CORE IMS holds covers around 100,000 patients per annum and in total around 250,000 patients.

Because the paper version of the tool can be disseminated informally (freely downloaded, photocopied and used), it is impossible to know fully who uses the instrument. However, it is possible to know who uses the software versions. At present around 30 Mental Health Trusts and around 75 PCTs use the software in psychotherapy and counselling services.

A study on the CORE National Research Database assessed completion rates for CORE-OM at time T1 and T2 and found pre-therapy CORE-OMs were completed by 83% of clients on average, while post-therapy forms were completed by 39%. These average figures did however mask large variations between different services ranging from 3% to 99% [99].

Overall these completion rates are much higher than for HoNOS. One of the reasons may be that CORE-OM has been a grass-roots development focusing on professional self-scrutiny, rather than a top-down focus on external evaluation.

9.5   Is there a time-series of CORE-OM data?

In principle therefore, there is a time-series of data available on CORE-OM from the CORE National Research Database held by CORE IMS. They hold data from 1999 onwards and this could in principle be linked to some activity data in their database where treatment or therapy has been included. Access to this database would need to be negotiated for any research purposes. Data quality in early years would need to be tested carefully.

10.   Can outcome measures be converted into a generic measure?

In order to produce a productivity index using either HoNOS or CORE-OM (or both measures), one would need to consider whether they could be converted into a generic measure like a Quality Adjusted Life Year (QALY) so that valuations of health benefits can be made across the whole health service.

The issue of conversion to a generic measure is still highly problematic in mental health. There is very little literature on the joint use of HoNOS and CORE-OM alongside a generic measure like EQ-5D (the tool used to produce a QALY), the valuation of either HoNOS or CORE-OM, or indeed mapping from either instrument to a QALY.

10.1 The use of generic measures in mental health

Resource allocation decisions are often based on cost-effectiveness criteria where interventions are assessed in terms of their cost per QALY. The QALY provides a way of measuring the impact of healthcare interventions and improvements in health related quality of life (HRQL). The most commonly used measure is the EQ-5D which has also been recommended by NICE. The EQ-5D covers 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has 3 levels: no problems, moderate problems, extreme problems. The EQ-5D provides a five-dimensional description of health status which can be defined by a five-digit number. For example the health state '11122' indicates no problems in mobility, self-care and usual activities, and moderate pain/discomfort and moderate anxiety/depression.

With 5 dimensions and 3 levels, approximately 3^5 = 243 different health states can be defined. EQ-5D is a preference-based measure, in other words, it is scored using values taken from the general population through a large-scale survey in the UK [100]. QALY weights are based on societal rather than patient values because resource allocation decisions are based on societal preferences.

EQ-5D is followed by a visual analogue scale (VAS) similar to a thermometer which ranges from 0 (the worst imaginable health state) to 100 (the best imaginable health state).

While the measurement of quality of life is a key concern in mental health services, the QALY has not been widely used as a scale [101]. Generic instruments designed to measure HRQL are generally considered to be less sensitive to changes in patient's health state compared to disease-specific instruments which focus on certain aspects of a given disease. Generic instruments are designed to be broadly applicable across different types and severities of disease [102].

The claim that EQ-5D and other generic preference-based measures are applicable to all interventions and patients has some merit with physical conditions where it has passed psychometric tests of validity and reliability. But for conditions such as visual impairment and hearing loss, it has not [103].
For mental health there is evidence that outcome measurement tools like EQ-5D [104] and other generic instruments (like SF-6D) [105] can be used to detect the impact of conditions such as mild to moderate depression and anxiety. But often generic measures like SF-36 are not practical with acutely disturbed patients [64, 106].

In a Dutch study both EQ-5D and SF-6D discriminated between severity subgroups and captured improvements in health over time [107].

In an RCT study by Sherbourne et al [108], they used utility weighted SF-12 or SF-36 measures and physical and mental health-related quality of life scores on a group of patients with depression. Several utility-weighted measures showed increases in utility values for patients in one of the interventions. However, QALY gains were small. They argue there are concerns about the use of direct single-item utility measures or utility measures derived from generic health status measures in effectiveness studies for depression. Utility measures that capture the mental health and non-health outcomes associated with treatment for depression are needed.

A study by Gunther et al [109] examined if the EQ-5D could be a valid measure in alcohol dependent individuals. They compared the EQ-5D against other measures of psychopathology and social

functioning (including HoNOS). Twelve percent of the individuals indicated "extreme problems" in at least one dimension. The EQ-5D VAS score, and the EQ-5D index showed moderate correlations with other scales. They found some evidence for EQ-5D's validity in the population group, but the EQ-5D showed a ceiling effect.

Evidence on how EQ-5D performs in groups of patients with chronic schizophrenia is more patchy. A similar study examined the psychometric properties of the EQ-5D in patients with schizophrenic, schizotypal and delusional disorders [102]. The patients again completed the EQ-5D alongside other subjective and objective psychopathology scales, and measures of functioning (including HoNOS). For almost all EQ-5D dimensions, different response levels were associated with significantly different scores of measures used for comparison. EQ-5D did not correlate well with a condition-specific instrument for schizophrenia. Again the EQ-5D showed a ceiling effect, although they concluded it seems to be reasonably valid in this patient group.

Another study found that EQ-5D and its visual analogue scale (VAS) failed to pick up changes in a group of patients with chronic schizophrenia [110]. They concluded that its focus on physical components were not suited to the psychiatric context. Its use as the core measure in economic evaluation in the field of psychiatry therefore seems less appropriate. The content of generic measures do not seem to capture the impact of severe psychotic and more complex non-psychotic conditions.

In the Gilbody et al [9] review of outcome measures used in clinical trials and outcomes research in psychiatry, they found that the dominant method of outcomes measurement in randomised trials was the symptom based psychopathology scales. The use of generic patient based measures is still largely absent in psychiatry. They argue that there is no robust research evidence to support their use as a routine outcome measure in psychiatric settings.

Generic measures may be difficult to apply to mental health patients since they often concentrate on physical functioning and ignore aspects of social functioning that are important in mental illness. They may therefore have large numbers of irrelevant questions making them unacceptable to respondents and they may therefore also be insensitive to detecting changes in health status. Generic outcome measures are essentially designed to identify health status changes at a population level and the rating scores are often un-interpretable at an individual patient level.

While a generic measure may be considered useful for supporting health policy decisions, or indeed calculating productivity indices across the health services, or for cost-effectiveness analysis, they are not considered sensitive enough or useful enough at the individual patient-level. Interviewee responses to the introduction and use of a generic measure like a QALY in the service are explored further in section 11.3.2.

10.2 Valuing and mapping HoNOS and CORE-OM

10.2.1 Valuing HoNOS and CORE-OM

None of the disease-specific instruments in mental health research have been scored using preference values so that they can be used to derive QALYs. They are generally not suitable for valuation using preference elicitation methods given their size and complexity.

If one considers the complexity of valuing EQ-5D with 5 domains and 3 levels as a dimensionality problem of 3^5 giving approximately 243 items, the equivalent calculation for HoNOS would be 12 domains and 5 levels (5^{12} = 244, 140,625 items + death) and for CORE-OM would be 5^{34} (even more items). These are too complex to create valuations for. Even if one took the 10-item version of HoNOS and the CORE-10, this would still amount to finding valuations for approximately 5^{10} = 9.7 million items for each).

The twelve dimensions of HoNOS are given again in Table 7. An example of a valuation of a HoNOS health state from these twelve dimensions is given in Table 8.

Table 7: Dimensions of HoNOS

1.	Overactive, aggressive, disruptive behaviour	7.	Problems with depressed mood
2.	Non-accidental self-injury	8.	Other mental and behavioural problems
3.	Problem-drinking or drug-taking	9.	Problems with relationships
4.	Cognitive problems	10.	Problems with activities of daily living
5.	Physical illness or disability problems	11.	Problems with living conditions
6.	Problems associated with hallucinations & delusions	12.	Problems with occupation and activities

Table 8: HoNOS Health State 321214435253

1.   Occasional aggressive gestures, pushing or pestering others; threats or verbal aggression; lesser damage to property (e.g. broken cup, window); marked overactivity or agitation [3]
2.   Fleeting thoughts about ending it all but little risk; no self-harm [2]
3.   No problem with drinking or drug taking [1]
4.   Minor problems with memory or understanding, e.g. forgets names occasionally [2]
5.   No physical health problem [1]
6.   Marked preoccupation with delusions or hallucinations, causing much distress and/or manifested in obviously bizarre behaviour, i.e. moderately severe clinical problem [4]
7.   Depression with inappropriate self-blame, preoccupied with feelings of guilt [4]
8.   Other mental and behavioural problem clinically present, but there are relatively symptom-free intervals and patient/client has a degree of control, i.e. mild level [3]
9.   Severe and distressing social isolation due to inability to communicate socially and/or withdrawal from social relationships [5]
10. Minor problems with activities of daily living; e.g. untidy, disorganised [2]
11. Accommodation is unacceptable; e.g. lack of basic necessities, patient is at risk of eviction, or 'roofless', or living conditions are otherwise intolerable making patient's problems worse [5]
12. Limited choice of activities; e.g. there is a lack of reasonable tolerance (e.g. unfairly refused entry to public library or baths, etc); or handicapped by lack of a permanent address; or insufficient carer or professional support; or helpful day setting available but for very limited hours [3]

Using time-trade-off or other similar methods, someone would be asked to imagine themselves in this health state for, say, the next 40 years before they die. If they could trade this for full health now, how many years would they be prepared to live in full health to equate to the 40 years in poor health? If they respond, say, 30 years, this would give a utility weight of 30/40 years or 0.75.

It would be very difficult however for the general population, whose values are typically used, to imagine what this health state may be like. This is often very problematic in mental health conditions. Some of the problems described are alien to many members of the public or very difficult to envisage. This may lead to a set of ill-informed values.

The HoNOS statements are also quite complex. For example, the five levels for dimension 6 (hallucinations and delusions) are given in Table 9.

Table 9: HoNOS – the 5 levels for dimension 6 (hallucinations and delusions)

1.	No evidence of hallucinations or delusions during the period rated
2.	Somewhat odd or eccentric beliefs not in keeping with cultural norms
3.	Delusions or hallucinations (e.g. voices, visions) are present, but there is little distress to patient or manifestation in bizarre behaviour, i.e. clinically present but mild
4.	Marked preoccupation with delusions or hallucinations, causing much distress and/or manifested in obviously bizarre behaviour, i.e. moderately severe clinical problem
5.	Mental state and behaviour is seriously and adversely affected by delusions or hallucinations, with severe impact on patient and/or others

In order to make the valuations, a set of notes accompany each of the dimensions. For example, the notes for item 6 on the HoNOS scale (hallucinations and delusions) would state in addition to the details given about the 5 levels that the ratings should:

● Include hallucinations and delusions irrespective of diagnosis.
● Include odd and bizarre behaviour associated with hallucinations or delusions.

- Do <u>not</u> include aggressive, destructive or overactive behaviours attributed to hallucinations or delusions, rated at Scale 1.

Given that the descriptive system for each health state is so complex, this would lead to cognitive overload for respondents.

Making a valuation and taking into account these accompanying notes, is therefore complicated. In order to make valuations, one might consider simplifying these descriptions, however this would negate the reliability and validity of the instrument.

There would in principle be three ways to value HoNOS or CORE-OM.

The first is the classical statistical inference approach which is typically used for something like EQ-5D. One would value a subset of the health states using techniques like those mentioned above, asking the respondent to imagine themselves in a particular health state and then trading that against perfect health [111] (time trade-off, standard gamble, discrete choice experiment, visual analogue scale or ranking). For each of the health state valuations, one would need around 30-40 respondents to make the valuations. One could then model the missing values. For EQ-5D this might have involved obtaining 30 valuations of different health states and then modelling the remaining 220 out of approximately 250 health states.

If however one were to obtain a similar tractable number of health state valuations for HoNOS, say 30 valuations, the rest of the 244 million health states would still 'theoretically' need to be modelled. Because there are so many permutations of health states in HoNOS, there would be a huge amount of uncertainty.

A second approach would be to try to value the full descriptive system using a multi-attribute utility function approach. In this method, one would get the public to value each of the domains by valuing the corner states (the worst state on each separate dimension). This would entail 12 items. One would then get them to value the levels in each domain and thus build up the score. This would entail a further 4×12 items. This method however relies on a lot of assumptions that one can treat the items independently and add them together. While it simplifies the valuation task for the individual, the valuation has less predictive value in terms of predicting responses.

A third method would be to try to obtain a condensed descriptive system. One would try to find a subset of questions that were considered most important, based on an analysis of existing HoNOS or CORE-OM data using various statistical methods. One could then value this subset of questions using one of the above approaches. This would make the problem more tractable. An example would be producing the SF6D from the SF-36. One can therefore reduce the dimensionality of the problem by looking at the extent to which certain domains or levels are redundant and reduce the descriptive system, to say 7 domains with 3 levels. This approach, called rasch analysis, is a way of condensing the descriptive properties because some domains will clearly have a bigger impact on HRQL than others.

Indeed for HoNOS it was found that Items 7-9 make the greatest contribution to the overall score, while Items 11 and 12 contributed very little [47]. A statistical data reduction process could therefore be possible in principle to make the problem more manageable.

One important consideration regarding HoNOS valuations is that HoNOS is supposed to be completed by the healthcare professional with reference to the more severe problems that occurred during a prior period. Therefore, if the patient had a normal mental state for the majority of the period of evaluation, but then had a severe psychotic episode in one day, HoNOS would take account of this significant episode. This gives a lower HRQL bound, unlike a QALY which gives the average rating over a period. Utility scores will therefore be lower.

Valuing HoNOS, or indeed CORE-OM, is therefore fraught with many difficulties.

10.2.2 *Mapping HoNOS and CORE-OM*

Given then that it seems unlikely that generic measures might be routinely used in psychiatry, how easily can the disease specific measures be mapped onto the generic measure? There is a growing consensus that a tool is required to capture patient based outcomes associated with mental health that can be used to map to EQ-5D and generate QALYs [108]. To date, there seems to be no established methodology to map tools like HoNOS onto generic outcome measures like EQ-5D.

One could therefore consider mapping across from one instrument to another using judgement or statistical inference. This approach (called cross-walking) would allow one to develop a look-up for a value of HoNOS, giving its equivalent EQ-5D score. This could be done by using both EQ-5D and HoNOS together on the same set of patients and generating equivalent mapped values on either instrument. Ideally it would entail the collection of: clinician-rated HoNOS scores, clinician-rated EQ-5D scores, and user-rated EQ-5D scores on the same set of patients. It is likely that there is a distinction to be made between the mapping of user rated scales and clinician rated scales onto generic outcome measures.

However, there are not many instances where HoNOS and EQ-5D have been used on the same set of patients. There were isolated examples in the literature, for example two German studies [102, 109]. However, as outlined in section 10.1, EQ-5D does not always accurately reflect mental health problems for some patient populations. No studies could be found using both CORE-OM and EQ-5D on the same populations.

There are also some concerns around mapping from a disease-specific measure (like HoNOS) to a generic measure (like EQ-5D), as outlined by Brazier et al [112]. Some of the criticisms around the different methods of mapping include its arbitrariness, and not estimating the uncertainty around the mapping, for the judgement approach. Using the statistical inference approach, the mapping functions which have been used to fit condition-specific to generic measures have also been found to be variable, with relatively low goodness-of-fit results. There are no clear-cut criteria for assessing the performance of the mapping. Performance depends on the degree of overlap in content between the instruments being mapped. The degree of overlap between HoNOS and EQ-5D is questionable.

While proponents of CORE-OM argue that it can more readily be mapped to EQ-5D than HoNOS, many of the same concerns regarding valuations and mapping would hold for CORE-OM too.

Brazier et al [112] argue that, in the end, mapping can only ever be a second best solution to using a generic preference-based measure in the study in the first place. Since no preference-based measures exist for mental health, there have been recent attempts to try to get funding for the development of a generic preference-based outcome measure in mental health (personal communication).

Given these immense challenges in mapping from one instrument to another, it is really not clear how mental health providers should attempt to obtain HoNOS equivalent scores for inclusion in the MHMDS when they use other morbidity measures, as recommended by the Fonagy report. It is not clear that Trusts would necessarily have the expertise to embark on a complex process of mapping from one morbidity measure to another for all the various rating instruments which clinicians might use in practice.

11. Themes emerging from interviews

During the course of this project, 17 interviews were held with various policy makers and academics involved in outcomes research. These included the Royal College of Psychiatrists Research and Training Unit (CRU), CORE IMS, the Department of Health and National Institute for Mental Health in England (NIMHE), Care Services Improvement Partnership (CSIP), the Mental Health Observatory, North East Public Health Observatory, and the Healthcare Commission.

In addition, 11 interviews were undertaken at 3 different NHS mental health provider organisations with managers, psychiatrists, and other clinical staff involved in the collection of outcomes data. Commissioner views were also obtained.

Ethics approval was obtained for the research from Leeds (East) REC. The sites were chosen on the basis of their HoNOS returns in the MHMDS. All sites had positive returns, but to varying degrees, ranging from around 2% to nearly 10%. Interviews were on a semi-structured basis and lasted about 45 minutes. Interviewees were assured of their confidentiality in the discussions. Interviews were recorded and the qualitative data was explored thematically. The interview schedule is attached in Appendix 3.

There were a number of themes emerging from the interviews, most notably these were around the incentives and barriers to the introduction of outcome measures, and the choice of the particular instruments.

11.1 Barriers to routine outcome measurement

11.1.1 No clinical benefits can be seen

Probably the most crucial barrier to the introduction of outcome measures is that clinicians are unable to see the clinical relevance or the clinical benefits. This is partly because they have not been given a clear rationale for the use of outcome measures, partly because they are simply told to complete scores, but primarily because they never receive any feedback on them. This lack of feedback is also associated with a lack of IT systems in being able to provide appropriate and useful clinical feedback.

Clinicians are weary of having to complete additional paperwork if they can't see direct benefits to themselves or their patients. Many view the completion of forms as akin to:

"pouring valuable clinical information into a black hole".

If no feedback is received, compliance and completion rates start to wane and the benefits are not appreciated. Often clinicians produce data which then gets used only at aggregate level, and is not fed back to them, hence it becomes extremely difficult to maintain enthusiasm for what is seen as a paper-filling exercise for managers.

Not only does there need to be clear communication with staff about what the clinical benefits are, but these would best be appreciated if feedback were received in a clinically useful and timely manner.

"(Clinicians)… don't fully understand what outcome measurement can potentially deliver."

In some instances, some senior clinicians have actively tried to derail attempts to measure outcomes. When clinicians are told to complete ratings without appreciating the benefits, they may covertly try to sabotage attempts with a hope that outcome measurement will 'go away'. Many clinicians believed that if they ignored outcome measurement, 'it would go away.' But most now realise that it won't go away and that they will probably have to engage with it in one way or another.

Clinician engagement is absolutely needed to get outcome measurement implemented. The best way to achieve clinician engagement, is for clinicians to receive meaningful reports back which can be seen as clinically relevant.

There were a number of examples of where feedback mechanisms had been introduced in services to varying degrees of success.

In one Trust, clinicians completed HoNOS scores, but the IT system only allowed them to receive back the total HoNOS score, not enabling any feedback about specific parts of the scale. The data was not linked to socio-demographic characteristics of the patients or to the team caring for the patient and the treatment plan, thus not allowing valuable feedback on how interventions are working for certain patient populations. These total HoNOS scores were not seen to be meaningful.

In another service, as per the DoH mandate, HoNOS ratings on enhanced CPA were completed for a number of years. However clinicians never received any feedback on all the data they were producing, and because they could not see any clinical benefits from this exercise, completion rates dropped.

In one Trust, a feedback session was observed where clinical teams were shown analyses of the data which they had been collecting on HoNOS over time, taken from the electronic patient record system. In this Trust, clinicians could not obtain feedback individually directly from the electronic system, and instead investment was made by the Trust to pay two clinicians with research and data analysis expertise to produce aggregate feedback of the sort shown in Figure 6 through Figure 9. On earlier versions of their electronic patient records system, teams could graph changes in HoNOS scores for individual patients over time, but that facility had been removed in the belief that it was not widely used.

Unfortunately, the feedback session was poorly attended, which may have been indicative of the level of interest in clinical teams in obtaining such information, although feedback sessions to other teams had apparently been attended very well.

No feedback was provided on individual clinicians or staff members, it was provided instead at Community Mental Health Team (CMHT) or specialist team level.

Figure 6 shows the average HoNOS ratings at time T1 for 3 CMHTs and gives a profile of the average level of severity seen by the teams. It suggests statistically significant differences between teams in terms of the average HoNOS scores, with lower scores suggesting lower severity.
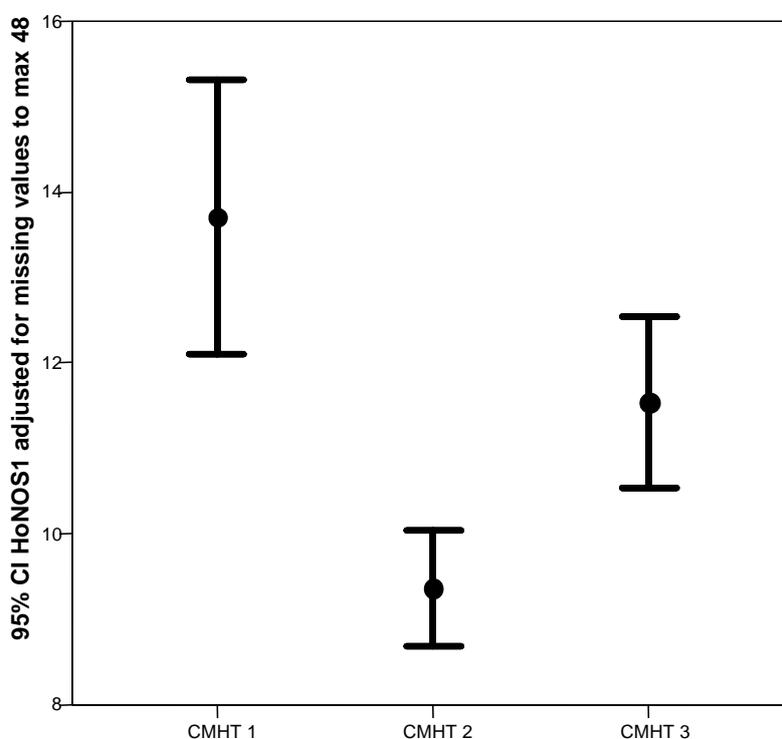
Figure 6: Example of feedback given to clinical teams at a Trust on HoNOS ratings at time T1

Out of all the episodes of care that were undertaken in the period of analysis, only 6.3% of all episodes had a paired HoNOS rating (in other words ratings at time T1 and T2). Given the small numbers, not a huge amount of analysis could be undertaken at the more detailed level, and hence more aggregate analyses were shown. Figure 7 shows the useable pairs of ratings over time (starting with zero).
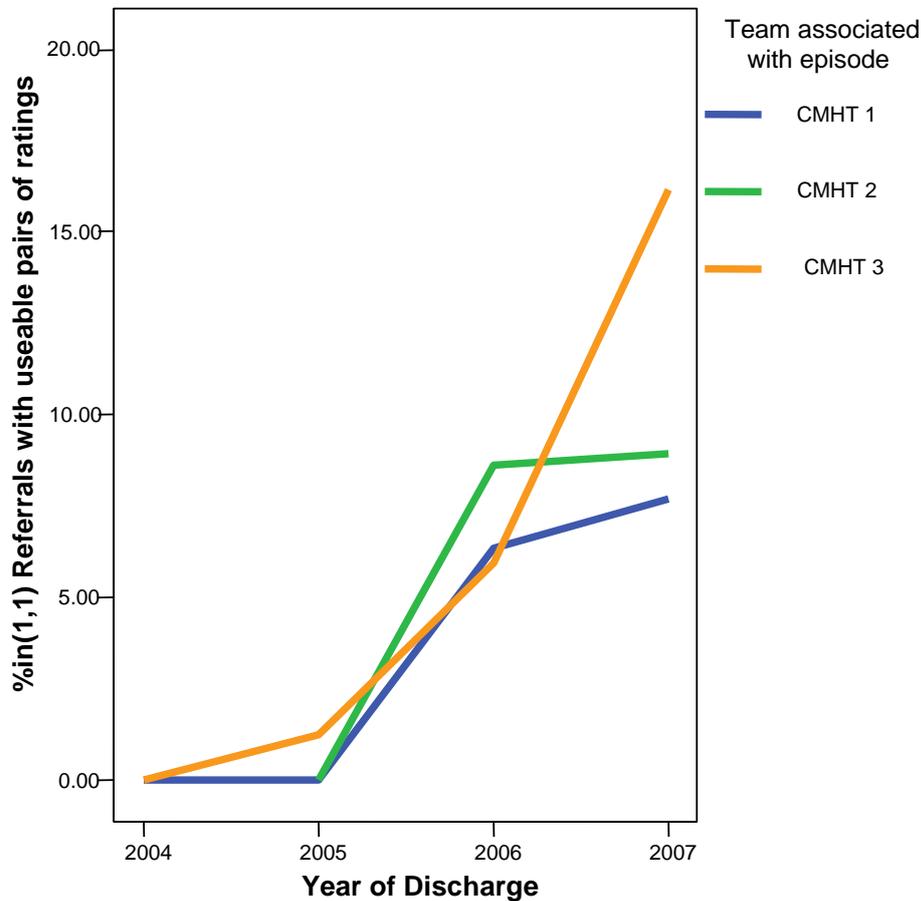
Figure 7: Example of feedback given to clinical teams at a Trust on paired HoNOS ratings over time

Figure 8 shows the aggregate change (for all CMHTs grouped together) in HoNOS ratings between time T1 and T2. This shows a statistically significant improvement in outcomes between the two time periods, on aggregate. However, the HoNOS scores aren't in fact paired, and don't take account of the fact that some may be going up and others down, while some may have a low starting value.
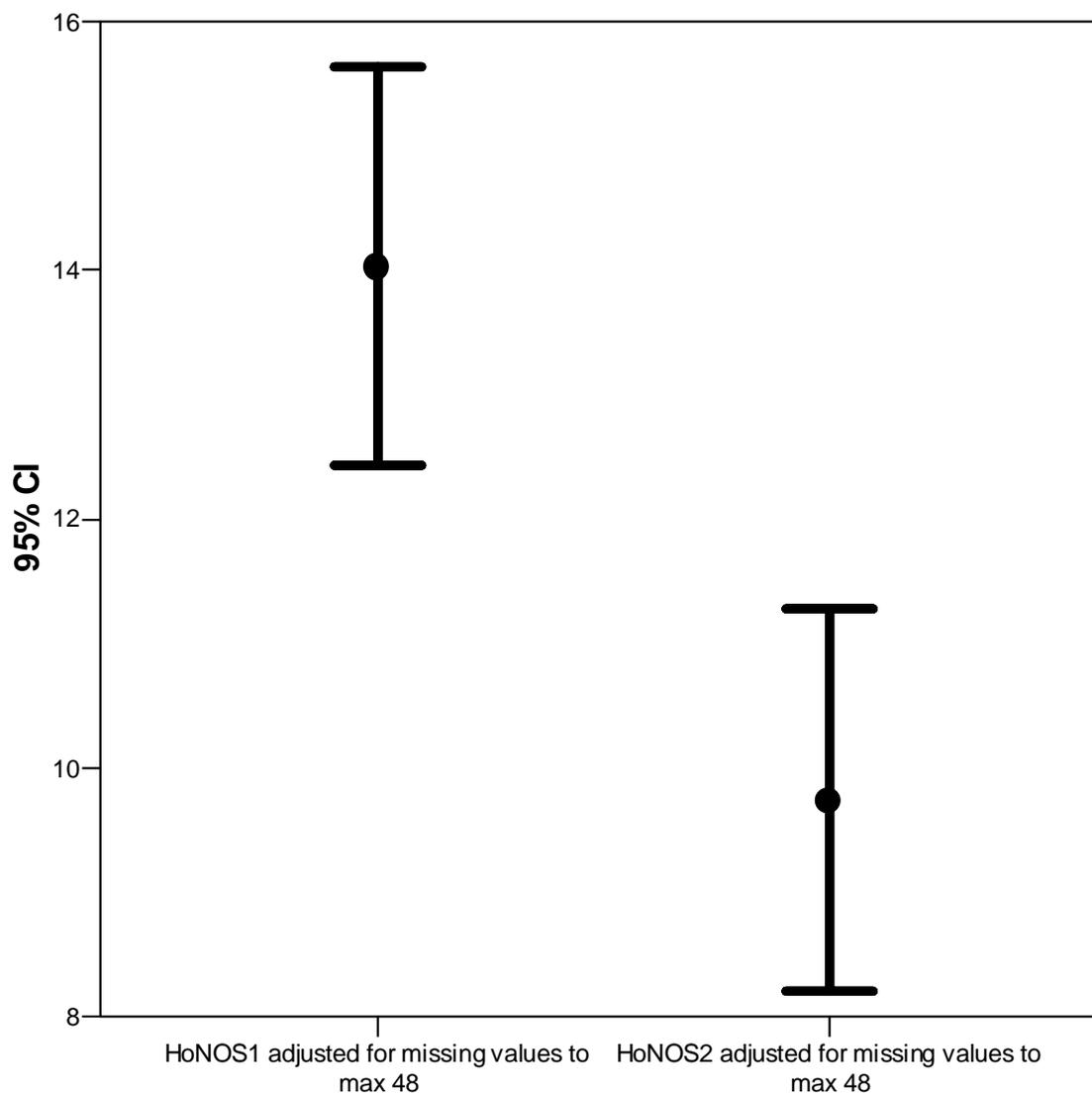
Figure 8: Example of feedback given to clinical teams at a Trust on changes in HoNOS ratings for all teams

Figure 9 shows again the change in HoNOS ratings between time T1 and T2, but shows the results broken down by clinical team. Therefore, in principle, only CMHT 3 shows a statistically significant improvement in outcomes between the two periods. This sort of feedback was considered very valuable to teams who could then examine their HoNOS rating data and why some teams were doing better than others (for example data quality, casemix, admissions policies, and so on).
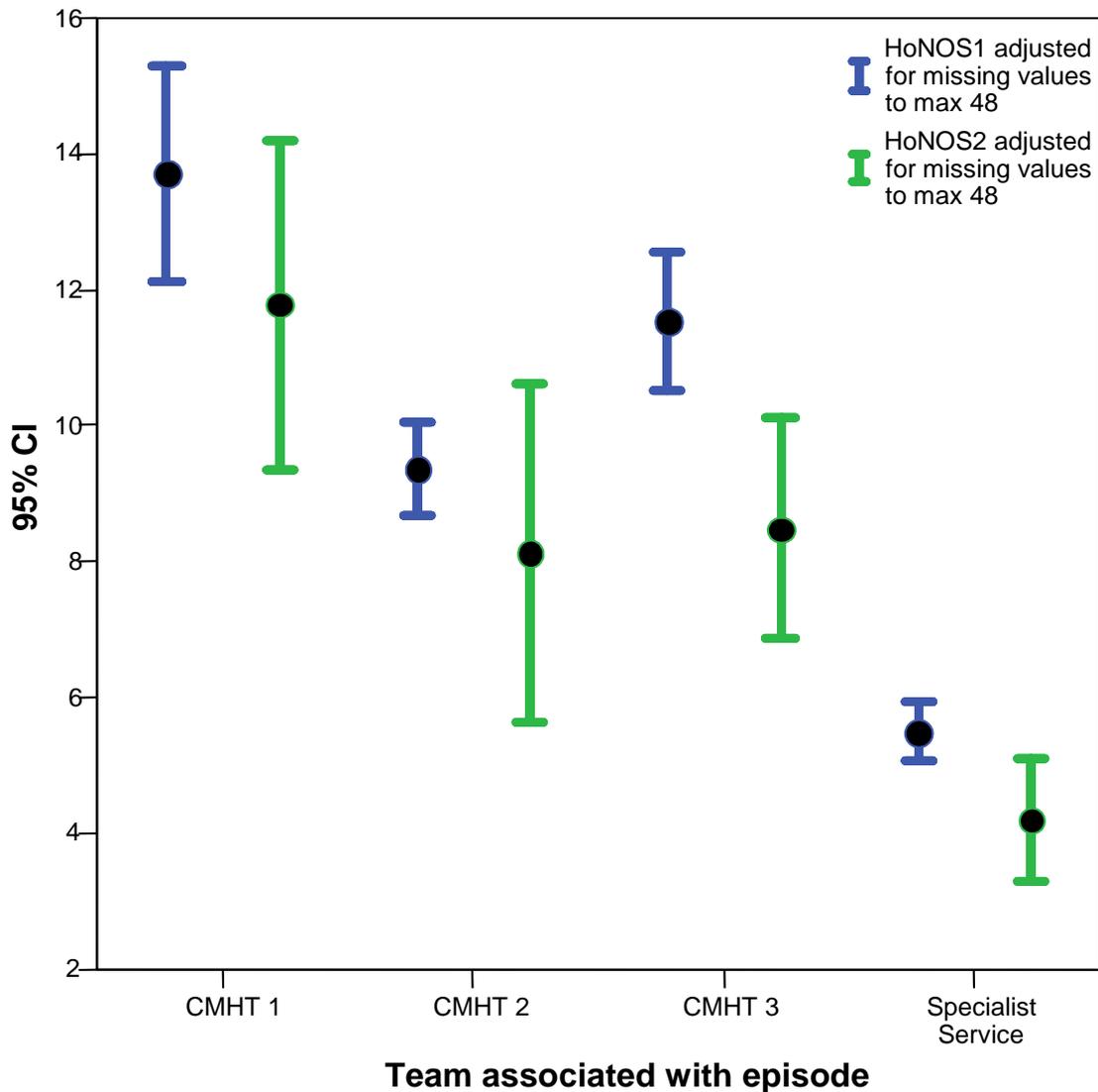
Figure 9: Example of feedback given to clinical teams at a Trust on changes in HoNOS ratings by team

Other Trusts had less developed feedback processes, though there were a number of other examples of the sorts of feedback which might be considered useful. One suggestion was the use of scatterplots of patients, so that outlier HoNOS scores could readily be identified. This was deemed to be clinically useful since it would allow clinicians to examine their practice and explore alternative forms of therapy for those which had outlier scores.

A more sophisticated version of this which was suggested as a useful way to present data was found to be the Reliably Significant Change diagram as shown in Figure 10. Pre- and post-treatment scores are plotted on each of the axes with all patients shown in a scatterplot. A 45-degree line would separate those patients who haven't changed (on the line), from those who have improved (below the line) and those who have got worse (above the line). This was seen as a good way of providing feedback in a simple way. Standard deviation lines could also be plotted on either side of the 45-degree line to show outliers.

"If you have a whole of dots at the top end you're telling clinicians you're not actually making people any better and in fact you're making some people worse. This grabs their attention".

Ideally one would want to impose lines based on a set of normal values for individuals with no mental health problems, thus standardising the data against the national population. This produces a reliably clinically significant change status, because it would show those patients who, for example, improve by more than 2 standard deviations. However, currently no HoNOS exists for people in the general

population with no mental health problems, so this cannot be produced at present. This data exists for CORE-OM, but not for HoNOS.

To plot just the basic diagram with the 45-degree line could be done in principle, but a mechanism would need to be found to generate these plots simply for clinicians so that only before and after scores are added and confidence limits are calculated automatically. There was no evidence of anything remotely this sophisticated being introduced at Trusts to provide feedback to clinicians.
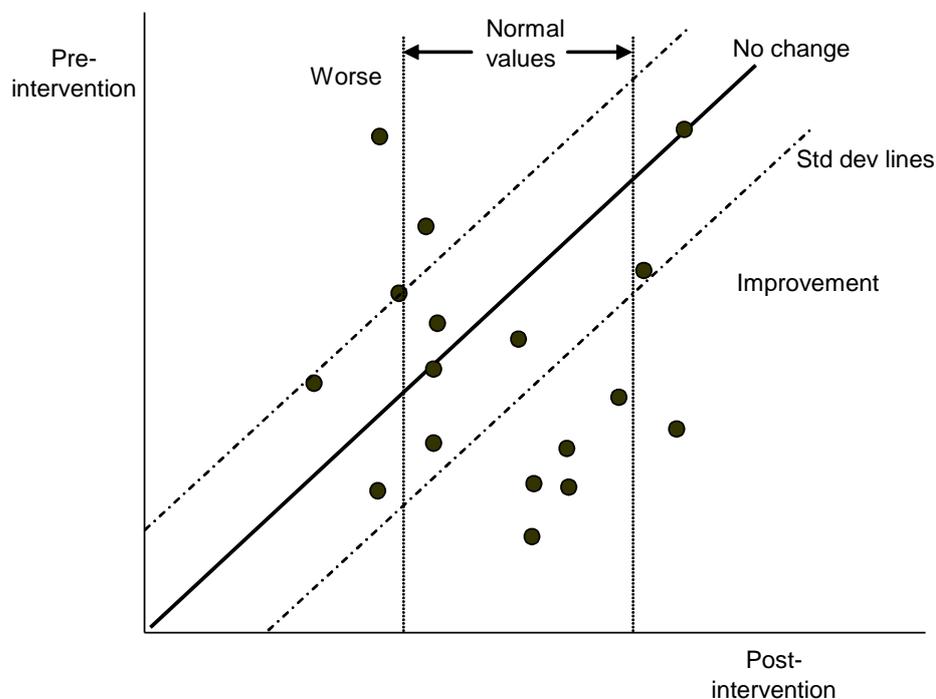
Figure 10: Reliably Significant Change diagram of patient scores

Some scepticism was however voiced about the level of sophistication that would be considered useful:

"If you look at some of the outputs from the DoH at the moment, they're increasingly using funnel plots and decision-control charts, … most of the people here wouldn't have a clue what they're about."

One Trust had capabilities within their electronic system to produce line graphs for clinicians as a HoNOS analyser. It could plot patient scores over time and connect the dots, but nothing more sophisticated than this. If a specific report was wanted, for example the HoNOS ratings for all the male patients which a clinician had seen in the previous year, they would need to phone their IT department who would then generate an individualised report. There was no mechanism in place for such reports to be drawn off the IT system by individual clinicians.

The IT department had found that they were producing over 500 different types of reports as a result, because each clinician would have a slightly different requirement for the way in which they wanted to see the data, with the result that the system was very inflexible. As a result, the IT system was undergoing redevelopment and a new more flexible system with instant feedback was being piloted in clinical teams. Different mechanisms of providing useful clinical feedback were being explored. The pilot scheme was running in two teams (one Community Mental Health Team (CMHT) and one inpatient ward) to make sure that a HoNOS rating was done for every patient admitted to the team and again on discharge so that a T1 and T2 rating would be available for each patient. The plan was that the data would be put onto an Excel database which the teams could then manipulate themselves so that it would be most useful for them in the format which they would want to see it. This could then be examined by age, sex, ethnicity, or diagnosis.

In a service where HoNOS was actively used and deemed extremely valuable, the HoNOS scores were also correlated against other background information from patients, such as age, gender and the team member involved in the patient's care. This enabled the clinical team to, for example, determine which types of patients had better outcomes under the care of certain members of staff. This helped individual staff members to better understand their own strengths and weaknesses, and also helped the team as a whole to identify areas of shortcoming. In response, specific team members could receive training geared to specifically develop the weaker areas identified. All HoNOS ratings were done as part of the clinical team meetings by using a projector to display the HoNOS rating system on the wall which would facilitate a team-based approach to HoNOS ratings.

One challenge in feeding back data on outcomes is that diagnoses in mental health aren't actually recorded routinely either. This means that one cannot readily track outcomes for cohorts of patients with similar diagnoses and examine how they've done.

Another reason why outcome measurement is often not considered clinically useful, is because there is a massive drop in completion rates of outcome scores between time T1 and T2.

"Only about 5% of patients will receive a second rating at time T2. This is because they (clinicians) feel they should be completing them, though they can't see it as useful if they're only completing them at T1."

Quite often, ratings are only done at the beginning of a patient's treatment episode and then never again. During the course of a patient's treatment they may see more than one clinician and one of those clinicians might be enthusiastic about completing HoNOS, while another will not. The lack of consistency contributes to the drop in completion rates at time T2 and hence the lack of clinical usefulness.

One clinical site had developed a clever 'glitch' in their software programme to ensure high HoNOS completion rates at time T2. Patients could not be logged off the system (and hence the caseload reduced) unless HoNOS was completed at time T2. This ensured 80% completion rates. This was however replaced with a Trust-wide system which simply prompts at time T2 whether a HoNOS has been completed, and if not asks for a reason, but any reason can be filled in and there are no repercussions from this. Completion rates dropped to below 70% again.

In terms of feeding back outcome scores to patients, there were mixed views about the value of this exercise. In psychological therapies and with the use of CORE-OM, it was seen as an absolutely key feature of the interaction. Clinicians feed back scores at the beginning, the middle and the end of an intervention and give patients a little booklet with a narrative discussing their scores. This is seen as very powerful for the therapeutic relationship. While CORE-OM was seen as a useful clinical tool when fed back to patients to track progress, this was seen as less the case with HoNOS.

In one clinical team, it was the role of the care co-ordinator to feed back the discharge plan to the patient, and this would sometimes include the HoNOS score of the patient. This was not done routinely however. HoNOS was deemed more often as being for the benefit of the clinician and the team in decision-making. The HoNOS was considered complex to interpret for patients.

"It may not mean much to them to go from a score of 35 to 7, while much of that change may in fact be driven by the fact that they have housing arranged, or other social interventions in place."

11.1.2 Poor information technology systems

Related to the key barrier of clinicians not receiving meaningful feedback on outcome measures, is the fact that mental health providers clearly lack up-to-date IT systems to produce appropriate feedback. There are still huge problems around informatics. Many IT systems crash often and are unreliable and require substantial investment.

"The networks are down around 20-30% of the time. People don't see IT as the solution, rather the problem."

For example, with the slow computing speeds in one Trust, it could take staff up to 45 minutes to load up HoNOS scores and risk ratings and this had implications for how much clinical time was then available.

Many clinical teams still exist almost entirely on paper-based systems. In mental health services generally there is still a large problem to get clinical staff to work online. Not all members of teams have access to IT infrastructure.

"We are anxious about letting go of our old ways. So currently information is recorded in three different ways: in the patient notes, electronically, and then we'll dictate a letter as well."

While the Trusts that were interviewed happened to have the capability to run electronic patient care records, they are the exception to the rule and even then, seem to run parallel systems. Until one could get to the stage where the system was wholly electronic, Trusts ended up running two systems (one electronic and one paper-based) in parallel.

Some clinicians also still have concerns about confidentiality of keeping information on electronic systems.

In addition, many Trusts have clinical sites and teams dispersed geographically and they have each evolved their informatics in a unique way, making it difficult for the management to keep on top of what they do, never mind introduce new electronic systems. Many clinicians still have to carry clinical notes around with them because they commute between sites and their 'office'.

There was a feeling that there was also a poor understanding between informatics teams and clinical teams about what the requirements were of the clinical teams from the IT system and what was deemed important and useful information. It was felt that informatics teams didn't necessarily see the value of getting the most out of the MHMDS with the HoNOS scores as part of it.

"One can't even get ones caseload on the electronic system, so there is no belief that one could get nicely labelled graphs of routine outcome measures."

Because of the poor perceived capability of existing IT systems, for clinicians to get the most clinical value out of outcomes data, often required them to design their own computer system.

One clinician designed an electronic care planning system to implement HoNOS. This was done on an Access Database so that changes in HoNOS scores between time T1 and T2 could be correlated with socio-demographics. This care planning database was then subsequently rolled out across the whole Trust to extend outcome measurement to all units.

In another service, a clinician developed an Excel-based spreadsheet where patients could be tracked over time. Again this was produced on their own initiative and there was no Trust support for this.

In another Trust, several new clinicians started developing their own databases for outcome measurement, again because of a lack of capability initially from the IT system. Each of their databases ran independently and data was not available more widely than to their own team.

These clinicians developed considerable expertise in setting up their own systems, for example using SPSS, which proved very effective. A workshop was then held between these clinicians and the Trust IT managers so that they could develop and co-ordinate a system that could work at the individual patient level, at the clinician and team level, and the management and strategic level. The Trust IT department then developed a system which could draw out data in excel spreadsheets from different teams into different formats as required for different purposes. Standard reports would be available for senior managers and the Trust board to examine data in more aggregate form, but for clinical teams the data could be drawn off into excel spreadsheets. These then needed to be manipulated by individual clinicians or psychologists to interrogate the data in the way they wished and required them to have the IT skills to do this in Excel.

While this development of the IT system is clearly an important step forward for routine outcome measurement, there is of course the concern that many clinical staff lack the necessary IT skills to effectively interact with informatics systems.

"There is a gross underestimation of IT skills in the NHS."

For example, despite one Trust having the capability to do HoNOS ratings directly onto their computerised system, many clinicians were still doing it on paper and then often getting secretarial staff to transcribe the data onto the electronic media.

Most clinicians also used paper versions to complete CORE-OM. This would then be transcribed into an electronic system after the fact. The Trusts that don't use CORE-PC, require IT support where the electronic versions of the CORE-OM have been linked to the electronic patient record. There is also no facility for patients to enter CORE-OM directly onto computer. In one of the Trust directorates, facilities were set up to scan CORE-OM directly from the paper-based version into electronic format. In another Trust which mandated CORE-OM, these were also completed on paper and then transcribed to the electronic system. The convention was that medical staff could get secretarial support to load these up whilst non-medical staff would have to do this themselves.

One suggestion related to the inputting and management of electronic data, was to see administrative staff more involved as information managers. The secretarial staff who for example log patients onto the Trust electronic database and manage referrals, could be involved in ensuring teams get regular reports on their outcomes and ensure the feedback is provided in a manner which is useful to the team. Instead of just inputting HoNOS scores, they could use the electronic database to inform teams about how many referrals they had, what HoNOS scores they had, what diagnoses the patients had, and so on. They could therefore provide regular governance reports to clinical teams. So in addition to the role of clinicians changing in that they also input data electronically and not just hand the job to secretarial staff, there would be a need to re-examine the role of administrative support in the NHS.

Another suggestion to make the IT systems more clinician-friendly was to make the electronic media look more like a set of patient notes so it would be intuitive to open and tab down to the different sections.

"The present system … is not that well integrated. You have to push lots of different buttons to get to the HoNOS rating. Clinicians are fundamentally resistant to change, so the way the system looks and the way one can interact with it is critically important."

Ultimately, both the upgrade of IT systems and the development of IT skills amongst clinical staff would require substantial investments which did not seem to be a high priority by many interviewees. Some Trusts found themselves in a position of a weak local health economy with overspent PCTs and thus a retracting budget. Therefore in a political milieu where Trusts have to cut back on services, investment on computers and IT systems was not seen as appropriate or advisable.

11.1.3 A change in culture needed

One of the words that emerged often when asking about barriers to outcome measurement was 'culture'. The main challenge seems to be to establish an outcome-oriented clinician culture.

One Trust was encouraging all their clinicians to publish their outcome measures on the website. But there was a lot of resistance, because:

"… people are afraid of others viewing their outcome measures – you want a culture where people aren't afraid. We're a long way off that."

It was suggested that peer pressure can be an effective means of increasing uptake when one for example sends around an email saying that clinician X has completed so many HoNOS scores while clinician Y has done far fewer. This does focus clinicians' competitive interests. In one Trust, all clinical teams were mandated to complete a governance report every year. They were given a template to do this and one of the questions was about whether they are completing HoNOS or not.

This governance report is then printed on the internet and becomes public information. This becomes a very powerful catalyst for change.

"When we make it mandatory for all clinicians to publish their outcome measures on the web, then we will see the ratings being used more consistently and more widely."

Also incorporating outcome measurement into the consultant appraisal has been seen as a useful way to get it focused on clinician's minds. However, all of these suggestions require a culture of trust and openness, otherwise fearfulness could lead to sabotaging of outcome measurement.

Many clinicians had concerns over outcome measurement being used for performance management. The fear was that outcome measures would be used to distinguish good clinicians from poor ones. Any kind of league-tabling approach will make clinicians very afraid.

"Using outcome measures for constructive purposes rather than for finger-pointing will be very very important."

Part of the fear is driven by the fact that other changes might be more readily attributed to the system and the system not functioning properly, thereby dissipating responsibility. But outcome measures were directly attributed to individuals and teams. Questions could then be asked about why team A hadn't seen the same improvements in outcomes at time T2 compared to team B.

"The culture has to be a data-oriented culture as well. People need to become willing to start looking at data and looking at impact and see how well they're doing."

Getting clinicians to use outcome measures is not seen as part of their practice, it is not something they've grown up with or have been trained in, so a main barrier is trying to get them to change their practice. Getting it into the culture of their practice is a key challenge. There is a huge cultural shift required.

"Teams aren't used to doing it."

"The challenge is a 'hearts and minds' exercise."

Clinicians inherently enter the profession because they want to be effective in terms of improving health outcomes. Incentives that are geared towards helping them to be more effective, rather than making them fearful that they aren't effective enough, would be seen as far more effective. If outcomes could help clinicians to see what treatments were most effective for which types of patients, this would be a strong incentive.

11.1.4 Time constraint

Interestingly, most clinicians did not see the time element as a major barrier to outcome measurement. It was felt that HoNOS was for the most part quick and easy to complete and should take at the very most 5 minutes additional.

"It doesn't actually add a huge burden to the clinical workload."

The CPA involves quite a lot of paperwork anyway, so adding these numbers on computer was not seen as hugely onerous.

"Once you get used to HoNOS it only takes a couple of minutes. Everybody thought it was worth doing. Once you've built into what you do, it's a doddle."

11.1.5 Training needed

Unlike the Australian case study, training did also not emerge as a major barrier.

Some felt that staff could readily be trained on HoNOS through the 'train the trainer' courses from the CRU. If each Trust had 5-10 individuals trained as trainers, there would be no reason why teams couldn't get the appropriate training in the use of HoNOS.

However, the cost of training was seen as an issue. Many staff using HoNOS hadn't been properly trained. While clinicians may not have seen training as a major barrier, inadequate training is of course a serious issue in terms of the way in which the instrument is used and the reliability of the ratings. In terms of staff learning about HoNOS, it was often the case that they may have had a brief chat with a colleague, or picked up a few pointers from someone, but lacked formal training.

CRU's perception of the level of interest in HoNOS, as gauged by the uptake of HoNOS training, was that the demand had been very little in the last 2 years. However, CRU's perception was also that this had been picking up in the last 3 months.

In terms of training requirements for CORE-OM, a few issues emerged around training users on this tool. While these didn't emerge as major barriers to CORE-OM, it was felt that for patients with language or literacy problems, it may be necessary for staff to talk users through the instrument before completion. This would however have repercussions for the psychometric properties. It was felt that support and advice might be needed for service users for self-completion. There may also be ways of making outcome measures more visibly accessible for patients with learning disabilities. If however existing scales were adapted, it would again affect the psychometric properties of the instrument.

11.1.6 Lack of interest from patients

There was one view that patients, once treated and better are not interested in completing self-rating questionnaires and this could prove a barrier. This view was not however a universal concern amongst respondents. This may however affect the ability to collect time T2 data and hence the value of the outcome measure.

11.2 Incentives for routine outcome measurement

In terms of incentives that might drive the implementation of routine outcome measurement, there were a number of different potential drivers. When these incentives emanated from clinicians themselves, they tended to be more successful in sustaining such initiatives. Whereas when the drivers came from management or external factors such as applications for Foundation Trust status, they were less successful, because once again no clinical benefits could be seen.

11.2.1 Clinical champions needed

It was seen by many that clinical champions enthusiastic about outcome measurement, were absolutely essential. Clinical champions at a reasonably senior level could drive forward outcome measurement, and this would result in a better uptake than a top-down push from management. Good practice examples help to show other clinicians what can be achieved and the value of outcome measurement. In this way outcome measurement is driven from a grassroots perspective. Often when such clinical champions left an organisation or a team, the collection of outcomes data ceased.

One Trust with a large uptake in CORE-OM had a particularly strong clinical champion in the area of clinical psychology and the introduction of CORE-OM was attributed to this person X, as the X-factor. They were absolutely paramount in getting widespread use of the measure. It was not because the Trust had mandated it, it was because X had encouraged it.

"Finding enthusiasts within the organisation and giving them support…"

could be one way to encourage uptake. There have to be demonstration sites and pilot sites and champions which have credibility to clinicians.

While one Trust had appointed a clinical lead on outcomes measurement who received additional funding for this role, they were not perceived to be a champion or innovator and were seen to simply go around the Trust telling clinicians to do outcome measurement – this was not perceived as useful.

On the other hand, another Trust appointed a clinical lead to an outcomes leadership role and they were primarily responsible for feeding back outcomes data to teams and garnering enthusiasm for outcome measurement. Since this individual had a personal interest in outcomes measurement, they had success at obtaining enthusiasm from clinical teams. In addition to being a clinical champion, management support was also a key factor.

11.2.2 Management drivers

A strong relationship between a clinician who is enthusiastic about outcome measurement and a team manager is absolutely essential. One clinician's view, who had been trying to get outcome measures implemented for some time but did not receive any management support initially, was that:

"However much you want something on the ground, unless you have support from the exec, its very difficult to make it happen."

Outcome measurement was before just seen by the senior management as a particular clinician with a bee in their bonnet, until their own priorities changed on the matter.

Senior management teams or Trust boards are faced with a huge raft of 'must-do's' and if outcome measurement is not on that 'must-do' list, it will not receive management support and it will not get done. It needs to become a management priority and be actively performance managed.

The down-side of this approach is that it will be driving outcome measurement for all the wrong reasons. The Trust Board and management will not push outcome measurement because they think it is clinically useful, they will push it because they feel they have to.

External drivers such as ticking the boxes to show that appropriate governance arrangements are in place, will often drive senior management teams. Their interest will be primarily in the aggregate data, to either pull out poor performing services, or to examine at an aggregate level what the value for money is that is being produced by the service as a whole. They will not generally be interested in clinician or patient feedback.

From the clinician's point of view, outcome measurement for these purposes is then seen as a form-filling waste of time.

In cases where outcome measurement is driven by management in a top-down way, HoNOS completion in CPA meetings is done by the lowest ranked member of the clinical team who will often not know the patient very well. The recording is then sporadic and it is not guaranteed to end up on the electronic patient information system. Compliance is therefore poor.

One of the Trusts interviewed had mandated the use of both HoNOS and CORE-OM for the past 2 years. HoNOS is mandated for patients on enhanced CPA in acute care and CORE-OM for patients in psychological therapies. While this had been mandated by the Trust management, it was not actually happening in practice.

Ironically, there was still some disagreement as to whether HoNOS is indeed mandatory or not within the MHMDS. Some Trusts had apparently sought clarification from the Department of Health and had been told that it is not (and presumably used this as a reasoning to not push for its implementation).

Another Trust has made it compulsory to collect outcome measures routinely, but they didn't specify which measure it should be. Standards were set but not always adhered to.

Command and control as a strategy to drive forward routine outcome measurement was therefore not always very successful. On the other hand, clinicians (clinical champions) trying to implement them without management support, also struggled to make the process work.

A successful initiative undertaken by one Trust management, was to run a workshop on outcome measurement to keep it current in clinicians' minds and keep the momentum going, and to re-iterate the value of doing outcome measurement. This was in particular because clinicians were very against using HoNOS and the conference got people who were using it successfully in other Trusts to

showcase what they were doing (clinical champions). This was seen as very useful. The uptake of HoNOS at time T1 and T2 in the 6 months after the conference increased from around 30% to around 80% so it had a huge impact.

11.2.3 The Healthcare Commission

A key theme which emerged is that routine outcome measurement is unlikely to be viewed as a mandatory activity, until it becomes a biting target in the Healthcare Commission's performance management regime.
CORE-OM is already quite widely used in psychological services, while HoNOS is slightly less well used, but the push for HoNOS is now coming from the Developmental Standards being set by the Healthcare Commission. The wording from the Healthcare Commission was considered by some to be quite "crafty" in that they state that Trusts should use objective measures to determine clinical effectiveness, though they do not specifically state that outcome measures have to be used routinely, or indeed, which instruments should be used. But it was viewed that realistically, the only way one can show at an aggregate level clinical effectiveness, is to use outcome measures or a battery of outcome measures. Therefore it is seen that in a veiled way the Healthcare Commission is expecting to see routine outcome measurement.

Many Trusts were interpreting the Healthcare Commission Developmental Standards in this way. The general view was that a poor set of ratings would contribute to a poor annual Health Check by the Healthcare Commission and therefore a reduced ability to apply for Foundation Trust status. Hence, the lack of using outcome measures could have serious repercussions.

"I think there will be trouble on the horizon. The Healthcare Commission will penalise Trusts through the annual health check if they don't have outcome measures."

To date there has been no penalty from the Healthcare Commission for missing data in the HoNOS returns in the MHMDS. But it is believed that once the Developmental Standards kick in, there will be penalties for not completing HoNOS fully. There is a perception that if they want to get outcome measurement done routinely, then the Healthcare Commission would probably have to get more 'heavy-handed' about it. These data would however only be useful in aggregate form, and could not be relied upon at patient level, because they might be completed under 'duress'. They should not be used for assessing individual patient progress. Trying to explain the individual patient-level variation at this stage was seen by some as too premature. Only aggregate variation could be examined.

If the carrot and stick approach were taken and Trusts were actively performance managed, the Healthcare Commission could incentivise uptake of HoNOS by including coverage in the MHMDS as a performance measure. This is currently the case for coverage on the ethnicity completion within the MHMDS.

Again, while a Healthcare Commission target would have implications for the Trust board and the executives, it would not be of interest to clinicians and would be seen as a form-filling burden, of no clinical relevance.

11.2.4 Foundation Trust status

Related to Healthcare Commission targets being a driver for increasing the uptake of routine outcome measurement from Trust management, the application for Foundation Trust status on the back of the Healthcare Commission annual health check, was seen as another key driver.

On the back of making an FT application, many Trust boards were producing directives on outcome measures. Some Trust boards were working towards finding a set of outcome measures, and other clinical governance and quality of care indicators, which they would support. They were considering mandating from teams a set of outcome measures and indicators which would be both clinically useful and would provide managers with useful information. This would allow for greater consistency across teams and would force them to become more 'business-like' in gearing up for FT status.

Trusts felt they had a strong incentive to show added value, effectiveness and quality. Being able to show which patients were benefiting from which services was seen as a key driver in the FT

application and the more market-oriented approach that was being pursued by Trusts. The Trust board at one Trust was not seen as terribly interested in outcome measurement before, but with the FT application, outcome measurement had suddenly become a priority.

There was speculation by one respondent that a letter had been sent to the Chief Executive of the Trust mandating them to use HoNOS in the context of their FT application. This was not able to be verified however. But there was no doubt a feeling from all respondents that application for FT status was a strong driver for uptake.

However, once FT status had been achieved it was not clear, what the arrangements were in terms of continuing to collect outcomes. Some interviewees felt that Trust boards had taken their foot off the pedal with continuing to push for outcome measurement implementation post-FT status. If anything, the FT status had weakened the incentives for routine outcome measurement.

11.2.5 Commissioners

While commissioners said that they were becoming more proactive in encouraging providers to use HoNOS and that it was high on their agenda, this was not the impression obtained from the providers. The drive for routine outcome measurement did not seem to be very strong from commissioners.

Key performance indicators requested by commissioners from providers did not include any outcome measures. Commissioners typically asked for data on readmissions, length of stay, efficiency indicators, data on physical health, and suicide rates, but not outcomes. They also didn't typically request data on patient satisfaction measures from providers.

Outcome measurement seems to becoming more of a priority and higher up on the agenda of commissioners, but at present many providers still struggle to supply commissioners with basic information about activity. Mental health providers on the whole seem to struggle with information and this can make for less effective commissioning.

There is however a hope that commissioners will, in future, drive the agenda for embedding routine outcome measurement into practice. There is a perception that commissioning is becoming more sophisticated and will therefore have an impact in the future on the use of outcome measures in practice.

Some commissioners are starting to look at building outcomes into their contracts with providers, but at present this seems to be a largely exploratory exercise. There seems to be a need to co-ordinate some of this at StHA level.

As part of the Local Delivery Plan reporting between Trusts and PCTs, providers have to supply a number of data items, for example numbers of patients receiving certain types of treatments (e.g. crises resolution or early intervention). The measures are largely process focused rather than outcomes focused. PCTs in turn submit that data to StHAs because they are also performance managed on these data. The focus for commissioners is therefore currently on these measures rather than outcomes.

There is a serious concern that commissioners wash their hands of mental health services because they can more readily see what value for money they're getting in other areas of health care so they will slowly divest themselves from mental health. Mental health commissioning has definitely received less of a priority than other areas of commissioning and commissioning for outcome measurement has definitely not been a priority.

11.2.6 Payment by Results

There were mixed views about whether Payment by Results (PbR) would be a potential incentive to drive the introduction of routine outcomes in mental health in the future. Some felt that it would incentivise only managers, but not clinicians.

"PbR would exercise the minds of the exec, but not the clinicians."

However, if outcomes were attached to the tariff, all teams would do them and completion rates would be high. If it were indeed possible to construct an appropriate tariff around improved outcomes in mental health, and HoNOS was one way of measuring that ...."I'm sure we'd see a massive uptake". There was however a question around whether this was the right incentive. It would get it done, but it may not necessarily deliver service improvement or good quality data. One would need to ensure that scores were not just being 'made up'. Some felt that if PbR were a route to the widespread introduction of outcomes, they would "surf this wave, however evil it might be".

Many felt they would prefer to see outcome measures introduced independently ahead of any PbR initiative, though they would want to ensure that the suite of measures they introduce were compatible with any PbR scheme. Others felt that using financial incentives such as PbR to achieve certain aims, such as the collection of outcomes, erodes what is left of the public ethos. If one started to incentivise people where they previously had the 'higher moral ground', it was seen as not very helpful. PbR could be used as another stick whereby if teams cannot show their effectiveness, they will cease to exist. But the general view was that this was the wrong way to go about it.

Some felt there would be a huge potential for perverse incentives with a clinician-rated measure like HoNOS. It would need to be run alongside an institutional audit scheme which could be very expensive. Attaching financial incentives to achieving better outcomes through something like PbR would run a massive danger of distorting the outcome measurement process. Other unforeseen problems would be created by making payment contingent on outcome.

The DoH point of view was that ultimately PbR cannot relate only to activity, it would need to take account of outcomes too. For commissioners, it was also seen as ultimately being paramount, because they should be purchasing improvements in health outcomes. However, linking the emerging practice of outcome measurement too early to a performance management system like PbR, could be dangerous. The first step is to win hearts and minds of clinicians that routine outcome measurement is a vehicle for service improvement. It should therefore rather be driven from a clinical perspective. If it is imposed on clinical teams and it doesn't quite fit with what they do, it cannot work.

Many felt there is a lack of clarity at present as to what PbR would look like in mental health and therefore how outcome measurement might be part of that process in terms of future commissioning arrangements. It was felt to still be quite a long way off for mental health.

Interestingly, a variant of HoNOS has played a key role in the development of PbR in mental health which is at a critical stage [113]. One of the key concerns is that Healthcare Resource Groups (HRGs) don't work in mental health. Mental health patients don't fit into the neatly defined biological categories that might be found in acute services.

Building on earlier work which tested the early development of both HRGs and HoNOS [114, 115], a modified version of HoNOS was used (called HoNOS Plus) in six pilot Trust sites in the Yorkshire area (Leeds; South West Yorkshire; Doncaster and South Humber; Humber; Tees, Esk and Wear Valleys; and Northumberland, Tyne and Wear) to explore the introduction of PbR in mental health and to come up with a number of activity categories [116]. PbR is being piloted around 13 care pathways. These don't have a full costing model attached to them as yet. HoNOS has been adapted as a needs assessment tool to HoNOS Plus to help determine which pathway or care package patients will receive. So it is used to determine when something is deemed of clinical significance (need). It seems to have helped determine the most workable set of care pathways which can be costed for PbR.

HoNOS Plus uses a similar scale to HoNOS but includes historical information and some additional items on engagement, child protection and vulnerability to help assess needs. HoNOS Plus was never tested though for reliability and validity. This is not an official version of HoNOS and since it was never designed as an assessment tool, this has been considered problematic. HoNOS Plus cannot be used as an outcome measure because the collection status has been changed.

11.3 The choice of the instrument

In terms of the choice of instrument, there were very mixed views when it came to HoNOS. Clinicians were either told to use it and did not choose it, or were otherwise very enthusiastic about it as an instrument.

CORE-OM on the other hand, was generally regarded as clinically useful and therefore had wide coverage in many services, though it was seen as not appropriate for patients with psychosis. CORE-OM was seen as a more sensitive measure to pick up changes, though some clinicians also found it to general and gross. Since the drive to use CORE-OM had come directly from clinicians, at a grassroots level, it tended to be a popular instrument and used voluntarily. Psychological therapies tended to be very comfortable with outcome measurement generally and various rating tools are a routine part of practice. As a result, a tool like CORE-OM is far more accepted.

An interesting point that was raised was that the growth in specialist services in mental health was to some extent at odds with the drive to use quite broad non-specific measures like HoNOS or CORE-OM which were deemed to be too general and sometimes not sensitive enough to detecting change.

The fact that HoNOS has been promoted by the DoH has not helped give it a good name. Even though it was developed by the Royal College Research Unit (CRU), it has been sponsored and endorsed by the DoH and that makes many people reluctant to use it.

HoNOS is rooted very much in a medical model of mental health problems. There is seen to be a big discontinuity with HoNOS between what users perceive to be important aspects of their care and what clinicians perceive to be important. HoNOS therefore potentially under-represents the user's perspective.

People seemed to have mixed views about HoNOS as to whether it is a sensitive measure or not. Many felt that HoNOS is quite a blunt or gross measure and didn't like the calibration. Others felt HoNOS picks up the patient's condition at an acute phase as well as when they are well again and discharged and is reasonably sensitive. However, it was not seen as sensitive in rating patients with anxiety or Obsessive Compulsive Disorder (OCD) because there is only one question covering this, and in some services this is the bulk of the patient population. HoNOS was seen by some as not having any clinical purpose at the individual patient level, rather it is seen as a strategic tool with benefits only at aggregate level.

Others felt HoNOS is the best validated, tested and socially relevant outcome measure out there. It is broad and covers a lot of questions on social circumstances such as accommodation.

"For us the main outcome measure is social inclusion."

Often the aim of a treatment may not be that of curing patients of their mental illness. Instead the goal is often to get individuals included in society at whatever level may be appropriate for them. The outcome will therefore depend on the patient's individual situation and what outcome would be deemed good for them. Often just maintaining them may be considered a good outcome. HoNOS was seen to be helpful in planning social interventions accordingly (e.g. accessing social care packages) and when patients returned (e.g. the patient had a relapse), the team caring for them would have a good idea of what outcomes were appropriate for the patient because they had rated the patient over time.

Some of the HoNOS questions on social functioning, were however seen as problematic. In one team, the last 2 questions (on living conditions and accommodation) were amended because the questions were not seen as relevant. Essentially they were completing the 10-item version of HoNOS, but entering it as a 12-item instrument in their data. This would be problematic for comparability and validity of the instrument if such amendments are being made to the questions routinely. These questions were felt not to pick up patient's mental health so much as society's attitudes to people with mental health problems.

Ultimately though, each clinician would have a different view on the instruments they liked and the ones they didn't like and no solution would ever be reached. It was argued by some that since HoNOS is a reasonably well validated instrument, "I don't see why we shouldn't draw a line in the sand, and get on with it…". One can never satisfy everybody, so provided the instrument is implemented properly and they have sufficient training, this should suffice.

In one Trust, while HoNOS was mandatory, clinical teams could explore what other measures they would like to use in addition to HoNOS. While there had been some unhappiness about the choice of

HoNOS, it was "not up for debate. You just have to use it." In addition, they could use any instrument they liked – these could be loaded up onto the Trust electronic database and used alongside.

The psychometric properties of various instruments were deemed important to some clinicians but not all. One view was that clinicians are not that aware of the psychometric properties of instruments. While clinicians may know that such concepts are important they may not be familiar with how to get such information on a tool or how to make a judgement about it.

Quite often clinicians would rather develop their own questionnaire or instrument specific to their own needs, which is totally un-validated, rather than use a ready made tool off-the-shelf which has been tested. Alternatively, they might just download a tool off the internet and try it out as a once off. They would therefore often rather reinvent the wheel, and use tools which would be unverifiable, and potentially of poor validity and reliability.

There is therefore not enough information available as to what instrument to use under what circumstances, what the instruments actually measure and which ones have strong psychometric properties. As a result, the Department of Health and Care Services Improvement Partnership (CSIP) are producing a Compendium of outcome measures which will be able to inform clinicians as to the measures available, their properties, and the practicalities of each tool, such as how time-consuming they will be (see section 11.4.2).

In terms of who did HoNOS ratings and where the information was gathered from, the majority of ratings were done by the key worker who would be involved with the patient at admission and discharge. In other circumstance, all staff would do HoNOS ratings as a team. In some cases it would make sense to do ratings as a team, but in others (most likely in the community) it would be impractical. In another service the HoNOS rating was also done in co-ordination between the lead clinician and the care co-ordinator for patients on enhanced CPA. If it was just for an assessment, only the clinician would do it.

In terms of the information, this would be drawn from a wide variety of sources, including family and other staff, visiting patients in their homes, and speaking to their carers. The time period used would also sometimes vary from the previous two weeks, to 3 days in the case of acute care, since to do ratings at two distinct non-overlapping intervals, two weeks would be too long. This convention is similar to that used in Australia. The accuracy and validity may however be questionable because there doesn't seem to be a standard way of doing the rating across services.

11.3.1 Clinician–rated versus self-rated instruments

Clinicians seemed to have mixed views on whether a clinician-rated or self-rated instrument was better. However policymakers and academics tended to view self-rated instruments far more highly, since the patients should in essence be judging their own outcomes. However a number of issues emerged.

Self-rated measures place less of a burden on clinician time, but may result in lower completion rates due to literacy, language, and cultural barriers. Completion rates for self-reports are likely to be considerably lower in mental health. Users should also receive training on how to complete the questionnaires, but there is currently no framework for that to happen and investment in that does not seem to be seen as a high priority.

While there are issues around HoNOS as well, as a clinician-report, in terms of its cultural sensitivity and appropriateness, it is argued that the person doing the rating should be able to make an appropriate judgement about the cultural issues to make an appropriate rating.

Self-rating instruments like CORE-OM were seen especially by staff in psychological services to be easily completable by the majority of their patients who are less severe and tend to be treated in the community. They felt user-rated measures gave patients a way of seeing their own improvement over time and it seemed more valid. It also gets patients more engaged more easily.

However for some patients, they would rate themselves as desperately poor when they weren't actually and vice versa, so self-rating instruments have limitations for some client groups. There may

be particular problems with self-report instruments for patients with personality disorders, for example, often only extreme dimensions will be circled on the questionnaire.

There was a perception that if clinician-rated tools were used, they would need to be audited. However, it is not clear how this might work for HoNOS if on occasion it is completed by a group or team rather than an individual clinician, or indeed how the rating can 'objectively' be verified.

11.3.2 EQ-5D

Views around running a generic measure like EQ-5D alongside other outcome measures had a mixed response. Many felt it would depend on the face validity of the instrument and its perceived usefulness. While the focus in mental health is very much on improving the quality of life with people with long-term conditions, clinicians would want a quality of life measure to pick up aspects of social exclusion and whether or not individuals can participate fully in their community. The fact that EQ-5D was a short instrument was considered a bonus. It would negate many concerns over having a time-consuming complex instrument.

There was however a concern that EQ-5D would not be appropriate in mental health services. It was felt that it would be inappropriate at the extremes of the age distribution, in elderly and CAMHS services, where interventions are geared not only towards the child or older person, but also towards parents, families or carers.

There was a feeling that from a practical perspective one would more likely be able to introduce an EQ-5D than an SF36 measure. The VAS on the EQ-5D might be particularly useful in some circumstances in mental health.

Concerns were expressed about EQ-5D in terms of its ability to measure scenarios where not improving patient's health status was considered a good outcome. Where patients are socially excluded or find it difficult to sustain relationships, the health service role may not always be to improve their mental state.

One view was that it had taken many years to implement clinician-rated measures like HoNOS and try to encourage their uptake, and so it could not be seen how EQ-5D would do any better. In terms of the priorities for mental health services, there was an apprehension about how low down the scale of priorities this might fall. It could not be seen that it would valuable, for example, to commissioners. They would rather see a full set of HoNOS scores. It was also questionable what benefit clinicians would get from it.

Introducing something like EQ-5D is seen as a very long way off being a viable consideration in mental health services.

11.3.3 Other performance data which is collected

The main driver of performance data collection in general is without doubt the Healthcare Commission's annual health check. However this data is collected largely at aggregate level or is at least aggregated up to Trust level (e.g. the annual surveys of patient satisfaction). Clinicians are in fact often unaware of the extent to which the Healthcare Commission monitors each Trust and the amount of data which is produced at aggregate. The pressure is keenly felt however by the Trust board rather than the clinicians, highlighting the disconnect between the two groups who clearly face different sets of incentives.

In some clinical teams in Trusts they had exit questionnaires where they asked patients about their satisfaction with the service and these snapshots were taken every few months. In addition, similar to the survey results by Gilbody [9], other data like readmission rates, suicide rates, length of stay, patients under the Mental Health Act and other performance indicators are also collected on the Trust electronic database and can be fed back to individual clinical teams. These are all collected routinely and used in clinical audit. The Developmental Standards being set by the Healthcare Commission have proposed a whole set of performance indicators, particularly around CPA. In one Trust, it was decided that these indicators will be used as the basis for what teams will collect and feedback on

routinely because they were thought to be a sensible set of measures and to reflect accurately the quality of the care provided.

The national patient satisfaction survey which is collected annually by the Healthcare Commission, may in future have additional aspects tacked onto it (rather than running another separate survey). This will include questions on quality of life. Since quality of life may not be attributable to the service provider, it cannot be used in a performance management regime, but it could be useful more generally for people planning or delivering mental health services.

11.4 The policy context for outcome measurement

11.4.1 The past policy initiatives – the Fonagy report

Many of the policymakers and clinicians interviewed, participated in the expert reference group set up by the DoH on outcome measurement and had diverse views on the process, the instruments and the conclusions drawn.

Given the diverse set of interests involved, Fonagy was selected as an independent 'broker' to chair the group. Within the reference group various parties were very concerned with which measures were going to be chosen and there was much party-politics involved. Many of the experts who participated in the expert panel disagreed widely about the choice of instruments. There was a strong debate at the time between two measures: FACE, which the Department of Health had given a lot of support to, and HoNOS, which had been supported by CRU. There were also financial interests on both sides, e.g. FACE was backed by a commercial company that was marketing a computerised evaluation system and CRU was marketing a training programme for HoNOS. While these interests were not made explicit, many people who backed each of these instruments had invested some significant resources in setting up each of these tools. Each party therefore wished to see their investment repaid in the choice which the UK took in terms of a tool to mandate for broad use in the service.

The DoH view was that mandating the collection of outcomes and the choice of instrument (i.e. HoNOS), would make it happen. A suite of measures were recommended but it was felt by many of the other experts that this was too ambitious - even getting clinicians to complete just one measure would already be considered a good outcome.

Despite considerable investment from the DoH on the pilot projects, and pressure on Trusts to start using outcomes, the uptake was poor and the barriers were considerable and so the findings in the report were quite pessimistic. The report was therefore not promoted strongly by the DoH. In fact, many people in the service never got to hear about it.

The recommendation that HoNOS be used as the main instrument and any other instruments be 'anchored' against it, was seen by some as a balanced view of the outcome of the reference group, but by others it was seen as interests being swayed and 'judicious editing' of the report. HoNOS was seen to be by some, on balance, the best measure, not necessarily psychometrically, but in terms of its position as a reasonably popular instrument. Whether this was a consensus view is not absolutely clear.

The aftermath of the Fonagy report left somewhat of a vacuum in terms of any initiatives or policy drives to encourage outcome measurement. Those who were involved in the process seemed to be tentative to continue with outcome measurement. Following this there were also some drastic financial cuts to mental health services across the board, and there were fears that outcome measures could be used as a way to justify these cuts, further fuelling them being shelved.

11.4.2 The current policy perspective

As a follow-up to the Fonagy report, the Care Services Improvement Partnership (CSIP) is trying to drive forward the implementation of outcome measurement in the service once again.

The Care Services Improvement Partnership (CSIP) is the umbrella body for the National Institute for Mental Health in England (NIMHE), which in turn is an arms length body of the Department of Health. The Department of Health and Care Services Improvement Partnership (CSIP), as part of the

National Institute for Mental Health in England (NIMHE) are running a National Outcomes Measures project [117-119]. They have surveyed all Mental Health Trusts in England and are establishing a compendium of outcome measures in mental health services. The project aims to find out about available tools and people's experience of using these tools in practice. A steering group has been put together by CSIP to oversee the practice of outcomes measurement, building on the pilot sites work from the Fonagy report. The steering group is trying to ensure the implementation of some of the recommendations from the Fonagy report.

This group has commissioned the development of a Compendium of outcome measures, which will outline the vast number of available measures in mental health. The Compendium will not mandate any suite of measures. The problems for CSIP of mandating specific measures which they would like providers to use, is that the measures may change, and there are always complex arguments about the advantages and disadvantages of different measures in different circumstances. From their perspective, this will always result in "an interminable debate with protagonists". The Compendium will therefore provide information on the psychometric properties of different measures, their uses and in what circumstances they can best be applied, and what the copyright issues are for their use. A draft Compendium will be available in Autumn 2007 and a more definitive version in 2008.

The only measure which is mandated, is HoNOS, as part of the MHMDS. From the DoH point of view the use of HoNOS is not as widespread as they would like. The technical guidance on the collection of HoNOS will be updated in the near future to make it more clear, but it is argued that the minimum requirements should be clear – for patients on enhanced CPA, a HoNOS score should be collected at least once a year. At present the guidance on completion of HoNOS in the MHMDS states that everyone on CPA should receive a HoNOS rating which is inappropriate since this means effectively everyone in mental health services and HoNOS would not be appropriate for all patients. Within the services, the notion of whether or not HoNOS is in fact mandatory, has not been perceived to be clearly communicated from the DoH and with this in mind, the guidance could certainly be clearer from the DoH.

Therefore, the mandate is that at a minimum providers should complete HoNOS for patients on enhanced CPA at least once a year, and over an above that they are free to use any outcome measure they wish which would be appropriate for patients without severe and enduring mental illness. The Compendium will therefore give providers the up-to-date information on other available measures.

There are two main reasons that CSIP wishes to see routine outcome measurement embedded in practice:

- The first reason is for service improvement, so that services can ensure that they are being effective and useful (i.e. for the benefit of providers).
- The second reason is for service innovation, so that providers can demonstrate benefits for users in order to receive appropriate levels of investment (i.e. for the benefit of commissioners).

The DoH and CSIP view is that while the use of different measures is patchy and not a lot of information exists on their coverage, HoNOS is probably the measure with the widest coverage. CORE-OM is widely used in psychological therapy services. While other measures such as FACE, CUES and MANSA were piloted in the Fonagy report, they are not that widely used in practice. These measures did not do that well in the pilot sites.

In terms of the policy priority that outcome measurement receives, the 5-year review of the NSF [19] set out the key policy priorities for mental health services. Outcome measurement was not one of the priorities. Of the ten priorities, the final one is information systems. This states that "information systems remain primitive in mental health and the information we collect is not sophisticated enough for a modern service". Outcome measurement is not explicitly mentioned within the focus on improving information systems either.

11.4.3 The Australian system

Many interviewees mentioned the advantages and occasionally disadvantages of the Australian system where with the help of a lot of government support, they have an excellent benchmarking

database system, which could potentially also be mimicked in the UK. They have fully supported the use of HoNOS in all services and data is collected through the National Casemix and Outcomes process. Many proponents of outcome measurement seemed to envy many aspects of the Australian system and the fact that they have gone so far down the road of collecting patient level outcomes data.

It was however seen as unlikely that in the current policy climate a similar amount of resources would be ploughed into the UK system to get outcome measurement off the ground. From the DoH's point of view it would require new or additional funding to go down a similar route to that in Australia, the policy priority would not be to detract from any existing service provision.

The Australian system is however used very strictly to inform practice and no payment mechanisms are attached to the outcomes. This was seen as a major advantage.

There was a view that the UK is more pre-occupied with evidence-based medicine than Australia, and hence there wasn't the same pressure to justify the Australian system with RCT evidence. An allied view was that the Australian workforce may be more 'compliant' than that in the UK.

While Australia has done an excellent job with looking at aggregated data, there was a view that some Australians might be feeling some disillusionment with the system in terms of it not having delivered all the benefits that had been expected.

Some work is being done between CRU and the Australian Department of Health (and the New Zealand equivalent) to examine potential changes to the HoNOS scale. This might include an additional item 13 on anxiety which will break up the existing item 8 on the scale more. Since the Royal College are the copyright holders to HoNOS, any changes to the instrument would need to involve them.

There is however a lot of nervousness around amending the existing HoNOS in Australia, and then rolling out any changes within the UK because the copyright to HoNOS would've changed.

There is a lot of interest in the use of HoNOS in Northern Europe and Canada as well.

11.5 A case study of successful outcome measurement - CORC

An example which was mentioned during the interviews of a voluntary outcome measurement system which has been delivering some successes was CORC. In Child and Adolescent Mental Health Services (CAMHS) a suite of measures has been recommended by the CAMHS Outcomes Research Consortium (CORC) (see www.corc.uk.net). This is a collaboration between different Child and Adolescent Mental Health providers. The aim is to develop and pilot a model of routine outcome measurement that can provide information on the outcomes of services for children, young people and their families. This data can then be used by service providers, commissioners and users.

In December 2006 CORC had 53 members across the UK and one member in Norway. CORC is a collaboration funded by membership subscriptions. Members pay a membership fee of £5000 for the first year and £3000 for following years.

CORC is a members-led organisation with a central committee elected by members. As agreed by members, CORC recommends three measures:

- The Strengths and Difficulties Questionnaire (SDQ) - to capture child and parent views of change in difficulties
- The Children's Global Assessment Scale (CGAS) - to capture practitioner views of change in psychological, social and school functioning
- The Commission for Health Improvement (CHI) Experience of Service Questionnaire (ESQ) - to capture parent and child views related to their experience of service

The Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) is used by several members and it has therefore been agreed that it can be used as an additional measure for assessing practitioner views of difficulties particularly for older adolescents and young adults. A self-

report version exists that is suitable for young people over the age of 13 years. While HoNOSCA is not an agreed core measure, it is likely to be relevant to those working with young people with more severe difficulties, in specialist settings and with older adolescents. For this reason data from HoNOSCA can also be sent to the central database.

CORC encourages members to collect data from their services at two points (at first contact and 6 months later), to measure changes for the child and family. The data has to be stored in a suitable database before it is sent to CORC (currently once a year). The data is then analysed by the CORC central team who put together an individual report based both on this outcome data and any other basic demographic data. Individual services have to invest in data management processes to conform to these requirements, for example hiring a full-time researcher to collect the data from patients and enter it into a database.

Figure 11 provides an example of the feedback which the CORC central team might provide to Service A on their Strengths and Difficulties Questionnaire. Information is provided on how to interpret the plots, the quality of the data, and issues like causality, confounding factors and chance.
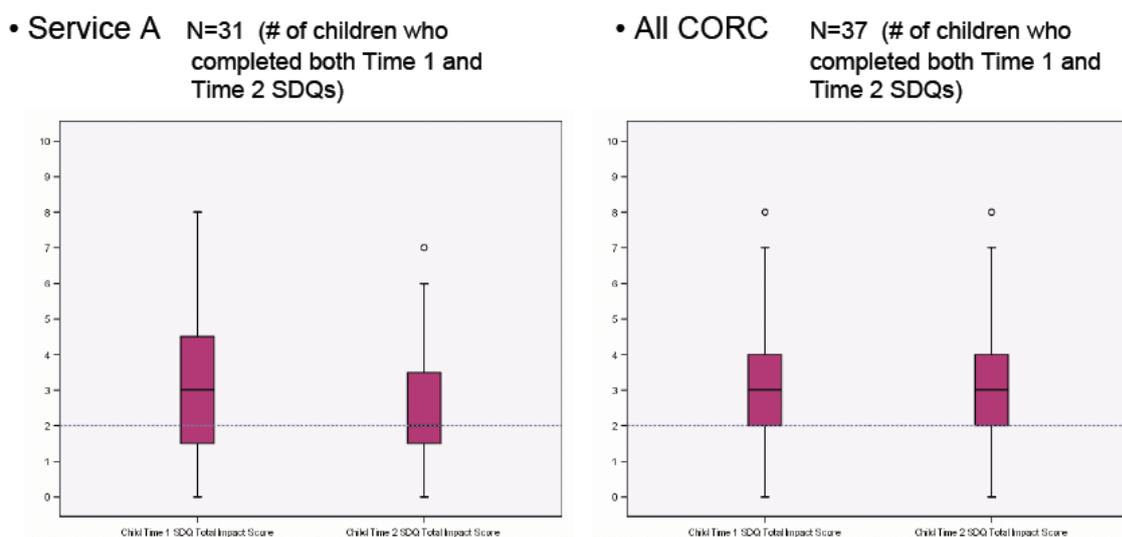
Figure 11: Example of CORC feedback provided to services

These CORC reports are fed back to clinicians every 6 months. So far, there are no clear examples of how the information is being fed back to patients. But in principle, they could be adapted by the service for use with patients, parents, or even commissioners.

Because CORC has been a bottom-up members-driven voluntary network, rather than a top-down centrally driven policy directive, it has been very successful in attracting members and getting routine outcome measurement into practice. The ethos of CORC is one of collaboration and learning from the experiences of all members. Many clinicians would feel a certain paranoia about measures being dictated to them from the top-down.

None of the members of CORC have left the group, which would suggest that they perceive it to be a useful network with valuable feedback processes. Paradoxically, there is no financial incentive to get the outcome measures off the ground, instead members have to pay for the scheme which seems to suggest that the incentive of valuable feedback is far more powerful than the financial incentive.

12.   Changes in the mix of treatments in mental health over time

The literature on calculating price indices for mental health care in the US is particularly relevant in examining the measurement of mental health outputs in the NHS. Although there are no price indices for NHS outputs, many of the issues in this literature are pertinent to how a productivity index might be constructed in the NHS setting for mental health. This literature specifically explores the change in the composition of treatments in mental health over time.

Rising expenditure on mental health has generated considerable interest in constructing price indices for this area of medical care, in particular: major depression, schizophrenia and bipolar disorder. What emerges from this literature is that it is important to focus on the direct medical costs of treating an episode of illness, rather than changes in the prices of the inputs used in treatment. For all three disorders, studies suggest that the price of treating an episode or individual have declined in recent years. This is contrary to many of the officially reported figures, for example from the Bureau of Labour Statistics (BLS).

This is because this literature improves on prior methods in that it attempts to define units of output of medical care that reflect the changing bundles of inputs that go into treating the illness. Output is also defined in a way that incorporates measures of the quality of treatment. The approach defines output as the services used in the treatment of disease, for example physician visits, hospital stays, prescriptions. This literature views these as inputs into the treatment of diseases. They view output instead as a course of treatment over a specified period, combining a number of treatment inputs which then produce health benefits. This conception of output allows for a change in the composition of inputs, or substitution among inputs as a result of technological change.

Many of the studies show that it is possible, even when input costs are rising, for treatment episode costs to fall, if there are changes in the composition of treatment. Berndt [120] argues that this can be explained by the fact that official (BLS) statistics do not make allowances for changes in the mix of treatment over time. These studies have found a considerable shift in the composition of treatment over time for depression [121-125], schizophrenia [126, 127], and bipolar disorder [128]. For example, these studies found a shift in the mix of treatment for depression which has taken place in recent years, with the combination of psychotherapy and tricyclic anti-depressants (TCAs), being replaced by the use of selective serotonin reuptake inhibitor (SSRI) drugs, sometimes also in combination with therapy.

This suggests that the gains from both pharmaceutical innovations and institutional changes have led to treatment for a greater number of episodes and individuals [120]. While this is an encouraging finding, uncertainty remains as to how quality of care has been affected. It is clear that there needs to be some adjustment for quality. Adequate data on quality and patient outcomes is not available from the medical claims data employed in most studies. Therefore, the authors use alternative approaches to account for quality.

12.1 Studies on depression

Depression accounts for approximately half of all mental health expenditure in the US [129]. There have been a number of studies investigating the pricing of its treatment. Although they do not use identical methods, there is a considerable degree of similarity in both the methods employed and the conclusions reached. The studies have focused on the episode of illness which involves pooling a number of treatments into 'bundles' which are *ex ante* expected to lead to similar outcomes.

These are used to construct demand and supply price indices for the treatment of major depression based on the cost of an episode of illness, for the period 1991-1995. Frank et al [125] identified nine major types of treatment in retrospective medical claims data from four large self-insured employers. Closely related treatment bundles were aggregated, resulting in the use of five bundles. An assumption of this methodology is that obtaining similar outcomes from alternative bundles provides similar expected utility levels. An additional assumption, is that the production function has a step-function form. So an individual receiving six psychotherapy sessions (barely meeting treatment guidelines) is seen as receiving "effective" treatment while an individual receiving four or five psychotherapy sessions (slightly less than treatment guidelines) is seen as receiving 'ineffective'

treatment. The proportion of episodes receiving 'effective' (guideline) care increased from 35% to 55% between 1991 and 1995 [129].

It is important to consider the substitutability of the treatment bundles. Fixed weight indices such as the base-period Laspeyre index and the end-period Paasche index, assume that there is no possibility of substitution among treatments. Therefore, in addition to these, the authors constructed a Cobb-Douglas index and a Törnqvist index (with more flexible assumptions about the substitutability).

All four indices showed that both the demand and supply price of treating depression had fallen between 28 and 32% between 1991 and 1995. The reported decline for the cost of treating depression is contrary to the increase in the BLS price indices over the same period. Frank et al [125] conclude that employing standard price indices – which do not allow for changes observed in treatment composition – may lead to changes in quantity being misinterpreted as changes in price, which is a cause for concern.

Several extensions to this work have since been carried out [122, 123].

Berndt et al [122] included a greater number of episodes and increased the number of treatment bundles to seven. They identified the bundles of treatment as: psychotherapy alone, tricyclic antidepressants (TCAs) alone, SSRI treatment alone, TCA treatment with psychotherapy, SSRI treatment with psychotherapy, combined TCA/SSRI treatment only, and combined TCA/SSRI treatment with psychotherapy. Substantial changes took place in the composition of treatment for acute phase major depression following the introduction of the first SSRIs. There was a large reduction in psychotherapy treatment alone but a more than doubling in the combination of psychotherapy treatment and SSRIs. Treatment involving the older generation antidepressant tricyclics (TCAs) declined. Thus there was substitution away from psychotherapy only, TCA only or TCA and psychotherapy combined, towards SSRI plus psychotherapy or SSRI only treatment. Using a Laspeyres price index, this resulted in a decline of only 0.6% per annum, compared to the 9.1% per annum decline reported by Frank et al [125].

Berndt et al [123] add to the previous research by using a method that combines clinical evidence and expert opinion. This approach has two advantages: First, it enables the calculation of treatment price indices that include variations over time in the proportion of 'off-frontier' production, which increases the number of observed episodes. Second it means that allowance can be made for different expected outcomes, and even for those patients put on the 'waiting list' or no treatment option, who often experience remission regardless. By doing so, the authors are able to assess the incremental outcome gains as a result of treatment compared to no treatment outcomes [123].

They use similar retrospective patient-level medical claims data containing information on prescription drug claims, inpatient hospital treatment, outpatient visits, diagnoses, procedures, and the demographic characteristics of all covered individuals. With an additional year of data (1991-1996), the authors identified and classified episodes of acute phase depression according to treatment and patient type. Information on expected treatment outcomes was gathered using a two-stage modified Delphi procedure. Ten panellists were asked to provide opinions on expected treatment outcomes, using the Hamilton Depression Rating Scale (HDRS). For each of the 120 patient treatment cells, a literature review of clinical research was provided. The experts were asked to consider a group of 100 patients and provide an estimate of how many patients would fall into four given categories: 'no depression', 'mildly depressed', 'moderately depressed' or 'no change'. They were also asked of the 100 patients meeting criteria for depression at initial visits, what number would fully respond to treatment, what number would evidence significant but partial response, and what number would not evidence any medically significant response. For 39% of the cells, substantial disagreement among the experts was noted. Each expert was then shown the group's rating for those cases where there was disagreement, and after meetings they were given the option to revise their own rating. After this process, no substantial disagreement remained.

The framework under which expected outcomes data were obtained allowed Berndt et al [123] to consider results for both full and partial remission. The probability of a full remission varies from 0.15 to 0.35, the probability of partial remission is higher. One particularly interesting finding is that approximately 40% of the treatment episodes make use of medical treatment with no or little effectiveness over the no treatment option, accounting for 20% of total spending on treatments.

Average spending per full or partial remission varies according to the treatment of those cases where there is no expected outcome rating. If these cases are excluded, or assigned the median outcome, average spending remains unchanged in 1996 from 1991. If they are assigned the worst possible outcome, average spending falls to 84 for full remission. For partial remission, using the median outcome also results in a decline in average spending to 90 in 1996. However, expenditure per incremental expected full or partial remission is preferred. Generally, the expenditure indices per incremental remission fall more rapidly than average expenditure.

A base-period Laspeyre, an end-period Paasche and a Fisher Ideal price index were constructed. The treatment price index falls for all three, but if incremental remission is considered, all three show a small rise. The trends in price indexes are slightly sensitive to how expected outcomes are treated. It may be that there have been changes in the patient population and treatment bundles.

To account for this potential change in patient mix, hedonic equations were estimated. Eight patient categories were identified, depending on the presence of medical comorbidity, gender, age if female and comorbid substance abuse. Three alternative hedonic equations for the price per expected full remission were estimated by Ordinary Least Squares (OLS). The results suggest that patient categories have significant and sizeable effects on treatment costs; comorbid substance abuse and comorbid medical conditions are particularly likely to increase treatment costs. Also, treatments having higher probability of full remission have higher costs *ceteris paribus*. All three equations estimate reductions in price. The results for partial remissions lead to similar findings. Estimating hedonic equations produces different results to those shown by the price indices, a reflection of the changing mix of patients over time. Overall, the cost of treating episodes of depression, conditional on expected outcome, has generally declined between 1991 and 1996. The results of the hedonic regressions indicate a fall of between -1.66% and -2.13% per annum over this period.

Berndt et al [123] conclude that using expected outcomes data gained in this way can help to construct price indices for medical treatments that allow for variation in expected outcomes and patient populations. However, it is acknowledged that the reliability of the outcome measures are a key determinant of the usefulness of this method, and validity checks are carried out to assess this. In conclusion, the composition change away from more costly psychotherapy-intensive treatment to less costly psycho-pharmalogical treatments had a significant impact on the average cost of treating an episode of acute phase major depression. Between 1991 and 1996, controlling for expected outcomes, the average cost of attaining expected full and partial remissions declined by about 12%. Since expenditures on depression are thought to be increasing over this period, it seems that the source of this is an increase in volume, (quantity) rather than price increases, as the cost of treating an episode of depression seems to have fallen. Quality also improved because over the period episodes meeting guideline standards increased from 42% to 49%.

12.2 Studies on schizophrenia

This approach of pricing an episode of care has also been implemented for other mental health disorders. Frank et al [126] investigate the changing costs of treating schizophrenia. Using data from two counties in Florida over the period 1994/95 to 1999/00, this study sets out to answer three main questions: what is the cost of treating an individual with schizophrenia on a per annum basis?; how have composition and quality of care altered over time?; and how are estimates of changes in treatment costs affected if quality is controlled for? As the 1990s saw the introduction of significant new treatments, the authors suggest that considerable changes in treatment patterns were likely to have taken place, impacting in turn on both the price and quality of care [126].

Producer price indices are constructed based on annual episodes of care. As the nature of schizophrenia is such that it would be considerably more complex for clinical experts to estimate probable outcomes, the authors use the recommendations made by the Patient Outcomes Research Team (PORT) to account for quality. Five of their recommendations referred to types of treatment observable from the claims data – the use of pharmacotherapy was found to be beneficial for example, whereas the evidence for vocational rehabilitation appeared to have little effect. The authors [126] then used these to construct treatment bundles, ranging from evidence-based to evidence-equivocal to no supporting evidence. Four bundles consisted of single treatments, such as any antipsychotic medication and others which comprised more than one form of treatment such as medication and therapy. An 'all other treatment' category was also formed. They defined output as the

course of treatment over an entire year, given that schizophrenia is a severe and persistent mental disorder. Each person-year observation was then categorised to one or more of the treatment bundles as appropriate.

The authors note that a shortcoming of measuring quality in this way is that it merely reveals whether a patient has received treatment, and provides no indication of whether the amount of treatment was sufficient. They also point out that in some sense this reflects only minimum standards of care, as it seems reasonable to expect that a patient should receive at least one form of treatment.

Multivariate regression equations were estimated, using the natural log of annual mental health direct medical costs as the dependent variable. Explanatory variables included patient characteristics, enrolment and medical histories and indicator variables for the treatment bundles. As this period saw significant changes in treatment, Chow tests were applied to test for parameter stability. This resulted in the rejection of parameter equality, and therefore the data were not pooled across years.

Quality and patient characteristics were held constant at a point in time, and predicted spending compared. This raises further questions over which year should be used to fix these factors, and various options were explored. An important finding is that, contrary to what one would normally expect, the fixed weight and chained Laspeyre price indices fall more rapidly than the comparative Paasche indices. Frank et al [126] interpret this as reflecting a rightward shift of the demand curve, as over time doctors accumulate knowledge of the efficacy of the newer pharmaceuticals, and alter their prescribing behaviour leading to positive correlations between price and quantity. They ran regression models with annual mental health spending as the dependent variable and combinations of treatment bundles as indicator variables. Their preferred price index revealed an average price decline of about 5.5% annually or a cumulative decline of about 25% between 1994 and 2000.

They found that compositional changes in the treatment of schizophrenia had been significant with various forms of psychosocial therapy and older pharmaceutical treatments being replaced by newer atypical antipsychotics, in line with PORT guidance. This leads the authors to raise the concern that a significant amount of treatment is not supported by clinical evidence.

Quality of care data was not as detailed as the Berndt et al [123] study, but this study demonstrates the considerable importance of controlling for quality. Without any adjustment for quality, mean mental health-related treatment costs for individuals with schizophrenia increased 2.4% over the period. When treatment quality is held constant, both the fixed weight and chained Fisher Ideal indices showed cumulative price falls of more than 22% over the period. The results show that without adjusting for quality changes, there has been a rise of 0.5% per annum in the individual annualised cost of treating schizophrenia. If quality of care and patient characteristics are held constant over time, mean treatment costs have fallen by 5.5% per annum between 1994/95 and 1999/2000, a cumulative price decline of around 25% over the period, with the greatest decline occurring in the first half of the 6-year period under consideration. Overall, they conclude that as the cost of treating an individual per annum has declined, the observed increase in overall expenditure is an indication that there has been an increase in the number of individuals being treated.

In a more recent paper by Frank et al [127], they examine whether the introduction of new medications to treat schizophrenia offset their costs by reducing spending in other areas. They use seven years of Florida Medicaid data and show that the new drugs result in a net increase in spending, which they attribute to increased adherence to treatment and higher levels of contact with treatment providers. The newer generations of atypical anti-psychotics have fewer side-effects, tend to be more effective against negative symptoms and are therefore better tolerated, and thus may reduce relapses which result in more hospital use. However, the new drugs have some side-effects of their own, namely weight gain and the risk of diabetes. They find rapid uptake of the new drugs but lack of cost-offset which may raise questions around their cost-effectiveness.

While this study doesn't specifically examine the compositional change of treatment bundles, spending is calculated by summing all the different components of treatment, for example psychotherapy and psychotropic drugs which would've changed over time. The study also does not measure patient outcomes on the two different generations of anti-psychotics.

12.3 Studies on bipolar I disorder

Research by Ling et al [128], discussed in Berndt [120] investigates the changing costs of treating bipolar I disorder. Like schizophrenia, there is at present no known cure for bipolar I disorder, so the focus of medical care in this case is to try and improve symptoms, including the prevention of recurrences. They defined four mutually exclusive and exhaustive treatment bundles: no treatment, psychotherapy only, mood stabilizers only and psychotherapy and mood stabilizers combined. In contrast however, the compositional changes in types of treatment for bipolar disorder have been more gradual than either depression or schizophrenia.

Ling et al [128] estimate a multivariate regression model, similar to that used by Frank et al [126], with indicator variables for treatment types and for patient characteristics. The results suggest a fall of 31% in the cost index between 1991 and 1995, compared to a decline of 12% if no allowance is made for changes in patient mix and treatment composition. The entire reduction occurred in the last 3 years. The average price decline was around 3.3% per annum.

12.4 Studies covering all mental health

Taking the evidence from the above studies, a recent study by Berndt et al [130] examined the level and composition of all mental health spending in the US. They apply quality adjusted price indices for several major mental disorders (anxiety, schizophrenia, bipolar disorders, major depressive disorders, and all others) to national mental health expenditure account estimates to examine changes in real output for the whole mental health sector. They use estimates on depression, schizophrenia and bipolar I disorder from the previous research, and then aggregate results across all categories of mental illness to arrive at overall price indices. These price indices reveal large gains in real output (70%-75%) relative to the ones used by the US Bureau of Labour Statistics (16%-17%).

12.5 Alternative approaches to productivity measurement

An alternative approach to calculating price and output indices, with Paasche, Laspeyre, Fisher, or Törnqvist indices, is to use a non-parametric approach such as Data Envelopment Analysis (DEA) to calculate a Malmquist productivity index.

A Malmquist productivity index has been calculated for Norwegian psychiatric outpatient clinics using DEA to examine whether any change is related to personnel mix, budget growth or financial incentives [131]. Bootstrapping methods are used to construct confidence intervals for the technical productivity index and its decomposition. A second stage regression is run on the productivity index to examine variables that may potentially be statistically associated with productivity growth. Overall they find substantial technical productivity growth. They find that personnel growth has a negative impact on productivity growth while growth in personnel with university educations increases productivity. Other than staff education as a proxy for staff quality on the input side, their study does not take account of any other changes in the quality of the output or interventions over time and they call for more research to explore this in the future.

12.6 Can these studies be reproduced in the UK with existing data?

In short, the answer to this question is, not without considerable difficulty.

The Mental Health Minimum Data Set covers mental health service usage and need [37, 75, 132]. The data set covers care received by service users during a spell, and includes details of clinical problems, treatments given, aspects of social care and outcomes. One of the key advantages of collecting data as spells of care is that they will include outpatient attendances which may extend over many years, as well hospital, community and day care attendances, which may commonly overlap. For each institution one would be able to track the characteristics of the patient, the health organisations involved, the nature of the problems including the range and severity, the amounts of different interventions delivered to the patient, the way these are combined as packages and scheduled over time, and how the patient's condition changes over time. This dataset does not however contain any data on the costs of treatments. This data set has only been in existence for three years and the quality and coverage is therefore still in its infancy.

The Mental Health Mapping data contains data on resource use in Local Implementation Teams (LIT), in particular staffing levels in different types of mental health teams. It provides a national picture on service provision in mental health services and covers 12 categories of service provision from public through to for-profit and not-for-profit and voluntary organisations. Alongside the Service Mapping data, a financial mapping exercise has been undertaken to identify NHS and local authority investment in mental health services. Each LIT is asked to provide information on service level revenue, capital and overhead costs. There is time series data available for both the Service Mapping and Financial Mapping data. These datasets are available for 6 years up to 2005/06. The Mapping data would however be unlikely to be able to provide detailed patient level costing of treatments, as it done at a service level.

Cost data could in theory be obtained from the PbR pilot sites which used HoNOS Plus to derive 13 groupings based on need and hence predicted resource use (rather than diagnosis typically used in HRGs). Each of these needs groupings were costed in the pilot project. However, there would not be an obvious way of linking this cost data to the MHMDS since no mapping was done between HoNOS Plus and HoNOS in the pilot projects, so this data cannot readily be integrated. Thus while output indices could in principle be generated from the MHMDS and there may even be the possibility of tracking compositional changes in treatment over time, there is no way of producing price indices equivalent to those generated from the medical claims data in the US. No costing data is at present readily or routinely available. Although Reference Cost data for mental health services is expanding every year, it is still insufficiently detailed to be linked to activity data to produce adequate 'price indices'.

13. Conclusions

The objectives of this report were to answer the following questions:

1.   What outcome measures are most commonly used in mental health?

From the literature and the accounts of interviewees, the consensus was that HoNOS and CORE-OM are probably the measures with the widest coverage in the NHS, even though this coverage is still patchy and in some areas non-existent.

2.   What are the practical issues about collecting these outcome measures routinely?

Some of the key issues involve ensuring a clear rationale for the introduction of routine outcome measurement with useful clinical feedback mechanisms to clinical teams so that the value of routine outcome measurement can be appreciated, the establishment of an outcome-oriented culture, strong management support and some positive incentives such as peer pressure. IT systems also absolutely need to support the easy entry of data and the easy access of feedback for clinical teams.

3.   Can these disease specific outcome measures be converted into a QALY or some other generic measure?

This cannot readily be done as valuation of either HoNOS or CORE-OM would be extremely complex and mapping would only be as good as the degree of overlap that existed between these measures and the generic measure. The evidence suggests that there will not be a strong overlap.

4.   Is there a time series of these outcome measures available?

Time series data on activity and outcome is available for both HoNOS and CORE-OM, but data quality is still a concern and access to CORE-OM would need to be negotiated.

5.   Is it feasible to examine changes in the mix of treatments in mental health over time?

In short, output indices could probably be established in the future with some examination of changes in treatments. However, the costing of activity for changes in treatments is still a very long way off.

In short, can mental health outcomes data be incorporated into a productivity index at present? The answer at this stage is probably still no. In essence with only two data points for MHMDS, this severely limits what can usefully be done at the present time. While CORE-OM may have longer time-series available at present, one cannot usefully produce an index without HoNOS at this point in time, since one cannot exclude the outcomes of patients with severe and enduring mental illness.

There are still some significant shortcomings in the form of costing data readily linked to activity data. The collection of cost data will commence when PbR comes on board for mental health, most likely in 2009. In addition, the coverage of the existing outcome data will need to improve for it to be used in a productivity measure. Some crude adjustments should then be possible. Outcome data collection needs to be improved, particularly at time T2.

What recommendations could be made to expand the coverage of routine outcome collection in mental health in the NHS?

1. From a policy perspective, the technical guidance on HoNOS within the MHMDS needs to be urgently rewritten. There is a lack of clarity within the service about what the minimum requirements actually are and whether indeed HoNOS is mandatory or not. This uncertainty is used as an excuse to not complete HoNOS.
2. If other morbidity measures are to be used alongside HoNOS, there needs to be clearer guidance (and a research effort) on how HoNOS equivalent scores can be obtained from other instruments for inclusion in the MHMDS. Mapping from one instrument to another is clearly a problematic and challenging issue even from a research point of view, so it is not clear how the service is supposed to obtain appropriate scores. The CSIP Compendium could be used to give guidance on this issue.

3. If indeed outcome measurement is to be a priority, it needs to be stated as such at the very highest levels and driven through documents such as the National Service Framework and the Healthcare Commission Developmental Standards. At present, the lack of clear directive leaves many services to think it is supposed to be a priority, but one which needs to happen 'in their spare time' and with no additional resources, and hence it is not regarded as a priority in practice.
4. IT systems in many Trusts are poorly developed and cannot possibly support any kind of routine outcome measurement system. If the drive for routine outcome measurement is to be pursued in earnest, this needs to be addressed urgently, with additional resources and clear guidance.
5. At provider level, feedback mechanisms need to be found which can supply quick and clinically useful feedback to clinical teams on outcome measures. A dialogue needs to take place about clinical requirements and informatics capabilities. Obtaining useful and timely feedback is a necessary (but not sufficient) condition for routine outcome measurement.
6. A necessary and sufficient condition is the addition of management support which is absolutely imperative. This involves providing the resources and support to embed outcome measurement into routine practice to support clinical decision-making in the first instance, rather than to solely support management decisions. Top-down drives to enforce routine collection will rarely be effective.
7. Finally, and most challengingly, an outcome oriented culture needs to be developed which is driven by the desire to learn about improving service quality rather than the desire to benchmark, league-table and remove 'failing' services. This culture may however incorporate some aspects of mild coercion and peer pressure provided this takes place in a developmental and learning culture.

14. Glossary

BASIS-32 - Behaviour and Symptom Identification Scale 32
BAI - Beck Anxiety Inventory
BDI - Beck Depression Inventory
BDI-I - Beck Depression Inventory I
BDI-II - Beck Depression Inventory II
BHS - Beck Hopelessness Scale
BLS - Bureau of Labour Statistics
BPRS - Brief Psychiatric Ratings Scale
BSI - Brief Symptom Inventory

CAMHS - Child and Adolescent Mental Health Services
CANE - Camberwell Assessment of Need for the Elderly
CGAS - Children's Global Assessment Scale
CIS-R - Clinical Interview Schedule – Revised
CMHT - Community Mental Health Team
CORE-5 - CORE-OM 5-item questionnaire
CORE-10 - CORE-OM 10-item questionnaire
CORE-A - CORE-OM Assessment
CORE-NET - CORE-OM version for networked system
CORE-NR - CORE-OM without risk items
CORE-OM - Clinical Outcomes in Routine Evaluation – Outcome Measure
CORE-PC - CORE-OM version for personal computers
CORE-SF - CORE-OM short-form
CPA - Care Programme Approach
CRU - Royal College of Psychiatrists' Research Unit
CSIP - Care Services Improvement Partnership
CUES - Carers and Users Expectations of Services

DEA - Data Envelopment Analysis
DoH - Department of Health
DRG - Diagnostic Related Groups

EQ-5D - instrument used to produce a QALY

FACE - Functional Analysis of Care Environments
FIHS - Factors Influencing Health Status
FOC - Focus of Care
FT - Foundation Trust

GAF - Global Assessment of Functioning Scale
GP - General Practitioner
GP-CORE - CORE-OM version for use in the general population

HADS - Hospital Anxiety and Depression Scale
HDRS - Hamilton Depression Rating Scale
HoNOS - Health of the Nation Outcome Scales
HoNOS-ABI - HoNOS for acquired brain injury
HoNOSCA - HoNOS for Children and Adolescent Mental Health Services (CAMHS)
HoNOS-LD - HoNOS for learning disabilities
HoNOS-secure - HoNOS for forensic services
HoNOS65+ - HoNOS for older people
HRGs - Healthcare Resource Groups
HRQL - health related quality of life

ICD - International Classification of Diseases
IIP-32 - Inventory of Interpersonal Problems - 32-item version
IIP-Av - Inventory of Interpersonal Problems - Avoidant
IMS - CORE Information Management Systems

IT - Information Technology

K-10+ - Kessler-10 Plus

LIT - Local Implementation Teams
LSP - Abbreviated Life Skills Profile

MANSA - Manchester Short Assessment of Quality of Life
MHI-38 - Mental Health Inventory
MHMDS - Mental Health Minimum Data Set

NHS - National Health Service
NICE - National Institute for Health and Clinical Excellence
NIMHE - National Institute for Mental Health in England
NSF - National Service Framework

OCD - Obsessive Compulsive Disorder
OLS - Ordinary Least Squares

PbR - Payment by Results
PCT - Primary Care Trust
PORT - Patient Outcomes Research Team

QALY - Quality Adjusted Life Year
QLS - Heinrichs Quality of Life Scale
QOF - Quality and Outcomes Framework
QOLI - Lehman's Quality of Life Index

RCT - Randomised Controlled Trial
RUG-ADL - Resource Utilisation Groups – Activities of Daily Living

SDQ - Strengths and Difficulties Questionnaire
SF-36 - generic short-form health survey with 36 questions
SF-6D - generic short-form version of SF-36
SFS - Social Functioning Scale
SSRI - selective serotonin reuptake inhibitor
StHA - Strategic Health Authority

TCA - tricyclic anti-depressant

US - United States

VA - Veterans Administration
VAS - visual analogue scale
VAT - Value Added Tax

YP-CORE - CORE-OM for young people

15.   Appendix 1 – Literature search strategy

Databases
PsycInfo
HMIC
MEDLINE

References found:
- after deduplication: 1504
- before deduplication: 1805
- actual references retrieved: 164

Search strategies

PsycInfo
Date searched - 16/04/07
Records found - 672
Filename - psycinfo.txt

Limits:
UK studies only, 2000-2007, English language only, human only, HoNOS terms added - these have not been limited by publication date or to UK studies

1. exp *Mental Health/
2. exp *Psychiatry/
3. exp *Mental Disorders/
4. mental health.ti,ab.
5. mental$ ill$.ti,ab.
6. mental$ ill-health.ti,ab.
7. psychiatry.ti,ab.
8. psychiatric.ti,ab.
9. mental$ disorder$.ti,ab.
10. or/1-9
11. exp *Treatment Outcomes/
12. exp *Psychological Assessment/
13. health status indicator$.ti,ab.
14. health outcome$.ti,ab.
15. ((measure$ or assess$ or score$ or scoring or index or indices or scale$ or monitor$) adj2 outcome$).ti,ab.
16. "outcomes measurement".id.
17. *"Treatment Effectiveness Evaluation"/
18. ((improv$ or measur$) adj (productivity or performance)).ti,ab.
19. ((output or price) adj (index$ or indices)).ti,ab.
20. or/11-19
21. (nhs or national health service or england or ireland or scotland or wales or united kingdom or great britain).mp.
22. (britain or british or great or england or wales or scotland or ireland or kingdom).lo.
23. 21 or 22
24. 10 and 20 and 23
25. 24
26. limit 25 to (human and english language and yr="2000 - 2007")
27. health of the nation outcome$ scale$.mp.
28. honos$.mp.
29. 27 or 28
30. 29
31. limit 30 to english language
32. 26 or 31

HMIC
Date searched - 16/04/07
Records found - 557
Filename - hmic.txt

Limits:
No limits. Includes DH-Data (Dept of Health database) and the King's Fund database, has a UK focus. Also includes 'grey' literature such as DoH reports.

1. mental health.ti,ab.
2. mental$ ill$.ti,ab.
3. mental$ ill-health.ti,ab.
4. psychiatry.ti,ab.
5. psychiatric.ti,ab.
6. mental$ disorder$.ti,ab.
7. or/1-6
8. health status indicator$.ti,ab.
9. health outcome$.ti,ab.
10. ((measure$ or assess$ or score$ or scoring or index or indices or scale$ or monitor$) adj2 outcome$).ti,ab.
11. ((improv$ or measur$) adj (productivity or performance)).ti,ab.
12. ((output or price) adj (index$ or indices)).ti,ab.
13. or/8-12
14. 7 and 13
15. 14
16. limit 15 to (yr="2000 - 2007")
17. health of the nation outcome$ scale$.mp.
18. honos$.mp.
19. 17 or 18
20. 16 or 19

MEDLINE (Including MEDLINE In-Process)
Date searched - 16/04/07
Records found - 576
Filename - medline.txt

Limits:
'Routine' terms limit, 2000-2007, English language only, human only. Search terms added to retrieve records on output indices and price indices.

1. exp *Mental Health/
2. exp *Psychiatry/
3. exp *Mental Disorders/
4. mental health.ti,ab.
5. mental$ ill$.ti,ab.
6. mental$ ill-health.ti,ab.
7. psychiatry.ti,ab.
8. psychiatric.ti,ab.
9. mental$ disorder$.ti,ab.
10. or/1-9
11. exp *Treatment Outcome/
12. *"Outcome Assessment (Health Care)"/ or *"Outcome and Process Assessment (Health Care)"/
13. *Health Status Indicators/
14. health status indicator$.ti,ab.
15. health outcome$.ti,ab.
16. ((measure$ or assess$ or score$ or scoring or index or indices or scale$ or monitor$) adj2 outcome$).ti,ab.
17. ((improv$ or measur$) adj (productivity or performance)).ti,ab.
18. (output adj (index$ or indices)).ti,ab.
19. (price adj (index$ or indices)).ti,ab.

20. exp Psychiatric Status Rating Scales/
21. ((outcome$ or process$) adj2 assessment$).ti,ab.
22. or/11-21
23. routine$.ti,ab.
24. 10 and 22 and 23
25. limit 24 to (humans and english language and yr="2000 - 2007")
26. health of the nation outcome$ scale$.mp.
27. honos$.mp.
28. 26 or 27
29. limit 28 to (humans and english language)
30. 25 or 29

16. Appendix 2 – The CORE-OM

The CORE-OM is a 2-page 34-item questionnaire as illustrated below.

Page 1:

Page 2:

Over the last week	Not at all	Only Occasionally	Sometimes	Often	Most or all the time	OFFICE USE ONLY
15 I have felt panic or terror	0	1	2	3	4	P
16 I made plans to end my life	0	1	2	3	4	R
17 I have felt overwhelmed by my problems	0	1	2	3	4	W
18 I have had difficulty getting to sleep or staying asleep	0	1	2	3	4	P
19 I have felt warmth or affection for someone	4	3	2	1	0	F
20 My problems have been impossible to put to one side	0	1	2	3	4	P
21 I have been able to do most things I needed to	4	3	2	1	0	F
22 I have threatened or intimidated another person	0	1	2	3	4	R
23 I have felt despairing or hopeless	0	1	2	3	4	P
24 I have thought it would be better if I were dead	0	1	2	3	4	R
25 I have felt criticised by other people	0	1	2	3	4	F
26 I have thought I have no friends	0	1	2	3	4	F
27 I have felt unhappy	0	1	2	3	4	P
28 Unwanted images or memories have been distressing me	0	1	2	3	4	P
29 I have been irritable when with other people	0	1	2	3	4	F
30 I have thought I am to blame for my problems and difficulties	0	1	2	3	4	P
31 I have felt optimistic about my future	4	3	2	1	0	W
32 I have achieved the things I wanted to	4	3	2	1	0	F
33 I have felt humiliated or shamed by other people	0	1	2	3	4	F
34 I have hurt myself physically or taken dangerous risks with my health	0	1	2	3	4	R

THANK YOU FOR YOUR TIME IN COMPLETING THIS QUESTIONNAIRE

Total Scores

Mean Scores
(Total score for each dimension divided by number of items completed in that dimension)

(W) (P) (F) (R) All items All minus R

Below is the look-up table of CORE-OM scores and severity levels [87].

Non-clinical range

Total score	Clinical Score	Simple score	Severity Level
1	0.3		
2	0.6	0	
3	0.9		
4	1.2		
5	1.5	1	
6	1.8		
7	2.1		
8	2.4	2	
9	2.6		Healthy
10	2.9		
11	3.2		
12	3.5	3	
13	3.8		
14	4.1		
15	4.4	4	
16	4.7		
17	5.0		
18	5.3	5	
19	5.6		
20	5.9		
21	6.2		
22	6.5	6	
23	6.8		
24	7.1		
25	7.4	7	
26	7.6		
27	7.9		Low level
28	8.2		
29	8.5	8	
30	8.8		
31	9.1		
32	9.4	9	
33	9.7		

Clinical range — Mild, Moderate, Moderate-to-severe

Total score	Clinical Score	Simple score	Severity Level
	Clinical cut-off level		
34	10.0		
35	10.3	10	
36	10.6		
37	10.9		
38	11.2		
39	11.5	11	
40	11.8		
41	12.1		
42	12.4	12	Mild level
43	12.6		
44	12.9		
45	13.2		
46	13.5	13	
47	13.8		
48	14.1		
49	14.4	14	
50	14.7		
51	15.0		
52	15.3	15	
53	15.6		
54	15.9		
55	16.2		
56	16.5	16	
57	16.8		
58	17.1		Moderate level
59	17.4	17	
60	17.6		
61	17.9		
62	18.2		
63	18.5	18	
64	18.8		
65	19.1		
66	19.4	19	
67	19.7		
68	20.0		
69	20.3	20	
70	20.6		
71	20.9		
72	21.2		
73	21.5	21	
74	21.8		
75	22.1		Moderate-to-severe level
76	22.4		
77	22.6	22	
78	22.9		
79	23.2		
80	23.5	23	
81	23.8		
82	24.1		
83	24.4	24	
84	24.7		

Clinical range — Severe

Total score	Clinical Score	Simple score	Severity Level
85	25.0		
86	25.3	25	
87	25.6		
88	25.9		
89	26.2		
90	26.5	26	
91	26.8		
92	27.1		
93	27.4	27	
94	27.6		
95	27.9		
96	28.2		
97	28.5	28	
98	28.8		
99	29.1		
100	29.4	29	
101	29.7		
102	30.0		
103	30.3	30	
104	30.6		
105	30.9		
106	31.2		
107	31.5	31	
108	31.8		
109	32.1		
110	32.4	32	Severe level
111	32.6		
112	32.9		
113	33.2		
114	33.5	33	
115	33.8		
116	34.1		
117	34.4	34	
118	34.7		
119	35.0		
120	35.3	35	
121	35.6		
122	35.9		
123	36.2		
124	36.5	36	
125	36.8		
126	37.1		
127	37.4	37	
128	37.6		
129	37.9		
130	38.2		
131	38.5	38	
132	38.8		
133	39.1		
134	39.4	39	
135	39.7		
136	40.0	40	

Guidance notes

1. The original mean item score can be readily calculated by dividing the clinical score by 10.
2. The 'simple' score uses the first integer only of the clinical score as a rough guide.
3. The reliable change index is 5 points and the cut-off level is a clinical score of 10 (or .5 and 1 respectively if using the traditional scoring method).

Below is an example of the CORE appraisal tool from CORE-PC which can be used for performance management

Name:			Report Date:					Filter:	
Practitioner **X**			09-01-06						

					Min	Max	Avg	X	Comments
Clients Attending 1st Session					10	110	46	60	Above average
		National Range			Service Range			17	Comments
		Min	Max	Avg	Min	Max	Avg		
Use of CORE	OM1	61	91	82	0	100	88	50	May be struggling to use CORE with clients. Explore
	OM1 + 2	15	83	45	0	100	38	15	Very low proportion of outcomes. Explore
KPIs	Clinical and/or Reliable Improvement	52	88	70	0	100	70	100	Excellent figures for feedback
	Deterioration	0	6	3	0	100	3	0	Excellent figures for feedback
	Unplanned Endings	9	35	19	0	100	18	15	Lower than service average
	DNAs	7	34	17	0	98	21	14	Lower than service average
First Sessions	Not Suitable for therapy	0	8	1.5	0	100	3	29	Very high levels of selectivity. Explore
	Referred on	0	5	1.5	0	100	3	1	OK
	Single Session	0	7	4	0	100	8	0	No single session work. Explore
	Accepted for therapy	74	98	87	0	100	68	71	OK
Below cut-off at referrals		8	25	20	0	100	22	5	Very high clinical rates. Explore.
Avg waiting times		22	90	54	0	565	71	6	Very fast access. Explore

General Appraisal
Some excellent examples of client management and clinical effectiveness. Seems to be high levels of client selectivity and low levels of completion of the CORE Outcome Measure which we usefully explored.

Action Plan	
Give positive feedback	Great outcomes for those with measures and evident effective client selectivity
Focus 1	The introduction of the CORE measure at assessment which X feels uncomfortable about
Focus 2	The selection criteria that X uses could be used by other team practitioners. Explore further
Focus 3	Waiting time management is something X should share with rest of team
Focus 4	Some single sessions must be therapeutically beneficial but X reports not doing single session work. Explore further
Focus 5	Explore DNA management in next appraisal/mentoring session

17.   Appendix 3 – Interview schedule

1. What, if any, outcome measures are used in clinical practice in mental health?

2. Is the Health of the Nation Outcome Scales (HoNOS) collected and what are its advantages and disadvantages?

3. Is CORE, FACE, CUES, MANSA collected? What are the advantages and disadvantages of these outcome measures?

4. Are psychometric properties an important consideration in choosing an instrument?

5. How do you gather information to rate HoNOS (or any other instruments), e.g. direct observation, interviews, medical records, consultation with other staff, consultation with carers/ family? Is the rating based on the most severe problem during a period or is it based on the current view of the patient?

6. What do you think reactions might be to collecting a patient-reported generic measure like EQ-5D alongside disease-specific measures?

7. What are the practical issues about collecting outcome measures routinely? Who does it, and how often? Do you have IT infrastructure to support its routine use? Do you have management support? What are issues around training? How time-consuming is it?

8. What are the key benefits and barriers to the use of outcome measures? What incentives do you think might work to increase the routine use of health outcome measures (e.g. PbR)?

9. Do you feel outcome measurement actually leads to a change in outcomes/improvement in quality? Do you see its clinical relevance?

10. Does the Trust / PCT / commissioner require you to collect any health outcomes data? Do you collect / are you required to collect any other quality measures instead e.g. readmission rates, suicide rates? Does any of this data get fed back to you (or the patient), and if so, is it useful?

11. What do you see as future policy developments in this area?

18.   References

1.   Atkinson, A. (2005) *Atkinson Review: Final Report*, HMSO: London.

2.   Castelli, A., et al. (2007) A new approach to measuring health system output and productivity, *National Institute Economic Review*, 200: 105-117.

3.   Fonagy, P., Matthews, R., and Pilling, S. (2004) *The Mental Health Outcomes Measurement Initiative: Report from the Chair of the Outcomes Reference Group*, National Collaborating Centre for Mental Health: London.

4.   Clifford, P. (1998) M is for outcome: the C.O.R.E. outcomes initiative, *Journal of Mental Health*, 7(1): 19-24.

5.   Lelliott, P., et al. (2001) Carers' and Users' Expectations of Services-User version (CUES-U): A new instrument to measure the experience of users of mental health services, *British Journal of Psychiatry*, 179: 67-72.

6.   Andrews, G. and Peters, L. (1994) *Measurement of consumer outcome in mental health*, in *A report to the National Mental Health Information Strategy Committee*, CRUFAD, Editor, CRUFAD: Sydney.

7.   Gilbody, S.M., House, A.O., and Sheldon, T.A. (2002) Outcomes research in mental health: Systematic review, *British Journal of Psychiatry*, 181: 8-16.

8.   Slade, M. (2002) What outcomes to measure in routine mental health services, and how to assess them: a systematic review, *Australian & New Zealand Journal of Psychiatry*, 36(6): 743-53.

9.   Gilbody, S.M., et al. (2003) *Outcomes measurement in psychiatry: a critical review of outcomes measurement in psychiatric research and practice*, in CRD report; 24, University of York: York. pp115.

10.   Ford, D. (2005) *Measuring quality in mental health services in the United States*, Final Report, Commonwealth Fund and Health Foundation Harkness Fellow in Health Care Policy. pp60.

11.   Bower, P., Gilbody, S., and Barkham, M. (2006) Making decisions about patient progress: the application of routine outcome measurement in stepped care psychological therapy services, *Primary Care Mental Health*, 4(1): 21-28.

12.   Callaly, T. and Hallebone, E.L. (2001) Introducing the routine use of outcomes measurement to mental health services, *Australian Health Review*, 24(1): 43-50.

13.   Godfrey, M. and Wistow, G. (1997) The user perspective on managing for health outcomes: the case of mental health, *Health and Social Care in the Community*, 5(5): 325-332.

14.   Slade, M., et al. (2006) Pre-morbid IQ and response to routine outcome assessment, *Psychological Medicine*, 36(8): 1183-1191.

15.   Dickey, B. (2001) *Measuring quality: an overview*, in *Improving mental health care: commitment to quality*, Dickey, B. and Sederer, L., Editors, American Psychiatric Publishing: Washington, D.C.

16.   Kramer, T.L., et al. (2003) Relationship between self-report and clinician-rated impairment in depressed outpatients, *Community Mental Health Journal*, 39(4): 299-307.

17.   Garcia, R., et al. (2002) A comparison of parent and therapist ratings of outcome in a child mental health clinic, *Child and Adolescent Mental Health*, 7(4): 168-172.

18.   Huxley, P. and Evans, S. (2002) Quality of life routine outcomes measurement: Lessons from experience in the USA and the UK, *Epidemiologia e Psichiatria Sociale*, 11(3): 192-197.

19.   Appleby, L. (2004) *The National Service Framework for Mental Health – Five Years On*, Department of Health: London. pp84.

20.   Stedman, T., et al. (2000) The perceived utility of six selected measures of consumer outcomes proposed for routine use in Australian mental health services, *Australian and New Zealand Journal of Psychiatry*, 34(5): 842-849.

21.   Leach, C., et al. (2005) Assessing risk and emotional disturbance using the CORE-OM and HoNOS outcome measures at the interface between primary and secondary mental healthcare, *Psychiatric Bulletin*, 29(11): 419-422.

22.   Marks, I. (1998) Overcoming obstacles to routine outcome measurement *British Journal of Psychiatry*, 173: 281-286.

23.   Slade, M., Thornicroft, G., and Glover, G. (1999) The feasibility of routine outcome measures in mental health, *Social Psychiatry and Psychiatric Epidemiology*, 34(5): 243-249.

24.   Gilbody, S.M., House, A.O., and Sheldon, T.A. (2003) Outcome measures and needs assessment tools for schizophrenia and related disorders, *Cochrane Database of Systematic Reviews*, (1): CD003081.

25.   Gilbody, S.M., House, A.O., and Sheldon, T.A. (2001) Routinely administered questionnaires for depression and anxiety: systematic review, *British Medical Journal*, no 7283: Feb 17 2001.

26.   Gilbody, S.M., House, A.O., and Sheldon, T. (2002) Routine administration of Health Related Quality of Life (HRQoL) and needs assessment instruments to improve psychological outcome--a systematic review.[see comment], *Psychological Medicine*, 32(8): 1345-56.

27.   Ashaye, O., Livingston, G., and Orrell, M. (2003) Does standardized needs assessment improve the outcome of psychiatric day hospital care for older people? A randomized controlled trial, *Aging & Mental Health*, 7(3): 195-199.

28.   Slade, M., et al. (2006) Use of standardised outcome measures in adult mental health services: Randomised controlled trial, *British Journal of Psychiatry*, 189(4): 330-336.

29.   Priebe, S., et al. (2002) The impact of routine outcome measurement on treatment processes in community mental health care: approach and methods of the MECCA study, *Epidemiologia e Psichiatria Sociale*, 11(3): 198-205.

30.   Green, R.S. and Gracely, E.J. (1987) Selecting a rating scale for evaluating services to the chronically mentally ill, *Community Mental Health Journal*, 23: 91-102.

31.   Pirkis, J.E., et al. (2005) A review of the psychometric properties of the Health of the Nation Outcome Scales (HoNOS) family of measures, *Health & Quality of Life Outcomes*, 3(76).

32.   Evans, C., et al. (2002) Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM, *British Journal of Psychiatry*, 180: 51-60.

33.   Adams, M., et al. (2000) Health of the nation outcome scales for psychiatry: Are they valid?, *Journal of Mental Health*, 9(2): 193-198.

34.   Department of Health (1991) *Health of the Nation: A strategy for England*, HMSO: London.

35.   House of Commons (1990) *The National Health Service and Community Care Act*, HMSO: London.

36. Wing, J.K., et al. (1996) *Honos: health of the nation outcome scales: report on research and development: July 1993 - December 1995*, London: Royal College of Psychiatrists. pp116.

37. Glover, G., et al. (1997) The development of a new minimum data set for specialist mental health care, *Health Trends*, 29: 48-51.

38. Gowers, S.G., et al. (1999) Brief scale for measuring the outcomes of emotional and behavioural disorders in children: Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA), *British Journal of Psychiatry*, 174: 413-416.

39. Gowers, S., et al. (2000) The Health of the Nation Outcome Scales for Child & Adolescent Mental Health (HoNOSCA), *Child Psychology & Psychiatry Review*, 5(2): 50-56.

40. Burns, A., et al. (1999) Health of the Nation Outcome Scales for Elderly People (HoNOS 65+), *British Journal of Psychiatry*, 174: 424-427.

41. Page, A.C., Hooke, G.R., and Rutherford, E.M. (2001) Measuring mental health outcomes in a private psychiatric clinic: Health of the Nation Outcome Scales and Medical Outcomes Short Form SF-36, *Australian and New Zealand Journal of Psychiatry*, 35(3): 377-381.

42. Bech, P., et al. (2003) Health of the Nation Outcome Scales (HoNOS): Implementability, subscale structure and responsiveness in the daily psychiatric hospital routine over the first 18 months, *Nordic Journal of Psychiatry*, 57(4): 285-290.

43. Andreas, S., et al. (2007) A Psychometric Evaluation of the German Version of the 'Health of the Nation Outcome Scales, HoNOS-D': On the Feasibility and Reliability of Clinician-Performed Measurements of Severity in Patients with Mental Disorders, *Psychopathology*, 40(2): 116-125.

44. Gigantesco, A., et al. (2007) Discriminant Ability and Criterion Validity of the HoNOS in Italian Psychiatric Residential Facilities, *Psychopathology*, 40(2): 111-115.

45. Barbui, C., et al. (2002) Prescribing Second-Generation Antipsychotics and the Evolving Standard of Care in Italy, *Pharmacopsychiatry*, 35(6): 239-243.

46. Audin, K., et al. (2001) Value of HoNOS in assessing patient change in NHS psychotherapy and psychological treatment services, *British Journal of Psychiatry*, 178: 561-566.

47. McClelland, R., et al. (2000) Validation of an outcome scale for use in adult psychiatric practice, *Quality in Health Care*, 9(2): 98-105.

48. Brooks, R. (2000) The reliability and validity of the Health of the Nation Outcome Scales: Validation in relation to patient derived measures, *Australian and New Zealand Journal of Psychiatry*, 34(3): 504-511.

49. Trauer, T., et al. (1999) Health of the Nation Outcome Scales: Results of the Victorian field trial, *British Journal of Psychiatry*, 174: 380-388.

50. Trauer, T., Coombs, T., and Eagar, K. (2002) Training in routine mental health outcome assessment: the Victorian experience, *Australian Health Review*, 25(2): 122-8.

51. Coombs, T., Trauer, T., and Eagar, K. (2002) Training in routine mental health outcome assessment: an evaluation of the Victorian experience, *Australian Health Review*, 25(3): 74-82.

52. Rock, D. and Preston, N. (2001) HoNOS: Is there any point in training clinicians?, *Journal of Psychiatric and Mental Health Nursing*, 8(5): 405-409.

53. Wing, J., et al. (1998) Health of the Nation Outcome Scales (HoNOS): Research and development, *British Journal of Psychiatry*, 172: 11-18.

54. Shergill, S.S., et al. (1999) The validity and reliability of the Health of the Nation Outcome Scales (HoNOS) in the elderly, *Journal of Mental Health*, 8(5): 511-521.

55. Orrell, M., et al. (1999) Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community, *British Journal of Psychiatry*, 174: 409-412.

56. Bebbington, P., et al. (1999) Validation of the Health of the Nation Outcome Scales, *British Journal of Psychiatry*, 174: 389-394.

57. Trauer, T. (1999) The subscale structure of the Health of the Nation Outcome Scales (HoNOS), *Journal of Mental Health*, 8(5): 499-509.

58. Trauer, T. (2000) Comment, *Australian and New Zealand Journal of Psychiatry*, 34(3): 520-521.

59. Amin, S., et al. (1999) Evaluating the Health of the Nation Outcome Scales: Reliability and validity in a three-year follow-up of first-onset psychosis, *British Journal of Psychiatry*, 174: 399-403.

60. Trauer, T. and Callaly, T. (2002) Concordance between patients and their case managers using the Health of the Nation Outcome Scales (HoNOS), *Australasian Psychiatry*, 10(1): 24-28.

61. Parker, G., et al. (2002) Assessing outcome in community mental health patients: A comparative analysis of measures, *International Journal of Social Psychiatry*, 48(1): 11-19.

62. Ashaye, K., et al. (1999) Do the Health of the Nation Outcome Scales predict outcome in the elderly mentally ill? A 1-year follow-up study, *Journal of Mental Health*, 8(6): 615-620.

63. Goldney, R.D., Fisher, L.J., and Walmsley, S.H. (1998) The Health of the Nation Outcome Scales in psychiatric hospitalisation: A multicentre study examining outcome and prediction of length of stay, *Australian and New Zealand Journal of Psychiatry*, 32(2): 199-205.

64. Boot, B., Hall, W., and Andrews, G. (1997) Disability, outcome and case-mix in acute psychiatric in-patient units, *British Journal of Psychiatry*, 171: 242-6.

65. Lambert, G., Caputi, P., and Deane, F.P. (2002) Sources of information when rating the Health of the Nation Outcome Scales, *International Journal of Mental Health Nursing*, 11(2): 135-138.

66. Taylor, J.R. and Wilkinson, G. (1997) HoNOS v. GP opinion in a shifted out-patient clinic, *Psychiatric Bulletin*, 21(8): 483-485.

67. Hunter, R., et al. (2004) The Scottish 700 Outcomes Study: A Comparative Evaluation of the Health of the Nation Outcome Scale (HoNOS), the Avon Mental Health Measure (AVON), and an Idiographic Scale (OPUS) in adult mental health, *Journal of Mental Health*, 13(1): 93-105.

68. Gallagher, J. and Teesson, M. (2000) Measuring disability, need and outcome in Australian community mental health services, *Australian and New Zealand Journal of Psychiatry*, 34(5): 850-855.

69. Sharma, V.K., Wilkinson, G., and Fear, S. (1999) Health of the Nation Outcome Scales: A case study in general psychiatry, *British Journal of Psychiatry*, 174: 395-398.

70. Stafrace, S. (2002) Doubts about HoNOS, *Australian and New Zealand Journal of Psychiatry*, 36(2): 270.

71. Stein, G.S. (1999) Usefulness of the Health of the Nation Outcome Scales, *British Journal of Psychiatry*, 174: 375-7.

72. Andrews, G. and Page, A.C. (2005) Outcome measurement, outcome management and monitoring, *Australian and New Zealand Journal of Psychiatry*, 39(8): 649-651.

73. Ashaye, O., Mathew, G., and Dhadphale, M. (1997) A comparison of older longstay psychiatric and learning disability inpatients using the Health of the Nation Outcome Scales, *International Journal of Geriatric Psychiatry*, 12(5): 548-552.

74. Eagar, K., Trauer, T., and Mellsop, G. (2005) Performance of routine outcome measures in adult mental health care, *Australian and New Zealand Journal of Psychiatry*, 39(8): 713-718.

75. Glover, G. and Sinclair-Smith, H. (2000) Computerised information systems in English mental health care providers in 1998, *Social Psychiatry and Psychiatric Epidemiology*, 35(11): 518-522.

76. James, M. and Kehoe, R. (1999) Using the Health of the Nation Outcome Scales in clinical practice, *Psychiatric Bulletin*, 23(9): 536-538.

77. Brooker, C., et al. (1999) Evaluating clinical outcome and staff morale in a rehabilitation team for people with serious mental health problems, *Journal of Advanced Nursing*, 29(1): 44-51.

78. Royal College of Psychiatrists (2007) *What is HoNOS?* http://www.rcpsych.ac.uk/crtu/healthofthenation/whatishonos.aspx,

79. Gilbody, S.M., House, A.O., and Sheldon, T.A. (2002) Psychiatrists in the UK do not use outcomes measures: National survey, *British Journal of Psychiatry*, 180(2): 101-103.

80. Callaly, T., et al. (2006) Routine outcome measurement in public mental health: results of a clinician survey, *Australian Health Review*, 30(2): 164-73.

81. Meehan, T., et al. (2006) Introduction of routine outcome measures: staff reactions and issues for consideration, *Journal of Psychiatric & Mental Health Nursing*, 13(5): 581-7.

82. Aoun, S., Pennebaker, D., and Janca, A. (2002) Outcome measurement in rural mental health care: a field trial of rooming-in models, *Australian Journal of Rural Health*, 10(6): 302-7.

83. Trauer, T. (1998) The health of the nation outcome scales in outcome measurement: a critical review, *Australasian Psychiatry*, 6(1): 11-14.

84. Trauer, T., Callaly, T., and Herrman, H. (2005) *Evaluation of adult mental health services using routine outcome measures*, in Unpublished report.

85. Eisen, S.V., Dickey, B., and Sidered, L.I. (2000) A self-report symptom and problem rating scale to increase inpatients' involvement in treatment, *Psychiatric Services* 51(3): 349-353.

86. Mellor-Clark, J., et al. (2006) Resourcing a CORE Network to develop a National Research Database to help enhance psychological therapy and counselling service provision, *Counselling and Psychotherapy Research*, 6(1): 16-22.

87. Barkham, M., et al. (2006) A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System, *Counselling and Psychotherapy Research*, 6(1): 3-15.

88. Cahill, J., et al. (2006) Convergent Validity of the CORE Measures With Measures of Depression for Clients in Cognitive Therapy for Depression, *Journal of Counseling Psychology*, 53(2): 253-259.

89. Lyne, K., et al. (2006) Dimensions of variation on the CORE-OM, *British Journal of Clinical Psychology*, 45(2): 185-203.

90. Barkham, M., et al. (2005) The CORE-OM in an older adult population: Psychometric status, acceptability, and feasibility, *Aging & Mental Health*, 9(3): 235-245.

91. Barkham, M., et al. (2005) Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings, *British Journal of Psychiatry*, 186: 239-46.

92. Trusler, K., et al. (2006) Waiting times for primary care psychological therapy and counselling services, *Counselling and Psychotherapy Research*, 6(1): 23-32.

93. Connell, J., Grant, S., and Mullin, T. (2006) Client initiated termination of therapy at NHS primary care counselling services, *Counselling and Psychotherapy Research*, 6(1): 60-67.

94. Connell, J., et al. (2007) Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R, *British Journal of Psychiatry*, 190(1): 69-74.

95. Leach, C., et al. (2006) Transforming between Beck Depression Inventory and CORE-OM scores in routine clinical practice, *British Journal of Clinical Psychology*, 45(2): 153-166.

96. Mothersole, G. (2006) The use of CORE System data to inform and develop performance assessment and appraisal: An experiential account, *European Journal of Psychotherapy and Counselling*, 8(2): 177-191.

97. Evans, R.B., et al. (2006) Developing the resources and management support for routine evaluation in counselling and psychological therapy service provision: Reflections on a decade of CORE development, *European Journal of Psychotherapy and Counselling*, 8(2): 141-161.

98. Mellor-Clark, J. (2006) Developing CORE performance indicators for benchmarking in NHS primary care psychological therapy and counselling services: An editorial introduction, *Counselling and Psychotherapy Research*, 6(1): 1-2.

99. Bewick, B.M., et al. (2006) Routine outcome measurement completion rates of the CORE-OM in primary care psychological therapies and counselling, *Counselling and Psychotherapy Research*, 6(1): 33-40.

100. Dolan, P. (1997) Modelling valuations for EuroQol health states, *Medical Care*, 35: 1096-1108.

101. Bowling, A. (2001) *Measuring disease: a review of disease-specific quality of life measurement scales*, Buckingham: Open University Press. pp395.

102. Konig, H.H., Roick, C., and Angermeyer, M.C. (2007) Validity of the EQ-5D in assessing and valuing health status in patients with schizophrenic, schizotypal or delusional disorders, *European Psychiatry: The Journal of the Association of European Psychiatrists*, 22(3): 177-87.

103. Barton, G.R., Bankart, J., and Davis, A.C. (2004) Comparing utility scores before and after hearing-aid provision, *Applied health economics and health policy*, 3(2): 103-105.

104. Kind, P., et al. (1998) Variations in population health status: results from a United Kingdom national questionnaire survey, *British Medical Journal*, 316: 736-741.

105. Brazier, J., Roberts, J., and Deverill, M. (2002) The estimation of a preference-based single index measure for health from the SF-36, *Journal of Health Economics*, 21(2): 271-292.

106. Kaplan, R.M. and et al. (1998) The Quality of Well-Being Scale: critical similarities and differences with SF-36, *International Journal for Quality in Health Care*, 10(6): 509-520.

107. Lamers, L.M., et al. (2006) Comparison of EQ-5D and SF-6D utilities in mental health patients, *Health Economics*, 15(11): 1229-1236.

108. Sherbourne, C.D., et al. (2001) Can Utility-Weighted Health-Related Quality-of-Life Estimates Capture Health Effects of Quality Improvement for Depression?, *Medical Care*, 39(11): 1246-1259.

109. Gunther, O., et al. (2007) The EQ-5D in alcohol dependent patients: Relationships among health-related quality of life, psychopathology and social functioning, *Drug and Alcohol Dependence*, 86(2-3): 253-264.

110. van de Willige, G., et al. (2005) Changes in quality of life in chronic psychiatric patients: A comparison between EuroQol (EQ-5D) and WHOQoL *Quality of Life Research*, 14(2): 441-451.

111. Mangalore, R. (2000) The utility approach to valuing health states in schizophrenia, *Mental Health Research Review*, May: 11-15.

112. Brazier, J., Yang, Y., and Tsuchiya, A. (2007) *Review of methods for mapping between condition specific measures onto generic measures of health*, in Report prepared for the Office of Health Economcis, School of Health and Related Research, University of Sheffield: Sheffield.

113. Department of Health (2007) *Options for the Future of Payment by Results: 2008/09 to 2010/11*, in Consultation Document Annex B: Specific Services, Department of Health: London. pp. 23.

114. Elphick, M., et al. (1997) *Mental health: casemix outcome resources needs: summary report*. Img; G7083, Winchester - Information Management Group, NHS Executive Headquarters, Highcroft, Romsey Road Winchester SO22 5DH: The National Casemix Office. pp9.

115. Elphick, M., et al. (1997) *Mental health report: casemix outcome resources needs*. Img; G7080, Winchester - Information Management Group, NHS Executive Headquarters, Highcroft, Romsey Road Winchester SO22 5DH: The National Casemix Office. pp137.

116. The Care Packages and Pathways Project (2006) *Mental Health: Payment by Results*, in Final report version 10, The Care Packages and Pathways Project. pp62.

117. Department of Health (2006) *[Mental health services: project on implementation of outcomes measurement]*, Leeds: Department of Health. pp. [1]p.

118. Selbie, D. and Great Britain. Department of Health (2006) *Mental health services: project on implementation of outcomes measurement*. Dear Colleague Letter; Gateway reference 7279: (20 October 2006), London: Dh. pp. [2].

119. Persaud, A., Pearson, S., and Great Britain. Department of Health (2006) *Mental health services: project on implementation of outcomes measurement*. Dear Colleague Letter; Gateway reference 7280: (25 October 2006), London: Dh. pp. [1].

120. Berndt, E.R. (2003) *Changes in the costs of treating mental health disorders: An overview of recent research findings*, in Tufts European Center Workshop: Preserving Access by Protecting Innovation. Taillores, France.

121. Berndt, E.R., Busch, S.H., and Frank, R.G. (1998) Price indexes for acute phase treatment of depression, *National Bureau of Economic Research*, (Working Paper 6799).

122. Berndt, E.R., Busch, S.H., and Frank, R.G. (2001) *Treatment price indexes for acute phase major depression*, in *Medical Care Output and Productivity*, Cutler, D.M. and Berndt, E.R., Editors, University of Chicago Press: Chicago.

123. Berndt, E.R., et al. (2002) The medical treatment of depression, 1991-1996: Productive inefficiency, expected outcome variations, and price indexes, *Journal of Health Economics*, 21: 373-393.

124. Frank, R.G., Berndt, E.R., and Busch, S.H. (1998) Price indexes for the treatment of depression, *National Bureau of Economic Research*, (Working Paper 6417).

125. Frank, R.G., Busch, S.H., and Berndt, E.R. (1998) Measuring prices and quantities of treatment for depression, *American Economic Review*, 88(2): 106-111.

126. Frank, R.G., et al. (2004) Quality-constant prices for the ongoing treatment of schizophrenia: An exploratory study, *Quarterly Review of Economics and Finance*, 44(3): 390-409.

127. Frank, R.G., McGuire, T.G., and Normand, S.L. (2006) Cost-offsets of new medications for treatment of schizophrenia, *National Bureau of Economic Research*, (Working Paper 12643).

128. Ling, D.Y.C., Busch, A.B., and Frank, R.G. (2004) *Treatment price indexes for bipolar disorder*, in Working paper, Harvard Medical School: Harvard.

129. Berndt, E.R., Busch, S.H., and Frank, R.G. (2001) Interpreting changes in mental health expenditures: Minding our Ps and Qs, *National Bureau of Economic Research*, Research Summary(Winter 2000/2001).

130. Berndt, E.R., et al. (2006) Real output in mental health care during the 1990s, *Forum for Health Economics & Policy*, 9(1): 1-17.

131. Halsteinli, V., Kittelsen, S.A.C., and Magnussen, J. (2005) Productivity growth in Norwegian psychiatric outpatient clinics for children and youths *Journal of Mental Health Policy and Economics*, 8: 183-191.

132. Glover, G.R. (2000) A comprehensive clinical database for mental health care in England, *Social Psychiatry & Psychiatric Epidemiology*, 35(11): 523-9.