Reference values for mental health assessment instruments: objectives and methods of the Leiden Routine Outcome Monitoring Study

Schulte-van Maaren, Y.W.M.; Carlier, I.V.E.; Giltay, E.J.; Noorden, M.S. van; Waal, M.W.M. de; Wee, N.J.A. van der; Zitman, F.G.

Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Reference values for mental health assessment instruments: objectives and methods of the Leiden Routine Outcome Monitoring Study

Yvonne W. M. Schulte-van Maaren MSc,[1] Ingrid V. E. Carlier PsyD PhD,[2] Erik J. Giltay MD PhD,[3] Martijn S. van Noorden MD,[4] Margot W. M. de Waal PhD,[7] Nic J. A. van der Wee MD PhD[5] and Frans G. Zitman MD PhD[6]

[1]Psychologist and Research Fellow, [2]Associate Professor and Research Director of the Centre for Routine Outcome Monitoring, [3]Associate Professor, Psychiatrist, [4]Psychiatrist and Research Fellow, [5]Associate Professor, Psychiatrist, [6]Emeritus Professor of Psychiatry, Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands
[7]Senior Researcher, Epidemiologist, Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands

Keywords
anxiety disorders, depressive disorder, reference values, routine outcome monitoring, instruments, somatoform disorder

Correspondence
Yvonne Schulte-van Maaren
Department of Psychiatry
Leiden University Medical Center
PO Box 9600
2300 RC Leiden
The Netherlands
E-mail: y.w.m.schulte-van_maaren@lumc.nl

Accepted for publication: 9 December 2011

doi:10.1111/j.1365-2753.2012.01830.x

Abstract

Rationale, aims and objectives: Routine outcome monitoring (ROM) was developed to establish the outcome of psychotherapeutic and pharmacological treatments through repeated assessments before, during and after treatment. Although standardization of psychiatric assessments and their reference values are essential for patient care, for various ROM instruments reference values are not available. The aim of the Leiden ROM Study is to generate reference values for 22 ROM instruments, covering generic and specific mood, anxiety and somatoform (MAS) disorders, for the general population. This article describes the extensive process of recruitment, as well as baseline characteristics of patient versus non-patient groups.
Method: Cross-sectional study in randomly selected participants aged 18–65 years from the Dutch population, included through general practitioners.
Results: Extensive demographic, psychosocial, mental health, and biological data from 1302 participants, recruited via general practitioners, were collected during a two-hour standardized assessment including observer-rated and self-report scales. These data will be compared with corresponding data from 7840 patients with psychopathology who were referred to secondary care. On-going quality control and calibration ensured maintenance of high quality during data collection.
Conclusions: This reference group study for mental health assessments is the first study of this size carried out in the Netherlands. The results of this study are expected to be of value to secondary psychiatric care because they allow the indication of progress in health, treatment effect and possible termination of treatment. Additionally, the reference values can be used by primary care physicians as decision threshold for referral to specialized mental health care and vice versa.

Introduction

Routine outcome monitoring (ROM) was developed to enhance the effectiveness of psychiatric care. ROM routinely measures treatment outcomes using different outcome measures that are both generic and disorder specific. It provides clinicians with information on the type and severity of psychopathology and feedback on treatment efficacy. Additional benefits are its use in research and benchmarking [1–3]. However, several ROM instruments lack reference values that provide optimal discrimination between the 'healthy' and the 'diseased', indicating whether the patient has progressed to a range of psychological health similar to non-patients, while not necessarily free of all symptoms. Also, with outcome variables often varying between different gender

and age groups, reference values are the key to determining whether a group or an individual scores above or below average for their gender and age [4,5]. Anchoring ROM instruments in population-based reference values makes clinical and scientific interpretations more meaningful and is consistent with practice in other areas of medicine [6,7]. Furthermore, reference values are useful to determine when primary care doctors could refer their patients to secondary care and vice versa.

In order to study the relationship between psychosocial factors, genetic variation, the effect of the hypothalamic-pituitary-adrenal (HPA) axis stress system, and the occurrence and course of mood, anxiety and somatoform (MAS) disorders, the Leiden Routine Outcome Monitoring Study was designed to generate a large ROM database [8,9].

The present ROM Reference Group Study was designed to provide reference values for 22 ROM in the general practice population in the Netherlands. This may help to facilitate assessment of a clinically significant change of treatment effects, defined as returning to normal functioning.

A secondary aim was to collect saliva from a large general population control group in order to facilitate research on genetic characteristics (DNA) and the HPA axis stress system in relation to the development and course of MAS disorders. Genetic factors and a deregulated HPA axis are involved in the aetiology of MAS disorders. Twin studies [10,11] have shown that mood and anxiety disorders are for 30–40% determined by hereditary factors. Furthermore, dysregulation of the HPA axis is believed to be linked with the pathophysiology of depression [12–14] and anxiety disorders [15,16].

The present study describes the methods and objectives of the ROM Reference Group Study, as well as baseline characteristics of patient versus non-patient groups.

Methods

Participants

The ROM reference group was recruited to serve as a comparison for the ROM patient group. Therefore, the aim for this reference group was that it should be representative of the ROM population referred for suspected (but not necessarily diagnosed with) MAS disorders, treated at the psychiatric outpatient department of the Leiden University Medical Center (LUMC) or at the mental health clinics of Rivierduinen (RD; hereafter referred to as the 'ROM patient group'). The sample was stratified for gender, age and urbanization level to be representative of the ROM patient group [17].

A total of 1302 participants (18–65 years) was recruited, 1294 of whom provided complete datasets (Fig. 1). In order to recruit persons reflecting normal functioning with different levels of subthreshold psychopathology, recruitment took place via general practices. In the Netherlands, because 99.9% of the general population is registered with a general practitioner (GP) [18], the practice registers provide a convenient frame for sampling the local general population. Eight university-affiliated general practices with a total of ±14 000 enlisted patients in the vicinity of Leiden were involved. In order to form a *non-patient* control group and to secure the reliability and validity of the collected data, four exclusion criteria were formulated: (1) treatment in a secondary psychiatric care centre in the last 6 months for psychiatric problems and/or dependence on alcohol or drugs; (2) hearing impairment, limited cognitive abilities, such as aphasia, severe dyslexia or dementia; (3) illiteracy or insufficient mastery of the Dutch language; and (4) a terminal disease.

The study protocol was approved by the Ethical Review Board (ERB) of the LUMC and all subjects signed informed consent.

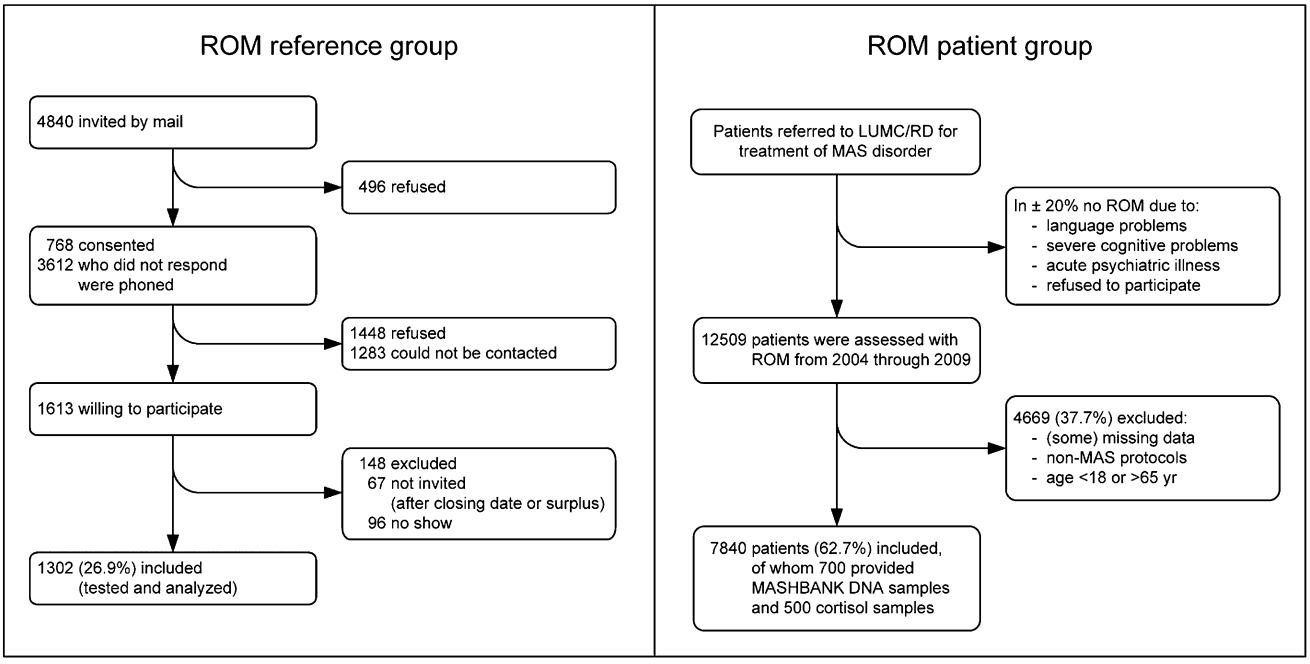

Figure 1 Flow chart depicting recruitment of the ROM reference and patient groups.

Since 2002, the LUMC and RD, serving a region of more than 1 million people, have implemented ROM [1]. ROM baseline assessments in the ROM patient group started in 2002 and are ongoing. Specially trained psychiatric research nurses assessed 80% of the patients (totalling 8357 ROM patients), 7840 of whom were aged 18–65 years. To facilitate research on genetic characteristics (DNA) and the HPA axis stress system (cortisol day curves), the MASHBANK (biobank for MAS disorders and the HPA axis) was founded at the LUMC and RD in 2007 after approval by the ERB of the LUMC. In this biobank, saliva samples are stored from ±1000 consenting MAS patients. Figure 1 shows the multistage recruitment flow of the ROM reference group, as well as recruitment of the ROM patient group.

Participants of the ROM reference group were offered the full set of generic instruments. Since the total number of instruments was too extensive and all participants were already asked to complete the depression instruments, random samples of 50% each were asked to complete the anxiety instruments or the somatoform instruments, with even ratios of males and females in each subgroup. Thus, four subgroups were established: males-anxiety; females-anxiety; males-somatoform; females-somatoform. A sample size of at least 120 per subgroup was considered to provide adequate power to yield reference values [19]. In genetic research, an adequate sample size is imperative because of the low frequency of several genetic variants and the problem of multiple testing. Furthermore, a sample size of 1000 DNA donating participants was deemed to be required [20]. With an anticipated response rate of 30%, about 4500 people were approached. In order to get a ROM-representative sample, four age groups were used: 18–25; 26–40; 41–55; 56–65 years, and the reference group was sampled accordingly.

Procedures

In order to recruit the ROM reference group, the eight participating GPs first screened their patient lists for those that met the inclusion/exclusion criteria. Subsequently, randomly selected appropriate persons were invited to participate by a letter (sent by regular postal service) by their GPs that was followed by an announced telephone call by the research team to ask for their participation. Objections against this call could be indicated on an enclosed reply card. To compensate for possible seasonal influences, recruitment took place all year long (between November 2009 and January 2011). Location was the LUMC clinic site and, if appreciated, at the participant's home or in the GP's practice. Similar to the ROM patient assessment procedures, dedicated Web-based computer software was used for the administration of all instruments and to prevent missing data within instruments. It was also used for data collection and storage, and for creation of summary variables [1]. Touch screens were used to accommodate computer-illiterate participants. A personal data entry program was developed in database software to organize identification codes for general, ROM and MASHBANK data, and to randomly assign the two specific instrument packets (depression and anxiety; depression and somatoform) to participants.

For participants of the ROM reference group, the interview started with an explanation of the study, and signing of the informed consent form. This was followed by a check and assessment of personal details and demographic data, general health, cognitive functioning and physical examination (i.e. body weight, height and blood pressure). Saliva samples were collected in participants who additionally consented to this biobank substudy. Next, computerized observer-rated and self-report questionnaires were completed. Finally, participants completed an evaluation form and received a gift voucher of €30 (for their time and cooperation) and a travel allowance.

In the ROM Reference Group Study, 3 psychiatric research nurses, 3 psychologists (Master's degree level) and 11 Master's students in psychology were extensively trained and tested at the start of and during the reference group study to ensure uniform and adequate quality and reliability. Topics were Mini-International Neuropsychiatric Interview Plus, version 5.0.0-R (MINI-Plus 5.0.0.) and abbreviated Comprehensive Psychopathological Rating Scale (vCPRS) interviewing methods, Global Assessment of Functioning Scale (GAF) scoring, use of QuestManager and additional knowledge about MAS disorders and MASHBANK. Three full days of training (by the primary investigator, SvM, two psychiatrists and two ROM-trained nurses) took place. Each interviewer also observed at least three interviews, and the first two interviews were carried out under supervision (one of which observed by the primary investigator). Supervision regarding interview techniques, problematic behaviour of the participants and scoring rating scales, to improve inter-rater reliability, took place every 2 months. Video recordings of interviews were used to further calibrate assessments between interviewers. Using a semi-structured scoring scale, a qualitative assessment was done and was found to be very good in all but one potential interviewer. This latter interviewer with insufficient skills was considered unsuitable and no longer took part. The ROM patient group was assessed by two trained ROM psychiatric research nurses; their training has been described in detail elsewhere [1].

Assessments

The ROM reference group assessment comprised measurement of physical health, saliva collection and observer-rated and self-report instruments. Measurement of physical health indicators comprised blood pressure, heart rate and body mass index, and health-related factors (i.e. general health, chronic diseases, smoking status and alcohol consumption).

From participants who agreed to participate in the MASHBANK substudy, saliva was collected enabling cortisol measurements and DNA isolation. HPA axis activity was assessed by free cortisol measurements using seven saliva samples per participant, self-collected at home. Procedures are similar to that described in detail elsewhere [14,16,21]. Saliva for DNA isolation was collected in DNA Genotek kits (Oragene, DNA Genotek, Ottawa, OT, Canada). Measuring cortisol and DNA concentrations in saliva has many advantages over measurements in blood samples. Saliva collection is non-invasive and can be repeated frequently. Furthermore, storage of the material requires no special treatment because DNA and cortisol levels remain stable at room temperature.

The assessments comprised 25 instruments concerning demographic and personal characteristics, psychosocial function, physical health and psychopathology (Table 1), 22 of which require reference values. Except for the 48-item Symptom Questionnaire (SQ-48), all tested ROM instruments are internationally used and

Table 1 Instruments used in the ROM reference and patient groups

Instrument	Full name	Domain	No. of items	Time (min)	Type	Public domain	References
Generic							
Personal							
DEMOG	Demographic inventory	Demography	12	2	SR	Yes	
CTQ	Child trauma questionnaire	Traumatic events childhood	28	5	SR	Yes	[35]
Psychosocial functioning							
GAF	Global assessment of functioning	General functioning	1	1	Obs	Yes	[36]
LOT-R	Life orientation test – revised	Optimism	10	5	SR	Yes	[37]
SF-36	Short form health survey 36	Physical health	36	6	SR	Yes	[38]
Psychopathology							
BSI	Brief symptom inventory	General pathology	53	8	SR	No	[39]
DAPP-sf	Dimensional assessment of personality pathology – short form	Personality	136	33	SR	No	[40]
IES-R	Impact of event scale – revised	Traumatic events	22	5	SR	Yes	[41,42]
MASQ-D30	Mood & anxiety symptom questionnaire –30	Mood and anxiety	30	5	SR	Yes	[43]
MINI-Plus 5.0.0.*	mini international neuropsychiatric interview plus 5.0.0.	General pathology	-	30	Obs	Yes	[44]
SQ-48	Symptom questionnaire –48 items	General pathology	55	4	SR	Yes	[8]
vCPRS*	Abbreviated comprehensive psychopathological rating scale	General pathology	25	10	Obs	Yes	[45]
WSQ	Web screening questionnaire for common mental disorders	General pathology	15	5	SR	Yes	[46]
Depressive disorder							
IDS-SR	Inventory of depressive symptoms	Depressive disorder	34	3	SR	Yes	[47]
BDI-II	Beck depression inventory version II	Depression, dysthymia & bipolar disorder	21	5	SR	No	[48]
Anxiety disorder							
AGO	Agoraphobia scale	Panic disorder	20	5	SR	Yes	[49]
PADUA/PI-r	PADUA inventory revised	Obsessive compulsive disorder	41	6	SR	Yes	[50]
PAI	Panic appraisal inventory	Panic disorder	45	10	SR	Yes	[51]
PSWQ	Penn state worry questionnaire	Generalized anxiety disorder	16	3	SR	Yes	[52]
SPS	Social phobia scale	Social phobia	20	5	SR	Yes	[53]
SIAS	Social interaction and anxiety scale	Social phobia	20	5	SR	Yes	[53]
WDQ	Worry domains questionnaire	Generalized anxiety disorder	30	3	SR	Yes	[54,54]
Somatoform disorder							
BICI	Body image concern inventory	Body dysmorphic disorder	19	4	SR	Yes	[55]
CIS20r	checklist individual strength	Chronic fatigue syndrome	20	5	SR	Yes	[56]
WI	Whitely index	Hypochondriasis	14	3	SR	Yes	[57]

*The MINI-Plus 5.0.0 and vCPRS are used for diagnoses; no reference values were established.

A list of all ROM instruments, including references of Dutch translations, is available at http://www.lumc.nl/psychiatry/ROM-instruments

Obs; observer-rated, ROM, routine outcome monitoring; SR, self-report.

validated. The generic self-report instrument SQ-48 was recently developed by our research group in order to assess mood, anxiety, somatoform symptoms, hostility and vitality.

Statistical analyses

Reference values will be calculated for all instruments, including subscales. Both for patients and for the reference group reference values will be determined for all subjects combined, as well as for four groups: young males (aged 18–40 years), older males (aged 41–65 years), young females (aged 18–40 years) and older females (aged 41–65 years). Means and SDs, 5th, 25th, 50th, 75th and 95th percentiles, and receiver operating characteristics (ROC) analyses (i.e. the cut-off score with the optimal sensitivity and specificity, and area under curve values) will be computed. Reference limits are often defined by two SDs below and above the mean if distributions are Gaussian. Since most distributions of total scores on the scales tested in the healthy reference group are expected to be strongly (positively) skewed, percentiles are more appropriate [22–24], with the lower interval bounded only by the 95th percentile being a common reference group [25]. However, trade-offs exist between the sensitivity and specificity, with a higher cut-off value (i.e. higher percentile boundary) having a relatively high specificity but low sensitivity, and vice versa (Fig. 2; left panel). ROC analyses will provide additional cut-offs reflecting discriminatory power [26].

Figure 2 (right panel) shows psychopathology expressed as the number of MINI diagnoses of MAS disorders in the ROM reference group and the ROM patient group.

Results

Figure 1 shows recruitment of the ROM reference group and the ROM patient group. A total of 1302 persons were interviewed and their data analyzed. The duration of the interview was shorter (range 1.5–2.0 hours) in participants without psychopathology and longer (range 2.5–4.0 hours) in participants with psychopathology. Although the interview was extensive, all participants finished the full assessment. Additional telephone calls after the initial mailing proved to have a motivating effect on the subsequent response rates. Patients from the first GP only received the invitation by mail (no telephone call) and showed a response of 16.3%. We tried to contact patients recruited from all other GPs by telephone. The response to the initial mail, before the telephone call by the research team, was 15.9% (768 of 4840). The response to the telephone call and the mail was 45.3% of those the research team managed to contact (1613 of 3557). A total of 67 responders were not included because of a surplus in some of the age groups, or because of logistical reasons at the end of the study. Therefore, the response of persons contacted was 37.3% (1302 of 3490). However, when taking into account the large group of 1283 persons that could not be contacted by letter or telephone, the response of persons mailed was 26.9% (1302 of 4840). A total of 148 persons were excluded: 36 who consented following the mail (treated in a secondary psychiatric care centre, or insufficient mastery of the Dutch language), 101 after a telephone call (for similar reasons) and 11 during or after the interview (for similar reasons, as well as severe dyslexia or cognitive impairment).

Table 2 presents the sociodemographic characteristics of the ROM reference group ($n = 1294$) and the ROM patient group ($n = 7840$), 543 of whom did not complete the demographic inventory. Gender and age distributions in both samples were similar, and the mean age in both samples was 2 years higher for men than for women. Compared to the ROM patient group, the ROM reference group less often lived in a rural area, was less often divorced, separated or widowed, was less often unemployed or disabled and had a higher educational level.

The aim for the ROM reference group was that it should be a 'normal' group but allowed for prevalent psychopathology that could be treated in the GP practices, and therefore, showed some (co-)morbidity of psychiatric illness but to a much lesser extent

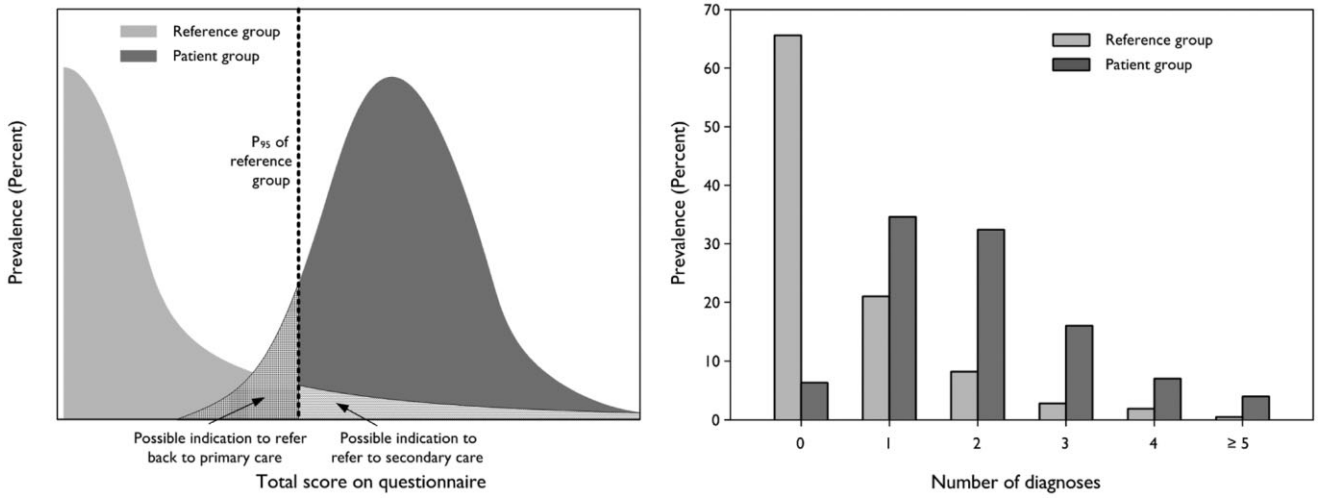

Figure 2 Left panel: the expected distribution of many of the 22 mood, anxiety and somatoform (MAS) disorder-assessment instruments in the ROM reference and patient groups; Right panel: the actual distribution of the number of MAS disorders in the ROM reference and patient groups. In the ROM reference group, above the 95th percentile (P_{95}; i.e. reference value) the probability is high for a person to meet the terms of psychopathology.

Table 2 Sociodemographic characteristics of the ROM reference group ($n = 1294$) and the ROM patient group ($n = 7297$)

	ROM reference group	ROM patient group
Gender		
Male	484 (37.4%)	2700 (37.0%)
Female	810 (62.6%)	4597 (63.0%)
Age (mean, SD) in years	40.2 (12.5)	37.9 (12.3)
18–25	194 (15.0%)	1508 (20.7%)
26–40	479 (37.0%)	2715 (37.2%)
41–55	448 (34.6%)	2370 (32.5%)
56–65	173 (13.4%)	704 (9.6%)
Urbanization level		
Urban	806 (62.3%)	3955 (54.2%)
Rural	488 (37.7%)	3342 (45.8%)
Marital status		
Married/cohabitating	890 (68.8%)	3721 (50.9%)
Divorced/separated/widow	78 (6.0%)	989 (13.6%)
Single	326 (25.2%)	2587 (35.5%)
Housing situation		
Living alone	200 (15.7)	1693 (23.2%)
Living with partner	902 (69.7)	3762 (51.6%)
Living with family	192 (14.8)	1842 (25.2%)
Educational status		
Lower	295 (22.8)	3133 (42.9%)
Higher	999 (77.2)	4164 (57.1%)
Employment status		
Employed part-time	508 (39.3%)	1737 (23.9%)
Employed full-time	554 (42.8%)	1702 (23.3%)
Unemployed/retired	197 (15.2%)	2118 (27.1%)
Work-related disability	35 (2.7%)	1874 (25.7%)
Ethnic background		
Dutch	1160 (89.6%)	5981 (80.0%)
Other ethnicity	134 (10.4%)	1316 (18.0%)

ROM, routine outcome monitoring.

than the ROM patient group (Fig. 2). According to the MINI-Plus, 9.4% of the ROM reference group met criteria for one or more MAS disorders compared to 74.5% in the ROM patient group. A single MAS diagnosis was present in 7.8% participants and in 47.9% ROM patients. In the ROM reference group, anxiety disorders were most prevalent followed by somatoform disorders. In the ROM patient group, major depression was the most prevalent disorder followed by anxiety disorders. Thus, the ROM reference group showed lower co-morbidity than the ROM patient group and reflected psychiatric morbidity within the general population (Table 3, Fig. 2).

Discussion

This cross-sectional study in a randomly selected sample from a Dutch general population (aged 18–65 years) aimed to provide reference values for ROM instruments (and to serve as a control group for the biobank) for patients with MAS disorders. It is the first reference group study for mental health assessments of this size carried out in the Netherlands. The large sample size and extensive assessment of psychopathology provide data that, by comparison with data from ROM patients, is expected to yield reliable reference values for ROM instruments (across a wide age range) that are not yet available. Genetic and HPA axis data enable further biological research into MAS disorders.

Comparison of the demographics of the ROM reference and patient groups showed a similar gender and age distribution, as expected given the sampling frame. There was a slightly (unintentional) different urbanization level. However, the effects of urbanicity on psychopathology are generally of limited significance in international [27] and Dutch (NEMESIS; Bijl *et al*. 1998) co-morbidity studies. Moreover, differences between rural and urban areas are declining in the Netherlands. Compared to the ROM patient group, the ROM reference group showed higher levels of education and less unemployment or work-related

Table 3 Mood, anxiety and somatoform (co-)morbidity in the ROM reference group ($n = 1302$) and the ROM patient group ($n = 7840$)

	ROM reference group		ROM patient group	
	Frequency	Per cent	Frequency	Per cent
None	1193	90.6	1998	25.5
Anxiety	54	4.1	1568	20.0
Mood	7	0.5	1682	21.5
Somatoform	42	3.2	500	6.4
Anxiety and mood	7	0.5	1377	17.6
Anxiety and somatoform	9	0.7	209	2.7
Mood and somatoform	1	0.1	275	3.5
Anxiety and mood and somatoform	2	0.2	231	2.9
Total anxiety	72	5.5	3385	43.2
Total mood	17	1.3	3565	45.5
Total somatoform	54	4.2	1215	15.5
Total	1302	100.0	7840	100.0

Anxiety disorders comprise panic disorder with or without agoraphobia, agoraphobia without history of panic disorder, specific phobia, social phobia, obsessive–compulsive disorder, post-traumatic stress disorder and generalized anxiety disorder, and anxiety disorders NOS.
Mood disorders comprise major depressive disorders, bipolar disorder, dysthymia.
Somatoform disorders comprise somatization disorder, undifferentiated somatoform disorder, pain disorder (chronic), hypochondriasis, body dysmorphic disorder and conversion disorder.
ROM; routine outcome monitoring.

disability. Accordingly, both co-morbidity studies [4,27] reported the highest morbidity rates for those with the lowest levels of education, and the lowest morbidity rates for those with the highest levels of education. Mental disorders were reported to be least prevalent among people in paid employment. Overall, morbidity and co-morbidity were strongly associated with occupational disability and unemployment.

As expected, morbidity of any current MAS disorder in the reference group was much lower than in the ROM patient group. Anxiety disorders were equally prevalent in the ROM reference group compared to a study in the general practice population ($n = 1778$) in the Netherlands (5.5%) [28]. Mood disorders were less prevalent in the reference group (1.3%) than in the general practice population (4.1%) as well as compared to prevalence rates in various European countries, ranging between 4.6% and 7.4% [29]. The current prevalence rate for somatoform disorders was 4.2% in our ROM reference group, compared to 16.1% in a general practice population [28]. This discrepancy can probably be ascribed to differences in the recruitment procedure, as the latter study included consultation seeking patients, whereas we included a random sample of the general practice population. Also, in our study, most interviews took place in hospital versus home interviews in the study of De Waal *et al*. Another explanation could be differences in the ascertainment of depressive and somatoform disorders (MINI-Plus 5.0.0. in our study versus the Scan diagnostic interview in the study of De Waal *et al*.). Moreover, selection and non-response bias may have occurred in our study, as depressed people are often less inclined to participate because of fatigue or loss of energy. Co-morbidity rates of psychopathology in the reference group were similar to those reported in the Dutch co-morbidity study [4] and very low compared to the ROM patient group.

Several issues need to be considered when analyzing reference values for psychiatric assessment scales from healthy populations. Reference values need to be accurate and reproducible. First, in samples derived from the general population many of the total scores do not have a bell-shaped Gaussian distribution, but rather an asymmetrical, right-sided, skewed distribution. When log-Gaussian curves are also not normally distributed, means with (1.96 times) SDs cannot be used to yield the central 95% of the reference population of subjects. Rather, percentile values (e.g. 97.5th, 95th or 90th) can be used, as this non-parametric method makes no specific assumption regarding the distribution from which the data are obtained. Nevertheless, extreme values can still have a profound effect in defining reference values, and therefore, sample sizes (in subgroups) of at least 120 are needed (for 90% confidence intervals) to reduce the amount of uncertainty [19,30,30]. Second, outliers can be removed before the analysis, using outlier detection methodology. For example, if the difference between the extreme and the next most extreme value exceeds 1/3rd of the range, the extreme value can be deleted (i.e. the Dixon test method) [19]; this may yield better reference values. However, an attempt should first be made to determine whether these extremes are errors in the assessment procedure. Third, there may be a profound influence from healthy and non-healthy (psychiatrically ill individuals) individuals on the estimation of reference values. About 10% narrower reference intervals will be derived from samples that excluded non-healthy subjects [31] but could make the reference range unreasonably narrow. Therefore, we chose to study a 'control' group rather than a 'healthy' group. Overall, there are many trade-offs between the different parametric, transformed parametric and non-parametric methodologies.

Reference values for psychiatric instruments are essential for patient care. In this ROM reference group, data were collected enabling the calculation of reference values for 22 ROM instruments that often lack these values because recruiting valid groups of reference subjects is costly and time intensive. These reference values are of major clinical importance because they can help to weigh the severity of symptoms and provide criteria that signify the transition from illness to health, and potential treatment termination. They can also be used by primary care physicians for referral to secondary care, and vice versa. Additionally, reference material to facilitate research on genetic characteristics (DNA) and the HPA axis stress system was collected.

Our study has specific strengths. First, to yield reliable and stable reference values, the group has to be of sufficient size and representative for the patient group of interest. Tests for decisions at the individual level such as therapy indication or monitoring require a sample size of at least 250 subjects per reference group standardized for age and sex [32,33]. The size of the group and four subgroups surpassed this number and the previously described size of the 120 recommended participants [19,30], even when partitioning the test subjects by sex and age groups. Second, the diagnostic interview was structured leading to better identification of diagnostic co-morbidity than unstructured interviews [34]. Next to self-report data, observational data were collected using the MINI-Plus. This approach provided comprehensive clinical information according to international standards (DSM-IV). Third, standardization of the interviews was assured, as both observation scales and self-report questionnaires were administered via a Web-based computer program, implying a fixed order in administration of instruments with no instruments skipped or data missing, and no errors because of manually entering data. Fourth, recruitment through GPs allowed for a good description of the sample characteristics. Furthermore, contacting possible participants by telephone presumably increased the response rate. Finally, an ongoing quality control and calibration among interviewers ensured that a high quality was maintained during data collection.

The present study also has some limitations. First, because recruitment of the ROM reference and patient group took place in the Dutch region of Leiden, reference values may not be directly internationally generalizable. Moreover, because ethnic participants formed a minority, generalizability of reference values to other countries and ethnicities is limited. Second, children and elderly were not included, thus requiring their own reference group studies. Third, non-response was significant, involving a possible, unknown bias. Finally, information about the characteristics of those who did not participate is lacking. It is unclear whether non-responders differed in a systematic way from the participating subjects.

In conclusion, we succeeded in collecting extensive data from 1302 persons from the general population, enabling the calculation of reference values for 22 ROM instruments. The results of the reference values are expected to become available within the next 2 years and will be useful for current and future diagnostic and research purposes in patients with MAS disorders.

Authors' contributions

YS participated in the study design and coordination, carried out the study and statistical analyses and drafted the manuscript. IC participated in the study design and the coordination and helped to draft the manuscript. EG participated in the conception and design of the study, the statistical analyses and helped to draft the manuscript. MVN participated in the study design. MDW played a key role in the recruitment of the reference group. NVDW participated in the coordination of the study. FZ conceived of the study and participated in its design. All authors read and approved the final manuscript.

Acknowledgements and Funding

This study was supported by Rivierduinen, a regional mental health care provider in the Netherlands. The authors gratefully acknowledge Gea Corton for her coordination of the recruitment of the reference group. We are indebted to the interviewers for their help in collecting the data: Marian Lucas, Lilo Gerdessen, Lennie Geerlings, Anna Coppoolse, Bart Huisman, Roy de Kleijn, Nathalie Boot, Charlotte Flierboom, Anne Raterman, Elise van Holsteijn, Karlijn Noest, Lotte van Reuler, Margina Ruiter, Nicole Billingy, Vera Atema, Mirjam vd Hoorn and Cyntha Bogaart. Marion Streevelaar and Marieke de Wit were of great value to the data management. In addition, the essential contribution made by the participants of this study and their general practitioners is very gratefully acknowledged.

Competing interests

The authors have no competing interests.

References

1. 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.
2. Lambert, M. (2007) Presidential address: what we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research*, 17 (1), 1–14.
3. Miller, S. D., Duncan, B. L., Sorrell, R. & Brown, G. S. (2005) The partners for change outcome management system. *Journal of Clinical Psychology*, 61 (2), 199–208.
4. 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 (12), 587–595.
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 (1–3), 116–123.
6. Linnet, K. (2000) Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clinical Chemistry*, 46 (6 Pt 1), 867–869.
7. Overbeek, L. I., Kapusta, L., Peer, P. G., De Korte, C. L., Thijssen, J. M. & Daniels, O. (2006) New reference values for echocardiographic dimensions of healthy Dutch children. *European Journal of Echocardiography*, 7 (2), 113–121.
8. Carlier, I. V., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J. & Zitman, F. G. (2011) Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice* 18 (1), 104–110.
9. 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. (2010) Pre-adult versus adult onset major depressive disorder in a naturalistic patient sample: the Leiden Routine Outcome Monitoring Study. *Psychological Medicine*, 16, 1–11.
10. Sullivan, P. F., Neale, M. C. & Kendler, K. S. (2000) Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, 157 (10), 1552–1562.
11. Uhl, G. R. & Grow, R. W. (2004) The burden of complex genetics in brain disorders. *Archives of General Psychiatry*, 61 (3), 223–229.
12. Carroll, B. J., Cassidy, F., Naftolowitz, D., Tatham, N. E., Wilson, W. H., Iranmanesh, A., Liu, P. Y. & Veldhuis, J. D. (2007) Pathophysiology of hypercortisolism in depression. *Acta Psychiatrica Scandinavica. Supplementum*, 433, 90–103.
13. Pariante, C. M. & Lightman, S. L. (2008) The HPA axis in major depression: classical theories and new developments. *Trends in Neurosciences*, 31 (9), 464–468.
14. Vreeburg, S. A., Hoogendijk, W. J., Van Pelt, J., Derijk, R. H., Verhagen, J. C., Van Dyck, R., Smit, J. H., Zitman, F. G. & Penninx, B. W. (2009) Major depressive disorder and hypothalamic-pituitary-adrenal axis activity: results from a large cohort study. *Archives of General Psychiatry*, 66 (6), 617–626.
15. De Kloet, E. R., Joels, M. & Holsboer, F. (2005) Stress and the brain: from adaptation to disease. *Nature Reviews. Neuroscience*, 6 (6), 463–475.
16. Vreeburg, S. A., Zitman, F. G., Van Pelt, J., Derijk, R. H., Verhagen, J. C., Van Dyck, R., Hoogendijk, W. J., Smit, J. H. & Penninx, B. W. (2010) Salivary cortisol levels in persons with and without different anxiety disorders. *Psychosomatic Medicine*, 72 (4), 340–347.
17. Kendall, P. C., Marrs-Garcia, A., Nath, S. R. & Sheldrick, R. C. (1999) Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, 67 (3), 285–299.
18. Poortvliet, M. C., Lamkadden, M. & Deville, W. Niet op naam ingeschreven (NONI) bij de huisarts. Inventarisatie en gevolgen voor de ziekenfondsverzekerden. NIVEL; 2005.
19. Horowitz, G. L., Altaie, S., Boyd, J. C., Ceriotti, F., Garg, U., Horn, P., Pesce, A., Sine, H. E. & Zakowski, J. (2008) Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline – Third Edition CLSI Document C28-A3. Wayne, PA: Clinical and Laboratory Standards Institute.
20. Nannya, Y., Taura, K., Kurokawa, M., Chiba, S. & Ogawa, S. (2007) Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project. *Human Molecular Genetics*, 16 (20), 2494–2505.
21. Penninx, B. W., Beekman, A. T., Smit, J. H., *et al*. (2008) The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17 (3), 121–140.
22. Bourdon, K. H., Goodman, R., Rae, D. S., Simpson, G. & Koretz, D. S. (2005) The strengths and difficulties questionnaire: U.S. normative data and psychometric properties. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44 (6), 557–564.
23. Noerholm, V., Groenvold, M., Watt, T., Bjorner, J. B., Rasmussen, N. A. & Bech, P. (2004) Quality of life in the Danish general population – normative data and validity of WHOQOL-BREF using Rasch

and item response theory models. *Quality of Life Research*, 13 (2), 531–540.
24. Reedtz, C., Bertelsen, B., Lurie, J., Handegard, B. H., Clifford, G. & Morch, W. T. (2008) Eyberg Child Behavior Inventory (ECBI): Norwegian norms to identify conduct problems in children. *Scandinavian Journal of Psychology*, 49 (1), 31–38.
25. Solberg, H. E. (2008) Establishment and use of reference values. In Fundamentals of Clinical Chemistry (eds C. A. Burtis, E. R. Ashwood & D. E. Bruns), pp. 229–238. 6 edn St. Louis, MO: Saunders Elsevier.
26. Barnabei, L., Marazia, S. & De, C. R. (2007) Receiver operating characteristic (ROC) curves and the definition of threshold levels to diagnose coronary artery disease on electrocardiographic stress testing. Part I: the use of ROC curves in diagnostic medicine and electrocardiographic markers of ischaemia. *Journal of Cardiovascular Medicine (Hagerstown, Md.)*, 8 (11), 873–881.
27. Kessler, R. C., McGonagle, K. A., Zhao, S., 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 (1), 8–19.
28. De Waal, M. W., Arnold, I. A., Eekhof, J. A. & Van Hemert, A. M. (2004) Somatoform disorders in general practice: prevalence, functional impairment and comorbidity with anxiety and depressive disorders. *The British Journal of Psychiatry; The Journal of Mental Science*, 184, 470–476.
29. Smith, A. L. & Weissman, M. M. (1992) Epidemiology. (ed E. S. Paykel), pp. 111–129. Handbook of Affective Disorders. New York: Guildford.
30. Reed, A. H., Henry, R. J. & Mason, W. B. (1971) Influence of statistical method used on the resulting estimate of normal range. *Clinical Chemistry*, 17 (4), 275–284.
31. Horn, P. S., Feng, L., Li, Y. & Pesce, A. J. (2001) Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47 (12), 2137–2145.
32. Angoff, W. H. (1971) Scales, norms and equivalent scores. In Educational Measurement (ed. R. L. Thorndike), pp. 508–600. Washington, DC: American Council on Education.
33. Bechger, T., Hemker, B. & Maris, G. (2009) Over Het Gebruik Van Continue Normering. Arnhem: Cito.
34. Zimmerman, M. & Mattia, J. I. (1999) Psychiatric diagnosis in clinical practice: is comorbidity being missed? *Comprehensive Psychiatry*, 40 (3), 182–191.
35. Bernstein, D. P., Stein, J. A., Newcomb, M. D., *et al*. (2003) Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, 27 (2), 169–190.
36. Endicott, J., Spitzer, R. L., Fleiss, J. L. & Cohen, J. (1976) The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, 33 (6), 766–771.
37. Scheier, M. F., Carver, C. S. & Bridges, M. W. (1994) Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem) – a reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67 (6), 1063–1078.
38. Ware, J. E., Snow, K. K. & Kosinski, M. (1993) SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: New England Medical center, The Health Institute.
39. DeRogatis, L. R. (1975) The Brief Symptom Inventory. Baltimore, MD: Clinical Psychometric Research.
40. Livesley, W. J. & Jackson, D. N. (2002) Manual for the Dimensional Assessment of Personality Pathology – Basic Questionnaire (DAPPBQ). Port Huron, MI: Sigma Press.
41. Horowitz, M. J., Wilner, N. R. & Alvarez, W. (1979) Impact of Event Scale: a measure of subjective stress. *Psychosomatic Medicine*, 41, 209–218.
42. Creamer, M., Bell, R. & Failla, S. (2003) Psychometric properties of the impact of event scale – revised. *Behaviour Resarch and Therapy*, 41 (12), 1489–1496.
43. Clark, L. A. & Watson, D. (1991) Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, 100 (3), 316–336.
44. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R. & Dunbar, G. C. (1998) The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, 59, 22–33.
45. Asberg, M., Montgomery, S. A., Perris, C., Schalling, D. & Sedvall, G. (1978) A comprehensive psychopathological rating scale. *Acta Psychiatrica Scandinavica. Supplementum*, 271, 5–27.
46. Donker, T., Van Straten, A., Marks, I. & Cuijpers, P. (2009) A brief web-based screening questionnaire for common mental disorders: development and validation. *Journal of Medical Internet Research*, 11 (3), e19.
47. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. (1996) The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychological Medicine*, 26 (3), 477–486.
48. Beck, A. T. & Steer, R. A. (1984) Internal consistencies of the original and revised beck depression inventory. *Journal of Clinical Psychology*, 40 (6), 1365–1367.
49. Ost, L. G. (1990) The agoraphobia scale – an evaluation of its reliability and validity. *Behaviour Research and Therapy*, 28 (4), 323–329.
50. Sanavio, E. (1988) Obsessions and compulsions – the Padua Inventory. *Behaviour Research and Therapy*, 26 (2), 169–177.
51. Telch, M. J., Brouillard, M., Telch, C. F., Agras, W. S. & Taylor, C. B. (1989) Role of cognitive appraisal in panic-related avoidance. *Behaviour Research and Therapy*, 27 (4), 373–383.
52. Meyer, T. J., Miller, M. L., Metzger, R. L. & Borkovec, T. D. (1990) Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, 28 (6), 487–495.
53. Beidel, D. C., Turner, S. M., Stanley, M. A. & Dancu, C. V. (1989) The social phobia and anxiety inventory – concurrent and external validity. *Behavior Therapy*, 20 (3), 417–427.
54. Tallis, F., Eysenck, M. & Mathews, A. (1992) A questionnaire for the measurement of nonpathological worry. *Personality and Individual Differences*, 13 (2), 161–168.
55. Littleton, H. L., Axsom, D. & Pury, C. L. S. (2005) Development of the body image concern inventory. *Behaviour Research and Therapy*, 43 (2), 229–241.
56. Vercoulen, J. H., Swanink, C. M., Fennis, J. F., Galama, J. M., van der Meer, J. W. & Bleijenberg, G. (1994) Dimensional assessment of chronic fatigue syndrome. *Journal of Psychosomatic Research*, 38 (5), 383–392.
57. Pilowsky, I. (1967) Dimensions of hypochondriasis. *British Journal of Psychiatry*, 113 (494), 89–93.

