Menopause Rating Scale (MRS)
Development of the scale

Status report by Lothar A.J. Heinemann
March 2007

1. Conceptual framework and intended application

1.1. Identification of concept and domains to be measured

This review follows the structure of recommendations for Patient-Reported Outcomes (PRO)\(^1\) issued by FDA (USA), although this concept did not exist when the MRS scale was developed. However, the “state-of-the-art” psychometric concept for test (scale) development contained at this early time already all important methodological elements of the current recommendations for PROs, at least when the MRS-2 (the current, validated scale) was developed.

The interest of clinical therapy-related research in aging women increased remarkably after the 1980ies and thereby the interest to document the effect of therapeutic interventions. In the 1990s, the interest in validated instruments to measure health related quality of life became increasingly an issue. Another important trend to be considered was the increasing recognition in more recent years among clinicians and researchers of the role of patient-reported data as outcome measures for clinical and drug research. Health authorities are in support of this growing interest. As a reflection, we can observe multiple attempts for a state-of-the-art development of health-related QoL scales applicable to women in their menopausal transition.

The first widely accepted attempt to document the severity of menopausal complaints in women was the Kupperman Index\(^1,2\). The focus of this symptom listing was primarily severity and degree of relief of symptoms, assessed on the basis of the physician’s impressions of the severity of complaints and assisted by the index rather than let women respond to their perceived symptoms. An interview of women by the physician was the background for the Kupperman Index. The content of this symptoms list (dimension) was neither formally analyzed nor was there any formal validation.

A new Symptom List – the Menopause Rating Scale – was developed in 1992\(^3,4\). It was initially developed to provide the physician with a tool to document specific climacteric symptoms and their changes during treatment and was seen as an improvement over the commonly applied Kupperman Index. The questions were selected based on clinical experience, and reviewed by representatives of the German Association for Gynecology & Obstetrics. This resulted in a questionnaire with 10 items. A continuous rating scale from 0.0 to 1.0 (visual-analog scale) was used to document the severity of symptoms. This scale was broadly used in clinical studies in the early 1990s. However, it was not used as self-administrable questionnaire – it was completed by the physician who did the interview with the patient. Moreover, there was no binding guideline as how to perform the interview.

---

A critical assessment of this first questionnaire, however, disclosed methodological deficiencies, which, both in theory and practice, limited its use. This resulted in proposals to improve the physician-based scale such as: application of the symptom list in a representative sample of women after questionnaire revision; revision of the questionnaire such that women will complete it themselves; first of all because self-assessment is more sensitive, and second, a self-administered questionnaire would not limit future application; modification of the wording of items to a simple, laymen-appropriate form; proper psychometric evaluation of the revised scale based on a representative sample and development of simple-to-use standardized items with identification of the dimensions.

In early 1996, the MRS was formally standardized following up-to-date psychometric rules and using a normative sample of 689 German women aged 40 to 60 years. Finally, the MRS scale – as PRO scale - consisted of a list of 11 items to be answered; the respondent provides her personal perception by checking one of 5 possible boxes of “severity” for each of the 11 items. Research in the coming years provided evidence for a good reliability and validity (see later).

The concept of the new scale was to measure health-related quality of life (HRQoL) and certain domains specifically. Numerous publications showed that women experience during the period of menopausal transition very often impaired quality of life due to a bundle of complaints related to vegetative sensations, problems with bladder and urinary system, impaired sexual relations including discomfort during intercourse, increased problems with muscles and joints as well as heart sensations/problems. Several symptom listings, including the Kupperman scale, existed in the literature. Deriving from this knowledge it was anticipated that at least three domains could be found: The known menopausal symptoms are predominantly clustered in psychological, physical (vegetative) changes as well as sexual problems.

The purpose of the development was a self-administered scale to (a) assess the impact of symptoms or complaints related to menopause transition (not disease-related) in women, (b) compare the severity of symptoms over time or between different groups, and (c) measure changes pre- and post treatment. The scale was developed in response to the lack of fully standardized scales to measure the severity of menopausal symptoms.

1.2. Intended application

A self-administrable scale to evaluate patient-reported complaints was intended that is easy to apply, has a high acceptance in women, is easy to score/to analyse, and can be applied in clinical trials as well in observational studies to describe a group, to compare results among groups (different locations or over time), and can be utilized as outcome measure in clinical or drug research.

1.3. Intended population

The scale should be applied in women irrespective of age, but with preference for the menopausal transition. Later, after use in the practice it became clear, that ethnicity, social class (city type/population density), and health status obviously do not limit the application. This was observed when analyzing the results of MRS in observational
surveys in many countries worldwide.⁶ No major differences in the internal structure of the MRS scale across different countries were observed.

2. Creation of the MRS instrument

2.1. Generation of items

The initial step to collect the symptoms relevant for a scale to measure menopausal symptoms was a review of the pertinent literature, mainly from Kupperman and earlier publications of Werner⁷ who described symptoms experienced by men and women after 40 years of age. The selection process, however, was mainly based on informal discussions and clinical experience of representatives of the German Association for Gynecology & Obstetrics in the early 1990s and later on for the revision of the scale in the late 1990s. Unfortunately, there are no (un-)published minutes/documents from these discussions available. Although there were many studies available related to menopausal symptoms a specific study of patients was not done in direct preparation of the MRS scale in this early time. To a great extend, the initial selection of symptoms/complaints for the MRS scale was mainly based on expert opinion.

Using factorial analysis the various complaints or symptoms were put into perspective. Symptoms or complexes of inter-correlating symptoms formed the ‘dimensions’ or domains.

Three dimensions were identified explaining 58.8% of the total variance and got the following names: psychological, somato-vegetative, and urogenital dimension. A five-point rating of severity was applied to each of the eleven items of the MRS: 0 (no symptoms) or up to four scoring points (severe symptoms), i.e. depending on the severity of the complaints perceived by the women completing the MRS.

2.2. Choice of data collection

The scale was designed as paper-based, self-administrable scale. No investigations were performed, as far as we know, how an electronic or web-based administration would influence the results. We consider it as not likely that substantial difference of response behaviour could be found in other applications as the self-administrated, paper-based version; however, we cannot exclude the possibility.

2.3. Choice of recall period

The goal of the scale is the assessment of health-related QoL in the actual or recent time. We assumed that the time span of “at this time” varies according individual perception as does a fixed time period (such as “last month”). We resisted following many critical remarks coming from cognitive debriefing or otherwise during the translation process into other languages to give a fixed time period, because the standardisation of the test was based on the period described as “recent/this time”. We ear-marked this issue for a possible revision of the scale. However, this should be followed by a new series of new or re-validations and re-translations into (currently) 24 language versions. This task was repeatedly postponed because of lacking sources, i.e. it will be very time consuming and costly.
2.4. Response options

A Likert scale (ranging from 0 to 4) is used to document the response. The patient/woman has to check for each of the items if the symptoms/complaints apply for him, and if so, how severe/intense or strong they were perceived. The appropriate box has to be marked. The direction of responses does not vary, it goes from 0 (= no, not applicable) to 4 (very severe). No formal investigation / analysis were done concerning floor or ceiling effects, although data are available. We always found in larger studies a few patients with the minimal number of scoring points, but almost never reaching the maximal possible number of scoring points. This analysis could be formally done, if requested, based on the large number of study databases available (not done or published yet).

2.5. Evaluation of patients understanding

Understandability of the final MRS items was not formerly investigated, i.e. not documented for the German version. However, there is experience collected in population survey and treatment studies with patients in Germany, but not documented or published. There was no indication to assume that the scale makes problems with understanding. In contrary, most patients completed the scale without interruption or queries or problems in less than 10 minutes.

The first formal cognitive debriefing was performed as part of the linguistic and cultural adaptation into English language, (12 women): 4 out of 12 women had comments for one or two items as how to phrase it better. No item was deleted (not permitted because the original scale was the German one) and none needed substantial re-phrasing, because suggestions were very marginal. The English MRS version was the preferred source for almost all translations into other languages. Almost all investigations in the frame of the linguistic & cultural validation into numerous languages had a cognitive debriefing as part of the procedure. However, this information is not openly accessible.

2.6. Development of format

The scale was planned as paper-based scale with a short instruction and a Likert scale from 0 to 4 to describe the personally perceived severity (or intensity) of the complaints (items). All items were phrased in a negative direction (complaints). The development of the MRS scale was described above. The formal standardization of the original German scale was done with a large normative population survey. Population survey were also done in several other countries around the world. Some of these surveys underwent also a factorial analysis that confirmed similarity of the structure (domains) with the initial German investigation (see methods publication).

2.7. Identification of preliminary scoring of items, domains

Following our general intention to develop a simple instrument for practical use, we decided to give each intensity (severity) grade of complaints one extra scoring point. If items were not applicable “No” (score =0) was coded, if the symptom was present, one of four intensity grades have to be chosen (scores 1…4).
The score marked in the scale is identical with the scoring points that are added up to get the total (or domain) scores. The **total score** of the MRS ranges from 0 (asymptomatic) to 44 (highest degree of complaints). The minimal/maximal scores vary between the three dimensions, depending on the number of complaints allocated to the respective dimension of symptoms:

- **psychological** domain: 0 to 16 scoring points (four symptoms: depressed, irritable, anxious, exhausted),
- **somato-vegetative** domain: 0 to 16 points (four symptoms: sweating/flush, cardiac complaints, sleeping disorders, joint and muscle complaints)
- **urogenital** domain: 0 to 12 points (three symptoms: sexual problems, urinary complaints, vaginal dryness).

### 2.8. Assessment of respondents and administrator burden

The scale is applied as self-administrable instrument. The majority of women can complete the scale in less than 10 minutes without problems. It is very unusual that women reject completion of the scale because it is too sensitive, or too private, or otherwise subjectively not acceptable.

As far as known, no burdens for the administrator of the scale were reported; the scale is supposed to be self – explanatory and easy to apply. Moreover, the scoring is simple, i.e. the score increases point by point with increasing severity of subjectively perceived complaints in each of the 11 items (severity 0…4 points). By ticking one of 5 possible boxes of “severity” for each of the items the respondent provides her personal perception. The composite scores for each of the domains (sub-scales) are based on summing up the scores of the items of the respective dimensions. The composite score (total score) is the sum of the domain scores. The three domains, their corresponding items and the evaluation are detailed in an evaluation sheet for paper-pencil evaluation (if not the computer analysis is preferred). There are reference (norm) values available to compare the results with. Norm values for several countries are available but not published.

The very broad dissemination of the scale, i.e. the translation in currently more than 25 languages (current status see [www.menopause-rating-scale.info](http://www.menopause-rating-scale.info)) speaks not only in favour of the methodological properties but also for easy application or little burden for patients and investigators.

### 2.9. Confirmation of conceptual framework & finalization of instrument

The 11-item scale meets the planned conceptual framework according to results of several factorial analyses. The domains found in the normative population survey mirror the expectations of the concept. This would be compatible with the notion that the MRS scale and its domains have good **face validity** (see validity).
3 Assessment of measurement properties

3.1 Reliability

The assessment of scientific measurements depends first of all on their repeatability, i.e. consistency of results and test–retest reliability. In contrast to systematic and random variation, reliability gives an estimate of method-related measurement error, which should be low, so as not to hide or dilute intended systematic changes – due to treatment for example.

The internal consistency measured with Cronbach’s alpha ranges between 0.6 and 0.9 across countries, for the total score as well the scores in the three domains. This is a very acceptable consistency of the MRS. Moreover, there is no evidence that the scale works differently in so many different countries of four continents.\textsuperscript{6}

The test–retest correlation coefficients (Pearson’s correlation) support the suggestion of a good temporal stability of the total scale and its three domains. The test–retest coefficients of the total score range between (Pearson’s correlation coefficient) $r = 0.8$ and $r = 0.96$ across Europe, North and Latin America, and Asia.\textsuperscript{6} This confirms earlier results.\textsuperscript{10,11} When it comes to the subscales with much fewer items, the variation increased and some of the coefficients went down to 0.5 (e.g. urogenital domain in Indonesia). Altogether, the test–retest stability over a time period of two weeks aggregated at the international level supports the notion of a very acceptable test–retest reliability of the total scale and their three subscales.

Although there is an impressive set of information currently available concerning the reliability of the MRS, there are also limitations: small sample sizes prevent a final conclusion regarding test–retest reliability in some of the language transfers, and no information is currently available yet for several other languages.

Altogether, the internal consistency reliability and the test-retest stability support the notion of a acceptably good reliability of the total scale and their three domains. This speaks also in favour of the quality of the linguistic and cultural adaptation in the respective languages.

3.2 Validity

While reliability assesses the consistency of measurement, validity estimates whether or not a HRQoL scale really measures what it intends to measure. In contrast to reliability, which can be determined straightforwardly with very few indicators, validation requires a continuing process (construct validation). It is a process of accumulating evidence for a valid measurement of what is intended.

\textit{Stability of internal structure of the MRS across countries}

Several analyses were performed since the first factor analysis in 1996 was applied to identify the dimensions of the scale. In three large samples studied until 2003, an astonishingly similar structure of the MRS was observed across time and different populations. Similar factor loadings (correlations) of the 11 items in the three domains suggest that the scale constantly measures the same phenomenon across regions.\textsuperscript{6}
Sub-scores and total score correlation

One important issue relates to the independence of the three subscales from the aggregate total scale: theoretically, the correlations between subscales (supposed to be independent due to the statistical model) would be closer to 0 than the correlations with the total score to which all subscales should significantly contribute. But only a somewhat lower correlation among subscales ($r = 0.4–0.7$) was observed, when compared to the correlations of subscales with the total score ($0.7–0.9$). This suggests that the subscales are not as independent from each other as one would expect them to be – based on a factor analysis with assumed orthogonal factors. The situation was similar in the four regions analysed, and in the individual countries belonging to these regions. This need to be considered when analyzing the results of the scale – often the total score might be preferred or utilized for interpretation.

Comparability of the scores among countries

Currently, MRS population reference values for severity of complaints were published only for the German population. Are these reference values applicable for other countries or cultures? The data of the large, multinational survey gave some indirect evidence: The mean values (and standard deviations (SD)) of the MRS total score and the three domains were not statistically significantly different between Europe and North America. Thus, there is no reason why MRS values between these regions may not be compared to each other.

However, the domains seem to slightly differ in some regions: Total, psychological and somatic scores were systematically higher in Latin America and systematically lower in Asia (Indonesia) than in Europe or North America. The urogenital scores were somewhat lower in Latin America or Indonesia than in Europe or the US. Obviously the perception of the prevalent symptoms depends on cultural factors – or the symptoms show real differences in prevalence. Thus, a direct comparison of MRS domain-scores between Europe or North America on the one side, and regions in Latin America and Asia on the other, should only be done with caution. This does not affect intra-individual comparisons (e.g. pre-/post-therapy), and it may have very little impact on the comparison of relative changes (pre-/post-treatment) among different countries. The latter is a working hypothesis and needs further research.

Concurrent validity: cross-validation with other scales

In order to further evaluate the MRS for scoring menopausal symptoms, it was compared with other instruments relevant for women in their menopausal transition: the Kupperman Index and the generic QoL scale SF-36.

A comparison of the MRS with the Kupperman Index produced a high correlation of raw scores (Pearson’s $r = 0.91$). The highest association of scores was found in the highest quartile (80%) of the MRS. The terms ‘mild’, ‘moderate’, and ‘severe’ for the degree of severity of menopausal symptoms reflect different contents in both scales, and are differently spread, i.e. they are not directly comparable.

There is a strikingly good association between the subscales of SF-36 and the MRS. The MRS correlates best in dimensions of the SF-36, which are highly relevant for women in
the menopausal transition.\textsuperscript{11} For that reason, the MRS can be utilized as an age- and condition-specific QoL instrument.

The MRS is a valuable modern tool for the assessment of menopausal complaints. It combines in practice excellent applicability, good repeatability, and there are ‘norm values for the population’ available. The MRS could serve as an adequate diagnostic instrument for menopausal QoL.\textsuperscript{6}

The next section reviews data about the utility of the scale to detect and measure treatment effects.

3.3. \textit{Ability to detect treatment-related changes}

\textit{Discriminative validity: detection of treatment effects}

A recent report collected data of 9300 women in a pre- and post-treatment experience with the MRS. This served as a basis to critically evaluate the capacity of the scale to measure the effect of hormone treatment independent from the severity of complaints.\textsuperscript{13}

The improvement of symptoms during treatment was equivalent to 36\% of the baseline score on average, and was similar for all three subscales. Thus, the MRS was successful in detecting treatment effects.

Was the MRS also sufficient to detect treatment effects in women with only few or mild symptoms? An improvement of complaints or QoL was seen with increasing degree of symptom severity: patients with few/no complaints before therapy improved by 11\%; those with mild complaints at entry by 32\%; those with moderate complaints by 44\%; and those with severe symptoms by 55\%. The scale does detect treatment effects in women with few complaints, albeit to a lesser degree.

That means that the MRS scale showed a convincing ability to measure treatment effects on quality of life across the full range of severity of menopausal complaints. The MRS scale was also tested whether it predicts the subjective, clinical impression/opinion of the treating physician. The sensitivity (correct prediction of a positive assessment by the physician) was 70.8\% and specificity (correct prediction of a negative assessment by the physician) 73.5\%. In other words, the MRS scale fits well with the subjective assessment of the treatment by the physician. The problem of this study was that the MRS scale was not self-administered but interviewed by the physician.

Consequently, a new study based on 3282 women with pre- and post- hormone -treatment data was published 2006.\textsuperscript{14} The application of the MRS scale differed compared to the previous study: self-administered in the latter vs. interview by physician in the former. The hormone-therapy related improvement of complaints relative to the baseline score was about or less than 30\% in total or domain scores, whereas it exceeded 30\% improvement in the old study.\textsuperscript{13} Similarly, the relative improvement after therapy, stratified by the degree of severity at baseline, was lower in the new than in the old study, but had the same slope. Although one cannot exclude different treatment effects with the study method used, this supports our hypothesis that the individual MRS interviews performed by the physician obviously biased the results.
towards over-estimation of the treatment effects. This hypothesis was underlined by the
degree of concordance of physician’s assessment and patient’s perception of treatment
success (MRS results): Sensitivity (correct prediction of the positive assessment by the
treating physician) of the MRS and specificity (correct prediction of a negative
assessment by the physician) were lower than the results obtained with the interview-
based MRS scale in the previous publication but still in a sufficiently good level.

Although the new study confirmed evidence for the capacity of the MRS scale to
measure treatment effects on quality of life across the full range of severity of
complaints before treatment, the authors concluded that the MRS scale should be used
as self-administered instrument in clinical studies to avoid overestimation of treatment
effects.

3.4. Choice of method for interpretation

Theoretically, there are two options to assist interpretation of findings with the MRS
scale:

- Comparison with norm (= reference) values from the respective population, i.e.
  how is the severity of complaints before treatment and thereafter compared with
  the population reference.

  The comparison with population reference values can be very useful for
  observational or comparative surveys, e.g., for prevalence studies. However, it is
  only the second-best approach for the interpretation of results of clinical trials with
  the MRS scale as outcome.

  Currently, there are norm values published for the German MRS version.
  Population-based surveys were also performed in other countries (see MRS
  website) but the respective reference values were not published or but could be
  done on request since data form the surveys are available.

- The evaluation of the (relative) improvement of AMS scores (compared with
  scores before treatment as described above) is the most relevant method,
  particularly for the interpretation of clinical outcome studies such as RCTs.

4 Modification of the instrument

Based on the experience gathered during the validation process of the current scale, a
modification of the MRS scale is not planned for the next future. This is not only due to
lacking sources but mainly caused by the overall good results of the validation studies of
the current version.


Werner AA. The male climacteric. JAMA 1939; 132:1441-3.

Schneider HPG, Heinemann LAJ, Thiele K. The Menopause Rating Scale (MRS): Cultural and linguistic translation into English. Life and Medical Science Online 2002;3: DOI:10.1072/LO0305326


