ARTICLE

Routine outcome monitoring in psychiatric clinical practice: background, overview and implications for person-centered psychiatry

M.S. van Noorden MD PhD[a], N.J.A. van der Wee MD PhD[b], F.G. Zitman MD PhD[c] and E.J. Giltay MD PhD[d]

a Psychiatrist, Leiden University Medical Center, Department of Psychiatry, Leiden, The Netherlands
b Psychiatrist, Leiden University Medical Center, Department of Psychiatry, Leiden and Leiden Institute for Brain and Cognition, Leiden, The Netherlands
c Psychiatrist, Leiden University Medical Center, Department of Psychiatry, Leiden, The Netherlands
d Psychiatrist, Leiden University Medical Center, Department of Psychiatry, Leiden, The Netherlands

Abstract

Routine Outcome Monitoring (ROM) is the systematic measurement of treatment outcomes in routine clinical practice. On the level of the patient and clinician, ROM may be a valuable source of information about patient's symptoms, treatment progress and psychosocial functioning. Furthermore, ROM can be used for purposes of research and benchmarking. The naturalistic character of the evaluations provide data that are more representative of 'real-world' patients than data derived from clinical trials. Despite these advantages, ROM has not been extensively implemented in psychiatry. The aim of the present qualitative review was to provide an overview of the conceptual background of the historical development, aims, methodological issues and potential applications of ROM in psychiatric outpatient settings and to consider their relevance to the development of person-centered psychiatry.

Keywords

Diagnosis, instruments, measurement, person-centered psychiatry, psychiatric epidemiology, psychiatry, randomized controlled trials, routine outcome monitoring

Correspondence address

Dr. Martijn S. van Noorden, Leiden University Medical Center, Department of Psychiatry, P.O. Box 7500, 2300 RC Leiden, The Netherlands. E-mail: m.s.van_noorden@lumc.nl

Accepted for publication: 31 July 2012

Introduction

Routine Outcome Monitoring (ROM) is the systematic measurement of treatment outcomes in routine clinical practice. ROM can be used as a tool for both the patient and the clinician in monitoring treatment progress. With ROM, depending on the choice of measurement instruments, detailed information about psychiatric diagnosis, several domains of symptoms and complaints and psychosocial functioning can be ascertained in every phase of treatment. Furthermore, on a group level, anonymised ROM data can be used for conducting epidemiological research, as well as for purposes of benchmarking.

ROM is a potentially important source of information regarding the *effectiveness* of treatment in daily or *real world* practice, in addition to the available information about *efficacy* of specific interventions derived from randomised controlled trials (RCTs) [1-6]. Despite these potential advantages, ROM has not yet been broadly implemented in psychiatry [7-10].

The aim of this qualitative review was to give an overview of the conceptual backgrounds of the historical development, aims, methodological issues and potential applications of ROM in patients with Mood, Anxiety and Somatoform (MAS) disorders.

Psychiatric diagnosis

An important precondition for a doctor to adequately treat an ill patient, is a reliable and valid diagnosis. This core criterion is true for all areas in medicine [11]. The study of symptoms and occurrence of diseases and hence the classification and definition of diagnoses, are within the scope of epidemiology. Ideally, knowledge of underlying pathophysiological disturbances is used for disease classification. Usually, a clinician gathers a medical history, physical examination and often laboratory tests and/or imaging tests to obtain a diagnosis [12]. Whenever a reliable and valid diagnosis has been established, a treatment plan can be proposed and informed consent of

the patient has to be obtained. After initiation, the effect of treatment has to be monitored. In theory, treatment effect can be measured in several domains: disease activity in terms of pathological processes or biological parameters, subjective symptoms as experienced by the patient, symptoms observed by the clinician (psychosocial) functioning and health-related quality of life [12,13].

Despite major research efforts in psychiatry during the past decades, knowledge about the pathophysiological mechanisms underlying most psychiatric disorders is still limited. This is in contrast with many somatic disorders, where large breakthroughs in understanding of pathophysiology have been accomplished. This lack of knowledge about pathophysiology of aetiology of psychiatric disorders has implications for both diagnosis and monitoring of treatment effect. Firstly, the value of laboratory tests and other biomarkers in psychiatric diagnostics in the individual patient is merely marginal [14]. The psychiatrist uses medical history taking, that is, the patient's report of internal phenomena and the systematic mental-state examination to ascertain the symptoms and complaints of the patient. Instead of laboratory or imaging tests, rating scales that measure psychopathology can be applied. Secondly, the monitoring of treatment effect is limited to standardised rating of symptoms and psychosocial functioning, because at present, no biological parameters (i.e., biomarkers) can be used as measures of disease activity. However, as mentioned above, monitoring outcomes on a routine and standardised basis with ROM has not yet become standard practice in psychiatry. The various reasons for this lack of implementation will be discussed later in this paper.

Until 1980, no well-defined, international accepted diagnostic criteria existed in psychiatry [15]. The need for reliable and valid diagnoses urged the American Psychiatric Association (APA) to introduce the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) in 1980 [16]. This document was based on validity field trials in the United States (US) and consensus of an APA task force and, contrary to the first two editions (DSM-I and DSM-II), it comprised detailed descriptions of symptom clusters and diagnostic criteria of psychiatric disorders. Since the introduction of DSM-III, psychiatric disorders are being classified based on the presence of symptoms, providing syndromal diagnoses. This is exactly what the influential German psychiatrist Emil Kraepelin proposed almost a century earlier [15]. The introduction of the DSM-III has caused a revolution in psychiatry as it dramatically increased the possibilities of conducting epidemiological research with results that were internationally applicable. The current version of the DSM, the DSM-IV (introduced in 1994 with a text revision in 2000; DSM-IV-TR), is the result of ongoing epidemiological research and consensus [17,18]. The publication of the DSM-V is due in 2013 (www.DSM5.org).

Measurement in psychiatry

In order to measure or classify psychiatric disorders, psychiatric symptoms preferably have to be assessed in an objective and reproducible, standardised manner [19]. Since most psychiatric symptoms have a large subjective component (e.g., delusive thoughts, hallucinations, disturbed mood and somatic sensations), objective measurement is a challenge. These symptoms are not easily observed or verified by an examiner [20]. The need for objective measurement of psychiatric symptoms has resulted in the development of psychometrics: the science of psychological assessment. A psychometric test is an instrument designed to produce a quantitative assessment of some psychological attribute(s). According to the psychometric principles, a psychometric test should be valid, reliable and free of bias [21]. Validity indicates that the test assesses the true state of the phenomenon being measured, reliability refers to the extent of reproducibility of the test and bias is a systematic error in the design of the test or study or in data analysis. For use in ROM, a test should also be sensitive to clinically important change over time [13]. Measurement in psychiatry can take place on the level of syndromal diagnosis, on the level of symptom severity, on the level of psychosocial functioning and on the level of health related quality of life. Ideally, a ROM test battery consists of measurement instruments that cover all these levels [8].

DSM Diagnostic measurement instruments

Syndromal classifications are potentially less fundamental than classifications that make use of clearly disturbed biological etiological processes, for example, the detection of tumor cells in cancer or the occurrence of a pathogen in infectious diseases. Nevertheless, the introduction of the DSM-II and its successive worldwide use has greatly facilitated the development of structured diagnostic measurement instruments, necessary for psychiatric epidemiologic research. Until the 1980's, psychiatric epidemiology was hampered by methodological shortcomings, most importantly because of fuzzy definitions of diagnoses and outcomes [22].

The 1980 DSM-III criteria were used for the development of the Diagnostic Interview Schedule (DIS), for use in the first large US community epidemiologic study on mental health: the Epidemiological Catchment Area (ECA) study, sponsored by the US National Institute of Mental Health (NIMH). This structured interview could be administered by lay interviewers because of its closed-ended questions that did not require clinical judgment [22,23]. Some years later, the Composite International Diagnostic Interview (CIDI) was developed in collaboration with the World Health Organisation (WHO) [24]. A modified version of the CIDI was used in the next large US community epidemiological study: the National Comorbidity Survey (NCS) [25]. After the development of these structured diagnostic interviews, reliable prevalence estimates of psychiatric disorders were possible. Because of the extensive format of the CIDI, which limited use in clinical practice, Lecrubier and colleagues developed in a European-US collaboration a short validated structured diagnostic instrument: the MINI International Neuropsychiatric Interview Plus (MINI-Plus). The MINI

plus was validated *versus* the CIDI with satisfactory results [26].

Symptomatic and functional measurement instruments

While diagnostic measurement instruments measure DSM diagnosis in a standardised manner, symptomatic and functional measurement instruments measure symptom severity and health status on a functional level. The latter two categories of instruments can be regarded as monitoring instruments, which may be applied at several time-points during treatment to evaluate progress of treatment and disease. Symptom-based scales may be generic or disorder-specific and self-report or observer-rated.

The Hamilton Depression Rating Scale (HDRS) and Montgomery-Åsberg Depression Rating Scale (MADRS) are examples of well-known disorder-specific, observer-rated rating scales that measure symptom severity in major depression [27,28]. Of these two scales, the HDRS has been predominantly used in RCTs, but the MADRS seems superior for outpatient use [29]. The Brief Symptom Inventory (BSI) is an example of a generic self-report rating scale that measures psychopathological symptom severity in several domains, for example, anxiety, depression, hostility and somatic complaints [30,31].

Domains of functional monitoring instruments include general health status, quality of life (QoL) and social and role functioning. These instruments are ordinarily regarded as QoL scales, although a distinction between Qol and (psychosocial) functioning can be made. An example of a widely used instrument that measures general health status is the self-report Short Form-36 (SF-36) [32].

Psychiatric epidemiology

The large-scale epidemiological community studies facilitated by the sophistication of psychiatric diagnosis and the subsequent development and validation of comprehensive diagnostic measurement instruments have provided valuable data on prevalence and incidence of psychiatric disorders in the general population. Furthermore, these studies have described in great detail clinical characteristics and correlates of most psychiatric disorders. A disadvantage of these epidemiological studies, however, is the fact that the responding subjects, do not necessarily reflect treatment seeking populations, even when they meet criteria for psychiatric disorders. This may limit generalisability or external validity of these findings to the daily clinical practice in psychiatric specialty care. Nevertheless, these studies have played a major role in the development of the field of psychiatric epidemiology and the development of current psychiatry. Psychiatric epidemiology can be defined as "the study of the distribution and determinants of mental disorders in specified populations and of the risk factors associated with their onset and course" [22]. In analogy with epidemiology in somatic medicine, psychiatric epidemiology can be subdivided in *community* psychiatric epidemiology and *clinical* psychiatric epidemiology.

In contrast with *community epidemiology*, which aims to describe disease phenomena and to estimate prevalence rates of diseases in the general population, the main goals of *clinical epidemiology* are to investigate the effects of presumed causal risk factors on the onset and course of illness in clinical patients, to evaluate the validity of diagnostic tests and to study predictors of treatment response that might be targeted in subsequent interventions [22,33]. In the last three decades, descriptive community psychiatric epidemiologic research, with the ECA and NCS as examples, has prospered. On the other hand, clinical psychiatric epidemiology has remained under-developed as compared to clinical epidemiology in other fields of medicine [33]. This is mostly because of the fundamental problem of establishing psychiatric diagnoses (assessment of caseness) as compared with somatic diagnoses, because of the limited validity of most psychiatric diagnoses. Another reason for this difference is the fact that the treatment of psychiatric disorders is diverse - despite the availability of evidence-based guidelines - making it more difficult to conduct clinical epidemiological research of naturalistic variation in treatment response [33]. Finally, because many patients with psychiatric disorders do not seek treatment, representative descriptive data of psychiatric disorders in the general population are not necessarily applicable to everyday patients in clinical practice [33,34].

Descriptive community epidemiological studies like the ECA and NCS have yielded valuable insights in prevalence rates and phenomenology of psychiatric disorders in the general population. Since these US community studies in the 1980's, replications have been conducted (National Comorbidity Survey-replication; NCS-R), as well as community studies in Europe. The first and second Netherlands Mental Health Survey and Incidence Study (NEMESIS 1 and 2) are examples of the latter [35,36]. The prevalence rates of most psychiatric disorders appear to be quite consistent over time and across continents (Table 1).

Contrary to community epidemiological studies, clinical psychiatric epidemiological studies have the important potential of evaluating interventions in daily clinical practice. Kessler [33] stated that "In addition to studying the aggregate magnitude of treatment effects, clinical epidemiological studies are needed to study the predictors of individual differences in treatment response". This type of work would ideally involve investigating baseline (i.e., as from the onset of treatment) predictors of course of illness in broadly representative clinical samples. Examples of large scale, truly naturalistic studies, are scarce nowadays. In order to conduct these studies, a ROM infrastructure could be used, in which outcome data of large naturalistic samples are collected.

ROM and epidemiological research

A well-implemented ROM infrastructure could provide anonymised data for epidemiological research in treatment-seeking patients with certain disorders. Baseline

Table 1 Benefits, current and possible future applications of ROM

Benefits of ROM
For patient
Detailed diagnosis and feedback/monitoring treatment progress is possible
In the case of poor response, problem areas can be identified
For clinician
Complementary, standardised information in addition to clinical judgement
Tool for providing feedback
Focus on problem areas
Allows easy interpretation by different clinicians
For institution
Benchmarking between departments or institutions
For research
Large datasets allowing clinical epidemiological research
Minimal selection criteria ensuring high generalisability to real-world practice
Current and possible future applications of ROM
Outcome and implementation studies
Development of risk profiles of poor outcome
CER in subgroups of patients (e.g. extra interventions in high-risk patients)
Guideline implementation and effect on outcome
Biological studies
Add-on research, e.g. biobanking (MASHBANK)
Psychometric studies
Translation of scales into other languages
Development and validation of freely available measurement scales
Calculation reference values for measurement instruments
Other applications
Expansion to patient groups with other diagnoses than MAS, e.g. SMI

Abbreviations: CER: Comparative effectiveness research; MASHBANK: Mood, Anxiety, Somatoform disorders and the Hypothalamus pituitary adrenal (HPA) axis Biobank; SMI Severe Mental Illness.

ROM data could be used for cross-sectional analyses. For example, clinical characteristics and comorbidity patterns of certain disorders could be investigated in 'real-world' patients [5,6,37]. Another possible application of cross-sectional ROM data is to use these data to assess the reliability and validity of measurement instruments in clinical practice [38].

Prospective ROM data could be used to investigate predictors of treatment outcome in daily clinical practice and to identify risk factors for poor outcome in real-world treatment settings.

Randomised Controlled Trials

The development of psychopathology measurement instruments several decades ago has also dramatically increased the possibilities of evaluating treatment efficacy by means of clinical trials. RCTs have widely been accepted as the gold standard in evaluating treatment efficacy in medicine [39,40]. For a new drug to be approved by the regulatory authorities, superior efficacy compared to placebo in RCTs is required. Indeed, the design of a typical RCT, in which 2 or more specific interventions with or without a placebo condition are directly compared in a double-blind way in a sample of patients with a specific disease, aims to maximise internal validity of the trial at hand. In other words, whenever an effect is found, it is most likely being explained by the intervention under study because confounding is largely eliminated through the randomisation process. This high level of internal validity can only be realised if both the disease under study is strictly defined, if the intervention is strictly defined and if the sample is homogeneous in terms of comorbidity and other clinical characteristics. This means that often a large and strict set of inclusion and exclusion criteria are being used in RCTs. If those strict conditions are met and the sample is large enough to detect clinically meaningful differences, in theory an RCT with maximised *internal validity* will provide the strongest possible evidence for superiority of a certain intervention [41].

However, in the real world, those perfect 'laboratory circumstances' do not exist. The RCT populations usually are highly selected and suffer from limited *external validity*. Yet, most evidence-based treatments in medicine are largely based on findings from RCTs. Of course, the findings from RCTs are a major leap forward in terms of 'evidence-based medicine' as compared to the mere descriptive 'clinical expertise' and case studies from the pre-RCT era.

A major disadvantage of RCTs, however, is the lack of external validity and generalisability [42-44]. In a recent study of our group we found that only 20-25% of our depressive outpatients would meet general inclusion criteria for RCT's [45]. In other words, 'real world' patients most likely differ from RCT patients. Apart from evidence about treatment efficacy, RCTs have also played a role in our knowledge about characteristics of psychiatric disorders (e.g., symptom profiles, comorbidity patterns). Typically, symptomatology of specific disorders has been analysed using only the baseline measurements in large RCT populations and reports about these clinical characteristics are being published secondary to the main paper describing the primary outcome of the intervention (see for example Marcus *et al.*, 2005 and Zisook *et al.*, 2007) [46,47].

Routine Outcome Monitoring

Historical perspective of Routine Outcome Monitoring

The limited generalisability of findings from RCTs and population studies to daily clinical practice and the lack of insight in processes and patient's experiences of treatment inspired Ellwood for his 1988 Shattuck lecture in which he pleaded for 'assessing routinely and frequently the health of patients using appropriate reliable and valid measurement instruments and to build large databases with these data' [1]. He predicted 'a new revolution in healthcare' and stimulated the systematic assessment of clinical, financial and health outcomes. Although this idea was well received in editorials [2,48], recent reviews have

shown only a limited number of published studies of routinely assessed outcomes or ROM in psychiatric specialty care [7]. Institutions that have adopted ROM usually used a slim test battery [49,50]. Several reasons for this lack of routine implementation of ROM in clinical practice are proposed: ROM is costly and time consuming and requires a relatively complicated technical infrastructure. More important, no consensus exists about the optimal choice of measurement instruments, so that outcomes are not easily comparable across clinics and across studies. Probably, parameters like treatment setting, patient population and the limited availability of measurement instruments free of copyright may contribute to this lack of consensus. Furthermore, the aim of ROM may vary, as several parties have different interests, for example, policy makers, insurance companies, patients, clinicians and researchers [7,51].

Aims and methodological issues of Routine Outcome Monitoring

In theory, ROM can provide both clinician and patient with valuable information about symptoms and treatment outcomes in daily clinical practice and *effectiveness* of treatments in 'real world' treatment settings. Evidence-based treatments are based on efficacy trials and may not be effective for every patient in the 'real world'. The main aim of ROM is improvement of the quality of patient care by measuring progress and giving feedback to the patient. Secondary aims of ROM are understanding mechanisms of disease and treatment, establishing cost effectiveness and benchmarking. For understanding the relationship between patients' health status (outcomes), disease status and treatment (process of care) it is necessary to have access to detailed information about the type of treatment [13].

If observer-rated measurement scales are being used, it is important that the interviewer is well trained because clinical interpretation of symptoms or complaints is essential for reliable and reproducible results. To increase objectivity, preferably, measurement instruments are applied by an interviewer who is not directly involved in the treatment of the patient. In addition, inter-rater variability between interviewers should be minimised by recurrent training sessions in which calibration takes place.

Ideally, ROM measurement instruments should be clinically relevant, sensitive to change, minimally burdensome to the patient, to the staff and to the institution in terms of costs of collection and data analysis [52]. This implies that a balanced selection of well-validated measurement instruments free of copyright is to be preferred. Given the fact that a substantial proportion of patients speak foreign languages, it is important that questionnaires are being validated in different languages. It is evident that different instruments may be applied in different patient groups. In the international literature, no consensus exists about the choice of measurement instruments, about the interval of measurement and about the groups of patients or treatment settings in which ROM may be applied.

Since the data gathering in ROM is naturalistic and observational, no experimental designs can be used if outcome data are routinely assessed. Hence, instead of causal inferences, only correlations can be established on group level [33]. Another methodological issue when analysing ROM data is the problem of confounding and selection bias, since the treatment that a patient receives will often be determined by a number of factors that are related to outcome, such as disease severity [9].

The use of patient-based measures of health may itself be useful in improving treatment outcomes, because of the possibility to provide feedback of ROM assessments to both patient and clinician. This may enable clinicians to detect problem areas in treatment that would have been missed without the use of data derived from ROM [7,53] and may increase patient's compliance to treatment protocols [7,54]. A limited amount of studies have demonstrated a positive impact of ROM on monitoring treatment and on the quality of communication between clinician and patient [7,55]. In the meta-analysis by Carlier *et al*. [7], a favourable outcome of feedback by ROM on mental health was found on the short term only.

ROM and Comparative Effectiveness Research

In the literature, research based on ROM data is often regarded as 'patient-centered research'. If treatment details are taken into account, ROM-data driven research could be used for Comparative Effectiveness Research (CER). CER is designed to improve the clinical decision-making process by providing research evidence on the effectiveness and risk-benefit profile on different therapeutic options for specific patient subpopulations [56,57]. In the US, the Patient-Centered Outcomes Research Institute (PCORI) has been established to facilitate CER. The mission of this Institute is to help people make informed healthcare decisions and to improve healthcare delivery and outcomes [58]. In Europe, to our knowledge, no comparable large-scale initiatives are being developed. In theory, ROM databases of different institutions could be merged in large collaboration efforts and used for CER. For such overarching initiatives to be successful, many consensus steps have still to be taken.

ROM critically appraised

When collected systematically and extensively, ROM data may be valuable for purposes of benchmarking. In theory, it would be possible to gain insight in treatment results of organizations, departments or even at the individual therapist level. Policy makers and health insurance companies have discovered ROM as a source of benchmarking data. However, a potential limitation or pitfall of ROM is important to consider. In this modern era of excessive growth of healthcare and inevitable health costs, the power of health insurance companies is growing

and the professional autonomy of medical specialists is increasingly under pressure. Understandably, policy makers like health insurance companies and governments demand more and more insight in costs of treatments and treatment processes to be able to control these costs. Since ROM is a potentially valuable source of information regarding these processes, many health providers have implemented ROM initiatives over the past years, with benchmarking as one of the major goals. However, benchmarking based on ROM assessments is possible, but may be a hazardous operation. First of all, some institutions have adapted a ROM system in which only a succinct set of outcome scales or only one scale is used. It would be hard to derive reliable benchmarking data because of important inter-patient differences that require complex statistics to be taken into account. For example, if in a certain clinic more MDD patients with comorbid personality disorders are being treated, outcomes may be worse compared to another clinic that uses the same guidelines, but where patients with less complicated complaints seek treatment. Our major concern would be that in the case of a limited ROM assessment battery, policy makers will draw conclusions based on insufficient or inadequately analysed data. For example, in tertiary care clinics or specialised secondary care clinics, typically patients with treatment-resistant complaints, somatic comorbidity, co-existent personality pathology or a combination of these are being treated. Those patients are likely to have worse treatment-outcomes, irrespective of the quality of treatment, as compared with less complicated patients in general psychiatric specialty care.

Future Perspectives

Apart for clinical epidemiological research, ROM data can serve as basis for research in other domains. Examples of these are biological and psychometric research. To further illustrate the potential of ROM, we will give some examples of current projects that use the Leiden ROM infrastructure.

First, the lack of well-known biological markers in psychiatric disorders complicates the borders of disease. When can someone with certain complaints be classified as a patient? The diagnostic classification system DSM-IV only partially answers that question by operationalising disorders by consensus definitions. Due to the absence of clear markers and borders, the line between ‘healthy’ and ‘sick’ will be hard to define. Validated measurement scales are helpful in defining and establishing that line. However, for many validated measurement scales used in ROM no reference values in the general population have been calculated. NormQuest is a study initiated in Leiden that aims to assess those reference values for the commonly used measurement scales in MAS patients [59].

Second, the ROM infrastructure with naturalistically obtained data also allows for add-on research. The Mood, Anxiety, Somatoform disorders and Hypothalamus pituitary adrenal (HPA) axis Biobank (MASHBANK) is an example of this type of research. The MASHBANK has been founded to investigate the link between genetic variants, functioning of the HPA axis and the phenotype in patients with MAS disorders and comparing those with patients from the general population. Patients routinely enrolled in ROM have been asked informed consent to donate DNA for the MASHBANK, after CME approval of the protocol. So far, almost 2000 samples of MAS patients and control subjects have been collected [59]. Additional examples of benefits, current and possible future applications are shown in Table 1.

Large-scale collaborations could result in development of risk-assessment based on CER and integration of biological markers in ROM. Although practical obstacles may have to be faced, lessons from the cardiovascular field for example, Framingham study [60], but also oncology demonstrate that large-scale collaboration may dramatically improve outcomes step by step in large groups of patients. For example, acute leukemia is the most common form of childhood cancer, comprising approximately 30% of all malignancies in children. Survival rates for Acute Lymphatic Leukemia have increased dramatically since the 1980s, with current 5-year overall survival rates of over 85% [61-63]. These improved survival rates are due to large-scale collaborations and RCTs of treatment of large groups of patients according to standardised research protocols and constant monitoring of outcomes. These protocols have evolved over and over according to outcomes of trials and findings of more biological studies [63,64]. In psychiatry, the establishment of the international schizophrenia consortium (ISC) has resulted in large-scale genetic studies in schizophrenia and bipolar disorder (e.g., Purcell *et al*., 2009) [65]. In Europe, the GENDEP consortium aims to use genetic profiles to predict outcome of antidepressant treatment (e.g., Uher, *et al*., 2010) [66]. These initiatives demonstrate that large-scale collaborations are possible.

Conclusions

Routine Outcome Monitoring, the systematic measurement of treatment outcomes in clinical practice, is a potentially valuable source of information about patient characteristics, disease characteristics and psychosocial functioning of psychiatric patients in ‘real world’ treatment settings. Although primarily used by patient and clinician for evaluating treatment progress, on an aggregated level ROM data can also be used for clinical epidemiological research and for benchmarking. Since ROM data are gathered routinely, findings are more representative of ‘real world’ patients than data derived from RCTs. In clinical psychiatry, large-scale ROM initiatives are still scarce, even though the standardized assessment of diagnosis and symptoms with validated measurement instruments may provide objectivity in diagnosis and treatment evaluation. Implementation of ROM provides multiple opportunities for research and improvement of patient outcomes and contributes actively to person-centered psychiatry.

References

[1] Ellwood, P.M. (1988). Shattuck lecture--outcomes management. A technology of patient experience. *New England Journal of Medicine* 318, 1549-1556.

[2] Holloway, F. (2002). Outcome measurement in mental health--welcome to the revolution. *British Journal of Psychiatry* 181, 1-2.

[3] Relman, A.S. (1988). Assessment and accountability: the third revolution in medical care. *New England Journal of Medicine* 319, 1220-1222.

[4] Zimmerman, M., Ruggero, C.J., Chelminski, I., Young, D., Posternak, M.A., Friedman, M., Boerescu, D. & Attiulah, N. (2006). Developing brief scales for use in clinical practice: the reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. *Journal of Clinical Psychiatry* 67, 1536-1541.

[5] van Noorden, M.S., Giltay, E.J., den Hollander-Gijsman, M.E., van der Wee, N.J. , van Veen, T. & Zitman, F.G. (2010). Gender differences in clinical characteristics in a naturalistic sample of depressive outpatients: The Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders* 125, 116-123.

[6] van Noorden, M.S., Minkenberg, S.E., Giltay, E.J., den Hollander-Gijsman, M.E., van Rood, Y.R., van der Wee, N.J. & Zitman, F.G. (2011). Pre-adult versus adult onset major depressive disorder in a naturalistic patient sample: the Leiden Routine Outcome Monitoring Study. *Psychological Medicine* 41, 1407-1417.

[7] Carlier, I.V., Meuldijk, D., van Vliet, I.M., van Fenema, E.M., van der Wee, N.J. & Zitman, F.G. (2012). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice* 18, 104-110.

[8] de Beurs, E., den Hollander-Gijsman, M.E., van Rood, Y.R., van der Wee, N.J., Giltay, E.J., van Noorden, M.S., van der Lem, R., van Fenema, E. & Zitman, F.G. (2011). Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology and Psychotherapy* 18, 1-12.

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

[10] Slade, M. (2002). What outcomes to measure in routine mental health services, and how to assess them: a systematic review. *Australia and New Zealand Journal of Psychiatry* 36, 743-753.

[11] Goodwin, D. & Guze, S. (1996). Psychiatric diagnosis. Fifth edition. New York: Oxford University Press.

[12] Fauci, A.S., Braunwald, E., Kasper, D.L., Hauser, S.L., Longo, D.L., Jameson, D.L., et al. (2008). The Practice of Medicine. 17th edition. New York: McGraw-Hill.

[13] Smith, G.R., Jr., Manderscheid, R.W., Flynn, L.M. & Steinwachs, D.M. (1997). Principles for assessment of patient outcomes in mental health care. *Psychiatric Services* 48, 1033-1036.

[14] Quinones, M.P. & Kaddurah-Daouk, R. (2009). Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. *Neurobiology of Disease* 35, 165-176.

[15] Mayes, R. & Horwitz, A.V. (2005). DSM-III and the revolution in the classification of mental illness. *Journal of the History of the Behavioural Sciences* 41, 249-267.

[16] American Psychiatric Association. (1980). Diagnostic and statistical Manual of Mental Disorders. Third edition. Washington, DC: American Psychiatric Association.

[17] American Psychiatric Association. (1994). Diagnostic and Statistical Manual of Mental Disorders. Fourth edition. Washington, DC: American Psychiatric Association.

[18] American Psychiatric Association. (2000) Diagnostic and Statistical Manual of Mental Disorders. Fourth edition-Text revision. Washington, DC: American Psychiatric Association.

[19] Zitman, F.G. (1990) Standaardisering van psychiatrische diagnostiek. In: Diagnostiek in de psychiatrie, mogelijheden en grenzen. (Abraham, R.E., Giel, R., Rooijmans, H.G.M. & Zitman, F.G. eds.). Leiden: Boerhaave Commissie.

[20] Gilbody, S., House, A.O. & Sheldon, T.A. (2003). Outcomes measurement in psychiatry: a review of outcomes management in psychiatric research and practice. Centre for Reviews and Dissemination report 24. University of New York, New York.

[21] Ishak, W.W., Burt, T. & Sederer, L.I. (2002). Outcome Measurement in Psychiatry: a critical review. Washington DC: American Psychiatric Association.

[22] Tohen, M., Bromet, E., Murphy, J.M. & Tsuang, M.T. (2000). Psychiatric epidemiology. *Harvard Review of Psychiatry* 8, 111-125.

[23] Robins, L.N., Helzer, J.E., Croughan, J. & Ratcliff, K.S. (1981). National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry* 38, 381-389.

[24] Robins, L.N., Wing, J., Wittchen, H.U., Helzer, J.E., Babor, T.F., Burke, J. et al. (1988). The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* 45, 1069-1077.

[25] Kessler, R.C., Mcgonagle, K.A., Zhao, S.Y., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U. & Kendler, K.S. (1994). Lifetime and 12-Month Prevalence of Dsm-Iii-R Psychiatric-Disorders in the United-States - Results from the National-Comorbidity-Survey. *Archives of General Psychiatry* 51, 8-19.

[26] Lecrubier, Y. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry* 12, 224.

[27] Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry* 23, 56-62.

[28] Montgomery, S.A. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* 134, 382.

[29] Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., Mors, O., Elkin, A., Williamson, R.J., Schmael, C., Henigsberg, N., Perez, J., Mendlewicz, J., Janzing, J.G., Zobel, A., Skibinska, M., Kozel, D., Stamp, A.S., Bajs, M., Placentino, A., Barreto, M., McGuffin, P. & Aitchison, K.J. (2008). Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychological Medicine* 38, 289-300.
[30] de Beurs, E. & Zitman, F.G. (2006). The Brief Symptom Inventory (BSI): Reliability and validity of a practical alternative to SCL-90. *MGV*, 61, 120-41.
[31] Derogatis, L.R. & Melisaratos, N. (1983). The Brief Symptom Inventory: an introductory report. *Psychological Medicine* 13, 595-605.
[32] Ware, J.E., Jr. & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care* 30, 473-483.
[33] Kessler, R.C. (2007). Psychiatric epidemiology: challenges and opportunities. *International Reviews in Psychiatry* 19, 509-521.
[34] Burger, H. & Neeleman, J. (2007). A glossary on psychiatric epidemiology. *Journal of Epidemiology and Community Health* 61, 185-189.
[35] Bijl, R.V., Ravelli, A. & van Zessen, G. (1998). Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Social Psychiatry and Psychiatric Epidemiology* 33, 587-595.
[36] de Graaf, R., Ten Have, M. , van Gool, C. & van Dorsselaer, S. (2012). Prevalence of mental disorders and trends from 1996 to 2009. Results from the Netherlands Mental Health Survey and Incidence Study-2. *Social Psychiatry and Psychiatric Epidemiology* 47 (2) 203-213.
[37] De Klerk, S., van Noorden, M.S., van Giezen, A.E., Spinhoven, P., den Hollander-Gijsman, M.E., Giltay, E.J., Speckens, A.E. & Zitman, F.G. (2011) Prevalence and correlates of lifetime deliberate self-harm and suicidal ideation in naturalistic outpatients: the Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders* 133, 257-264.
[38] Grootenboer, E.M., Giltay, E.J., van der Lem, R., van Veen, T., van der Wee, N.J. & Zitman, F.G. (2012) Reliability and validity of the global assessment of functioning scale in clinical outpatients with depressive disorders. *Journal of Evaluation in Clinical Practice* 18, 502-507.
[39] Atkins, D., Best, D., Briss, P.A., Eccles, M., Falck-Ytter, Y., Flottorp, S., Guyatt, G.H., Harbour, R.T., Haugh, M.C., Henry, D., Hill, S., Jaeschke, R., Leng, G., Liberati, A., Magrini, N., Mason, J., Middleton, P., Mrukowicz, J., O'Connell, D., Oxman, A.D., Phillips, B., Schünemann, H.J., Edejer, T.T., Varonen, H., Vist, G.E., Williams, J.W., Zaza, S. & GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *British Medical Journal* 328 (7454) 1490.
[40] Kaptchuk, T.J. (2001). The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? *Journal of Clinical Epidemiology* 54, 541-549.
[41] Rorty, R. (1977). Philosophy and the mirror of nature. Princeton: Princeton University Press.
[42] Licht, R.W., Gouliaev, G., Vestergaard, P. & Frydenberg, M. (1997). Generalisability of results from randomised drug trials. A trial on antimanic treatment. *British Journal of Psychiatry* 170, 264-267.
[43] Wells, K.B. (1999). Treatment Research at the Crossroads: The Scientific Interface of Clinical Trials and Effectiveness Research. *American Journal of Psychiatry* 156, 5.
[44] Zimmerman, M. (2002). Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *American Journal of Psychiatry* 159, 469-473.
[45] van der Lem, R., van der Wee, N.J., van Veen, T. & Zitman, F.G. (2010). The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychological Medicine* 16, 1-11.
[46] Marcus, S.M., Young, E.A., Kerber, K.B., Kornstein, S., Farabaugh, A.H., Mitchell, J., Wisniewski, S.R., Balasubraman, G.K., Trivedi, M.H. & Rush, A.J. (2005). Gender differences in depression: findings from the STAR*D study. *Journal of Affective Disorders* 87, 141-150.
[47] Zisook, S., Rush, A.J., Lesser, I., Wisniewski, S.R., Trivedi, M., Husain, M.M., Balasubraman, G.K., Alpert, J.E. & Fava, M. (2007). Preadult onset vs. adult onset of major depressive disorder: a replication study. *Acta Psychiatrica Scandinavica* 115, 196-205.
[48] Slade, M. (2002). Routine outcome assessment in mental health services. *Psychological Medicine* 32, 1339-1343.
[49] Burgess, P.M., Pirkis, J.E., Slade, T.N., Johnston, A.K., Meadows, G.N. & Gunn, J.M. (2009). Service use for mental health problems: findings from the 2007 National Survey of Mental Health and Wellbeing. *Australia and New Zealand Journal of Psychiatry* 43, 615-623.
[50] Lambert, M.J., Hansen, N.B. & Finch, A.E. (2001). Patient-focused research: using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology* 69, 159-172.
[51] Norquist, G.S. (2002). Role of outcome measurement in psychiatry. In: Outcome measurement in psychiatry. First ed. (Ishak, W.W., Burt, T. & Sederer, L.I., eds.). Washington, DC: American Psychiatric Publishing.
[52] Dickey, B. (2002). Outcome measurement from research to clinical practice. In: Outcome measurement in psychiatry. First ed. (Ishak, W.W., Burt, T. & Sederer, L.I., eds.). Washington, DC: American Psychiatric Publishing.
[53] Greenhalgh, J. & Meadows, K. (1999). The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *Journal of Evaluation in Clinical Practice* 5, 401-416.
[54] Hysong, S.J. (2009). Meta-analysis: audit and feedback features impact effectiveness on care quality. *Medical Care* 47, 356-363.
[55] Knaup, C., Koesters, M., Schoefer, D., Becker, T. & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *British Journal of Psychiatry* 195, 15-22.

[56] Mane, K.K., Bizon, C., Schmitt, C., Owen, P., Burchett, B., Pietrobon, R. & Gersing, K. (2012) VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *Journal of Biomedical Informatics* 45 (1) 101-106.
[57] Sox, H.C. & Greenfield, S. (2009). Comparative effectiveness research: a report from the Institute of Medicine. *Annals of Internal Medicine* 151, 203-205.
[58] Washington. A.E. & Lipstein, S.H. (2011). The Patient-Centered Outcomes Research Institute - Promoting Better Information, Decisions, and Health. *New England Journal of Medicine* 365 (15) e31.
[59] Schulte-van Maaren, Y.W., Carlier, I.V., Giltay, E.J., van Noorden, M.S., de Waal, M.W., van der Wee, N.J. & Zitman, F.G. (2012). Reference values for mental health assessment instruments: objectives and methods of the Leiden Routine Outcome Monitoring Study. *Journal of Evaluation in Clinical Practice* In press.
[60] Dawber, T.R., Meadors, G.F. & Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and Nations Health* 41, 279-281.
[61] Gatta, G., Capocaccia, R., Stiller, C., Kaatsch, P., Berrino, F. & Terenziani, M. (2005). Childhood cancer survival trends in Europe: a EUROCARE Working Group study. *Journal of Clinical Oncology* 23, 3742-3751.
[62] Pui, C.H., Sandlund, J.T., Pei, D., Campana, D., Rivera, G.K., Ribeiro, R.C., Rubnitz, J.E., Razzouk, B.I., Howard, S.C., Hudson, M.M., Cheng, C., Kun, L.E., Raimondi, S.C., Behm, F.G., Downing, J.R., Relling, M.V., Evans, W.E. & Total Therapy Study XIIIB at St Jude Children's Research Hospital. (2004). Improved outcome for children with acute lymphoblastic leukemia: results of Total Therapy Study XIIIB at St Jude Children's Research Hospital. *Blood* 104, 2690-2696.
[63] Pui, C.H. & Evans, W.E. (2006). Treatment of acute lymphoblastic leukemia. *New England Journal of Medicine* 354, 166-178.
[64] Lee, S.J., Earle, C.C. & Weeks, J.C. (2000). Outcomes research in oncology: history, conceptual framework, and trends in the literature. *Journal of the National Cancer Instiute* 92, 195-204.
[65] Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P. & International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748-752.
[66] Uher, R., Muthen, B., Souery, D., Mors, O., Jaracz, J., Placentino, A., Petrivic, A., Zobel, A., Henigsberg, N., Rietschel, M., Aitchison, K.J., Farmer, A. & McGuffin, P. (2010). Trajectories of change in depression severity during treatment with antidepressants. *Psychological Medicine* 40, 1367-1377.