Aging Males Symptoms Scale (AMS)
Development of the scale[1]

Status report by the developer of the scale (L.A.J. Heinemann).
June 2006

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 AMS 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.

The interest of clinical research in aging males increased in the mid-1990ies and thereby the interest to measure health-related quality of life.

We developed the scale under the assumption that men during aging are less aware of the fact that they undergo some kind of "transition" as women in the age when experiencing menopausal transition, i.e. men tend to overlook new symptoms usually, unless specifically ask for. Therefore, the first step was to carefully analyze the age-dependency of symptoms or complaints in large representative population surveys (we were involved in). The results confirmed our hypothesis and supported the development of a scale for aging males.[1]

Since aging is in later phases associated with a higher prevalence of lower testosterone level, we specifically looked for symptoms that are used in clinical practice. The only symptom list we found was the symptom list of Prof. Vermeulen (Ghent, Belgium). It had high reputation in clinical practice, was used by many physicians, but was neither validated nor published. Vermeulen, an internationally acknowledged expert in hypogonadism and androgen research, became later involved as expert in the development of the new scale. He had discussions with patients about symptoms of aging and decreased testosterone level on a daily basis and supported the right phrasing to be understandable for patients. He also insisted to get the symptoms relevant for hypogonadism included into the new scale.
The concept of the new scale to be developed was to measure health-related quality of life (HRQoL). The purpose of the development was a self-administered scale to (a) assess the impact of symptoms or complaints related to aging (not disease-related) in men, (b) compare the severity of symptoms over time or between different groups, and (c) measure changes pre- and post treatment. It was developed in response to the lack of fully standardized scales to measure the severity of aging symptoms and their impact on HRQoL in males, specifically.

[1] Drafted following the structure of: Guidance for Industry. Patient-Reported Outcome Measures: Use of Medical Product Development to Support Labelling Claims. FDA February 2006 (www.fda.gov/cder/guidance/5460dft.pdf; released on February 2, 2006)

It was anticipated that at least three domains could be found deriving from personal experiences gathered with standardized instruments in women, e.g. with the Menopause Rating Scale (MRS). The symptoms we found related with aging pointed in the direction of domains related with psychological and physical (vegetative) changes as well as sexual problems.

1.2. Intended application

A self-administrable scale was intended that is easy to apply, has a high acceptance in men, is easy to analyse, and can be applied in clinical trial 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 studies.

1.3. Intended population

The scale can be applied in men irrespective of age, but with some preference for men over 40. Later, after application in the practice it became clear, that ethnicity, social class (city type/population density), and health status obviously do not play a major role (see also further down; 2.9.). This was observed as when analyzing the results of AMS in observational surveys in many countries.[2]
No major difference in the internal structure of the AMS scale was observed between men with or without decreased testosterone level (hypogonadismus).[2,3]

2. Creation of the AMS instrument

2.1. Generation of items

The first step was to analyze the age-dependency of symptoms and complaints in large population-based observational surveys we were involved in some time ago, as mentioned above. The results were published later[1]. This led to the first draft list of possible items for a new scale. To our surprise the symptoms did not differ so much between men and women. On the other hand, this was already discussed decades ago: Unlike in women, where increasing prevalence of complaints with aging is widely accepted as associated with ‘menopause’, complaints in aging male's are rarely put into perspective of the hormonal involution, although Werner[4] reported already in the 1940s about the similarity of male complaints with those of women in this age span. Arguments pro and contra a male climacteric as well as compatibility with the female situation were discussed ever since then.[5,6,7] Later, we observed in a cross-sectional sample survey that the various types of sweating do not show differences in frequency between males and females during lifetime.[8]

The specific raw symptom inventory was based on (1) a short check list of complaints suggested by Vermeulen, and (2) additional symptoms with an age-dependent increase of prevalence observed in the pooled analysis of the population-based observational studies, mentioned above. This advanced list of symptoms and complaints was applied to 10 patients in a general practice in order to learn if they understand meaning and wording but no revision was needed (unpublished). This more comprehensive check list of symptoms or complaints consisted of 21 items. To identify symptoms related to aging and hormonal changes, not other diseases or conditions, we supplemented many other questions to the “Baseline Data Form” to be applied in a sample survey.[9]

We took a sample of medically well characterised aging males over 40 years from GP practices. Patients with minor diseases or early stages of medical conditions were included, but banal infections were preferred. These patients completed our draft symptoms inventory of 21 suspected 'aging males' symptoms'. In addition to these 'specific symptoms', we tried to characterize the group regarding health care utilization, attitudes toward a healthy life style or prevention, disease history, use of certain drugs, some socio-economic characteristics, and a some laboratory data, which might partly explain the so-called 'specific symptoms'. Almost 200 parameters were included into this survey database.[9,10] The survey data were used to develop the final scale and to preliminarily describe their psychometric characteristics.

Using factors analysis the various complaints or symptoms were put into perspective with aging, health problems or stressful situations the patient is exposed to and other parameters, to allow meaningful clustering of parameters/symptoms. Symptoms or complexes of inter-correlating symptoms formed the 'dimensions' or domains.

We found the following dimensions[9,10]: Psychological (5 items), somato-vegetative (7 items), and sexual symptoms (5 items) forming clusters with high loading in the domains (dimensions). Since these three factors correlated particularly with age and not with diseases or other conditions, we called the three "AMS factors". Other factors were found to be related with diseases or conditions and were considered as irrelevant for the AMS scale. Thus, the specific part of the raw scale was reduced from 21 to 17 items, the proposed AMS scale.

Degenhardt and Schmidt[11] reported for aging males using factorial analysis two factors similar to those we found: '*psychological syndrome* of energy loss' and a second factor '*vegetative* and vasomotor dysfunction' and mentioned in their paper that sexual symptoms had no clear association with the two factors mentioned.[9] Very similar dimensions were described for aging women (Menopause Rating Scale (MRS).[12] which has been successfully used in medical practice to monitor hormone replacement for more than a decade.

The final AMS scale was then applied to a population random sample of over 992 males in order to get reference (norm) values for German males aged 40-70 years. 959 questionnaires were available for analysis. More details were published.[9,10]

Understandability of the final AMS items was not formerly investigated, i.e. not documented, for the German version, with the exception of interviews with patients in the pilot phase prior to the execution of the normative sample survey in GP practice (see above). This was not a formal investigation and changes were not documented. However, there was rich experience collected in population survey and treatment studies with patients in Germany. 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 (19 males): 1-6 out of 19 males had problems with understanding of one or more items or provided suggestions as how to phrase it better. No item was deleted (not permitted because the original scale was the German one) but many items were re-phrased. The English AMS version was the preferred source for

almost all translations into other languages. The simple forward translation in the English original publication of the AMS scale[9] was never used in practice.

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 (depends on disease, treatment etc): give period as explicit as possible

The goal of the scale is the assessment of health-related QoL in the recent time. We assumed that the time span of "recent" 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 time". We ear-marked this issue for a future revision of the scale 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 1 to 5) is used to document the response. The patient has to check for each of the items if the symptoms/complaints apply for him, and if so, how severe/intense or strong it was perceived. The appropriate box has to be marked.
The direction of responses does not vary, it goes from 1 (= no, not applicable) to 5 (very severe).
No formal investigation / analysis were done concerning floor or ceiling effects. We always found in larger studies a few patients with the minimal number of scoring points, and even less that reached 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

The long raw scale was discussed with patients of a GP before the normative sample survey (see above). There was no formal analysis done for the German version of the final scale, but many investigations in the frame of the linguistic & cultural validation into numerous languages (cognitive debriefing). This and the consequences for a future revision of the scale were shortly discussed (see also methods publication[2]).

2.6. Development of format

The format of the raw and final scale was planned as paper-based scale with a short instruction and a Likert scale from 1 to 5 to describe the personally perceived severity (or intensity) of the complaints (items). All items were phrased in a negative direction

(complaints) following experiences with the development of the Menopause rating Scale (MRS)[12].
The development of the final AMS scale was described above. A patient sample was chosen to have well validated information also about diagnosis, treatment, laboratory data for the correlation with symptoms/complaints and thereby for the intended item reduction with factorial analysis. Four items that were considered as probably specific items were skipped in the multivariate item reduction process as not applicable but no other items were found in the questionnaire to be additionally included.[9,13]

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 =1) was coded, if the symptom was present, one of four intensity grades have to be chosen (scores 2…5).
The score marked in the scale is identical with the scoring points that are added up to the total (or domain) scores.

2.8. Assessment of respondents and administrator burden

The scale is applied as self-administrable instrument. The majority of men can complete the scale in less than 10 minutes without problems.
It is unusual that men reject completion of the scale because it is too sensitive, or too private, or otherwise inconvenient.

No burdens for the administrator of the scale were reported; the scale is supposed to be self - explanatory.
Moreover, the scoring scheme is simple, i.e. the score increases point by point with increasing severity of subjectively perceived complaints in each of the 17 items (severity 1…5 points). By checking one of 5 possible boxes of "severity" for each of the items the respondent provides his personal perception. The composite scores for each of the domains (sub-scales) is based on adding up the scores of the items of the respective dimensions.[2] 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 – unpublished except for Germany and France; see above).

The very broad dissemination of the scale, i.e. the translation in currently more than 21 languages[14,15,16] speaks not only in favour of the methodological properties but also for easy application or little burden for patient and administrator.

2.9. Confirmation of conceptual framework & finalization of instrument

The 17-item scale meets the planned conceptual framework according to results of several factorial analyses:
The items belonging to each of the three domains were highly correlated with the respective domain. Moreover, they were reasonable from the biological standpoint, i.e.

all complaints were known as associated with increasing age and decrease of male hormones. All items are relevant although the magnitude of correlation with the total score and with the respective domain scores differed. This would be compatible with the notion that the AMS scale and its domains have good *face validity*.

Later, an analysis of a French observational study (903 men)[17] demonstrated that the results are similar for men under and over 40 years of age, and not different from a large, pooled European database.[2, 14] Earlier it has been shown, based on German studies, that the internal structure of the AMS scale does not differ between men with and without hypogonadism.[2] In addition, the norm values for under and over 40-years old men seem not to differ if one account for the clear trend of increasing severity of complaints with increasing age.[14]
As mentioned above, it became clear, that ethnicity, social class (city type/population density), and health status obviously do not play a major role.[2] Increasing age however is significantly associated with higher AMS scores (total, somatic and sexual) as shown in a French population-based survey.[14] Higher family income, however, was significantly associated with lower AMS scores in the same study. BMI was not significantly associated with the AMS score according to the French study but showed a clear impact in an Austrian study[18] and a mixed Austrian/German study[19].

As mentioned above (cf. 2.4), the distribution of the response range in patients was not yet formally analyzed but can be done on request. The distribution of the responses was sufficiently varying across patients/respondents, i.e. the extremes are much less frequent than the middle part of the score distribution (unpublished).

We became not aware of problem related floor or ceiling effects although numerous surveys and also clinical studies were performed. However, since we are not aware of any formal investigation, we cannot exclude the possibility of such phenomena.

The psychometric characteristics of the AMS scale are good; there is rich information from many surveys in many countries partly included in a summary publication.[2]
The reliability was analysed as internal consistency reliability (Cronbach's alpha) as well as test-retest stability across many countries and found to be good (cf. next chapter, point 3.1). The same applies for the validity parameters (cf. 3.2). This includes also studies on responsiveness (utility as outcome measure) that are detailed further down (cf. 3.3).[20,21]

3 Assessment of measurement properties

3.1. Reliability

Reliability investigates to what extend measures are internally consistent and results of the scale more or less identical if the scale is repeatedly administered. In contrast to systematic and random variation, reliability gives an estimate of method-related measurement error which should be low not to hide/dilute systematic changes – due to treatment for example.

The internal consistency reliability measured with Cronbach's Alpha varied between 0.7 and 0.9 across countries and time periods.[2] This applied for total score as well as the three subscales. This is indicative for a very acceptable consistency of the AMS scale. No clear evidence was found that the scale works different across countries in Europe and Asia.[2]

The test-retest correlation coefficients (Pearson's correlation) support the notion of a good temporal stability of the total scale (r = 0.8-0.9) and its three sub-scales (r = 0.5-0.9), although most of the assessments across countries are based on very small numbers.

Altogether, the internal consistency reliability and the test-retest stability support the notion of an acceptable 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

Validity is a measure that describes to what extend a scale measures what it intends to measure. The initial information was available in 1999 when the scale was originally published.[1] But whereas reliability can be determined straight forward with very few indicators, the validity is almost always a continuous process (construct validation). It is a process of accumulating evidence for a valid measurement of what is purposed. Therefore, the currently available data are already fairly comprehensive and do pave the way for a focussed and continuing validation process.

Stability of the internal structure of the AMS scale

The first factorial analysis in 1996 was applied to identify the dimensions of the scale. Three dimensions of symptoms/complaints were identified [3]: a psychological, a somato-vegetative, and a sexual factor that explained 51.6% of the total variance. The three domains found in this analysis did fit theoretical expectations, seemed to be plausible, and met the conceptual framework – as discussed above. We concluded that the AMS scale have a good face validity (or content validity).

Since then, two independent, large community sample surveys were performed in German males, i.e. in 1997 and 2003.[2] The loadings of the 17 items on the 3 factors fully confirmed the structure found in the initial factor analysis. This suggests that the scale measures constantly the same phenomenon over time in different locations. However, two items did not contribute enough to the respective factors and showed differences: item 12 ("Feeling that you have passed your peak") and item 14 ("Decrease of beard growth"). This was ear-marked for a later revision of the scale, but kept currently unchanged, because otherwise standardization and norms will not be applicable anymore for a revised scale.

The comparison of the internal structure between the large German sample surveys and the samples from other countries showed mainly close similarities of the internal structure across countries but also a few potential problems (cf. methods paper[2]). The question is, whether this is a problem of very small samples with random error or expresses remaining cultural differences.
If some of the differences could be confirmed with larger samples this would be a reason to carefully re-phrase the items in a future revision.

The possibly slight differences in the internal structure of the AMS scale across countries suggest further research but does not invalidate the comparison in clinical studies across countries or even prevent pooling in multinational studies, because intra-

individual comparisons over time (before/after treatment) are the main criterion which might not be affected very much. It cannot be excluded however that the real efficiency of a given treatment measured with the AMS could be diluted and thereby underestimated. But this remains speculation until larger samples confirm the above findings.

All the same, the factor analyses across countries and time periods provided evidence of a sufficiently compatible internal structure of the AMS scale and suggest thereby good validity for daily practice.

However, it is also important to consider if the AMS scale is similarly applicable in patients with hypogonadism than in the normal male population. The internal structure of the scale was analysed in *males with androgenic dysfunctions* was obviously identical with the "normal population", but the three domains explained more of the total variance in hypogonadal men than in normal male population (65% as compared with around 55%). This is indicative for a particular sensitivity to measure health-related complaints in hypogonadism.

The results of the structural analyses are compatible with the notion that the AMS scale can be used in men with and without hypogonadism as well as under or above 40 years of age (see above).

Sub-scores and total score correlation

Another structural validity aspect is the total- domain- correlation, i.e. the correlation of the total score with the scores of the five domains. Ideally, there should be significant and high correlations between the total scale and all subscales forming the total scale. In contrast, the correlations among the individual subscales should be smaller, because the subscales are supposed to be "independent" according the factor-analytic model that was used.

But that is theory: We indeed found somewhat lower correlation among the three domains (r = 0.5-0.7) as compared with correlation of sub-scales with the total score (r = 0.8-0.9). This is less different than one would have wished. It suggests that the sub-scales are not as independent from each other as one would expect them to be – based on a factorial analysis with orthogonal factors. This was observed similarly in Germany, UK, Rest of Europe, and Asia.[2] We observed very similar correlation coefficients among countries or country groups. This is suggestive of pretty similar features of the AMS scale across the countries. It is even more important to underline that this pattern is true both for the situation in the normal population (community sample) and in hypogonadal patients with apparent androgenic dysfunctions.

Criterion-oriented validity: cross-validation with other scales

In fact, the comparison with other scales of similar purpose is crucial to justify the potential areas of application and the validity of interpretation of results obtained.

The first claim of the AMS scale is to measure "health related quality of life". This needs to be shown by cross-validation with scales known as generic QoL scales (e.g., SF36). There are also comparisons needed with standard scales that measure specific health-related aspects supposed to be measured by the AMS scale or domains, such as

instruments to measure symptoms in aging males (e.g. Finnish Turku scales), or with scales to screen for androgen deficiency in aging males (e.g., ADAM, MMAS, Smith's scale).

Comparison with generic QoL scale SF36

Since the AMS scale is a health-related QoL scale, comparisons with other QOL scales are meaningful. The AMS scale and the generic QoL instrument SF36 were applied at the same time in 116 German males aged 40 to 70 without serious health problems. The total score and the three sub-scores of the AMS scale were compared with two sub-scales of the SF36. The AMS total sum-score and the two sub-scales of SF36 were statistically significantly correlated: r= -0.49 (n= 116; p<0.0001). The correlation of the somatic sum-score of AMS and the somatic sum-score of SF36 was sufficiently high (r= -0.54; p<0.0001; n=116) as well as the psychological sub-scales of both instruments(r= -0.65; p<0.0001; n=116).[1,10] The correlation is inverse due to the fact that the sum-scores of the AMS increase with numbers (intensity) of symptoms/complaints whereas the sum-scores of SF36 increase with increasing well-being/happiness. But there is no comparator in the SF36 for the sexual sub-scale of the AMS.[10]

Comparison with health-specific scales

Finnish aging male's scales

Regarding health-specific validity of the AMS scale, a Finnish research group observed a strong and statistically significant correlation with their Turku 14-items scale that was proposed to be specific to measure complaints associated with testosterone deficiency (r= 0.8; n=95), and similarly promising results were obtained when comparing with their own "3-Item-Scale".[10] The two scales can be regarded as measuring the same phenomena and this speaks in favour of good test characteristics of the AMS scale.

Screening scales for androgen deficiency

Results of the AMS scale were compared with those of two screening instruments for hypogonadism in 2003: The ADAM scale of Morley at al[22] and the Screener of Smith et al[23]. These two scales were developed to screen for androgen deficiency, i.e. to detect persons where the determination of the testosterone level could be advisable.[24]
The associations between the AMS categories and the (ADAM) or (SMITH's) categories were significant. The AMS predicted the results of the two other tests quite good: AMS predicts ADAM: positive predictive value (pPV) = 92%, negative predictive value (nPV) = 50%, specificity = 97%, but sensitivity only = 29%. Thus, the AMS predicts a positive screening result of the ADAM scale quite well, but less good negative screening results of the ADAM scale. Similarly for the Smith's screener; the respective values for the comparison of AMS vs. Smith's screener are 65% (pPV), 49% (nPV), 87% (spec.), and 22% (sensitivity). The values for the prediction of ADAM results regarding the Smith' screener results are somewhat lower: 57% (pPV), 50% (nPV), 60% (sensitivity), and 46% (specificity), respectively. Just for comparison, Morley [22] reported for a positive ADAM result and a low bioavailable testosterone level (<70 ng/ml) a sensitivity of 88% and a specificity of 60%.

The Morley group published recently a own comparison between AMS, ADAM test and the Massachusetts Male Aging Study (MMAS) scale. [25] The sensitivity for the ADAM was 97%, for AMS 83% and for MMAS 60%, whereas the specificity was low for all instruments (30%, 39%, and 59% for ADAM, AMS, and MMAS resp.). The authors

concluded that the ADAM and the AMS may be useful screening tools for Hypogonadism across adult lifespan.

AMS in the context of a screening tool

Although the AMS scale was not developed as screening instrument for androgen deficiency many complaints included in the AMS scale are common features with androgen-deficient males. As discussed above, the AMS has obviously similar test characteristics as screening instruments[26]. In addition, Kratzik et al[18] observed in a population-based cross-sectional study in Vienna an impressive association between subscales of the AMS and free testosterone level when age and body mass index was taken into account. Recently, a Japanese research groups under Itoh et al[27] and Soh et al[28] observed a correlation between the AMS scores and the testosterone level. A Polish research group found a similar but less clear result[29]. Other studies, however, could not find associations of the AMS scores with testosterone level[30,31].
Although formally outside the conceptional framework, the associations found or missed do have an impact on the interpretation of results in clinical studies. Obviously, there are conditions that influence the outcome in terms of perceived complaints or improvement of symptoms. Age and body shape are among the variables the findings should be somehow adjusted for (or stratified). Another very important issue is the consideration of the "starting point before therapy", which is discussed later.

The aim to develop this composite screening tool (AMS, BMI, and age) was the creation of a simple instrument for pre-selection of subjects that need further medical attention (e.g., blood test for testosterone) or to self-assess the change of complaints (e.g. after therapy) by subjects wherever complaints don't lead to visit of a physician (e.g., self-assessment of patient or subjects).

The quality of this composite screening tool described with sensitivity and specificity to detect a low testosterone level. The sensitivity (correct prediction of positive subjects (=low TT value) was 48.3 and the specificity (correct prediction of negative subjects (unsuspicious, high TT value) was 73.7 in the Austrian sample where the test was developed. The respective values in the independent "validation sample" of patients in Germany were 69.9% and 64.4% for sensitivity and specificity, respectively.

The test was published as a graphic tool[19] but is also available for online use in the official AMS website.
The described test characteristics of the proposed composite "AMS- screener" will be acceptable for mass screening and pre-selection of subjects for further diagnostic work-up. There might be also an application of this tool in the INTERNET for educational purposes or self-assessment of interested patients. Likely it is less useful to apply this screening tool as "diagnostic tool" in the urological practice, where a series of blood tests for testosterone can be easily done. It is suggested to collect own experience with the practical application. This is to confirm or refute the described results obtained with this screener of AD or to narrowing down the scope of possible applications.

Cross-validation with depression scales

Reduced HRQoL is expected to be associated with depression. It can be assumed that the psychological domain of the AMS could measure it. This hypothesis was confirmed in two studies, one in France, and one in Japan.
The French group[32] observed a good correlation between the AMS score and the Hospital Anxiety and Depression Scale (HADS): Pearson r= 0.62. Similarly good correlation with the Beck Depression Inventory was observed from the Japanese Group[33]: Pearson r= 0.79 (total score), r = 0.79 (psychological domain), r = 0.65 (somatic domain), and r = 0.45 (sexual domain), respectively.

A cross-validation of the AMS total score and their sexual domain with scales to measure sexual dysfunction is still lacking.

In summary, the AMS scale was successfully validated in all relevant aspects. This is an important pre-condition for assessing the severity of complaints perceived by patients and thereby meets the guidelines for "investigation, treatment and monitoring of late-onset Hypogonadism in males" that was recently published.[34]
The next chapter reviews data about the utility of the scale to detect and measure treatment effects.

3.3. Ability to detect treatment-related changes

Discriminative validity: detection of treatment effects

In this section, we summarize what information became recently available concerning predictive or criterion-oriented validity, i.e. the ability of the AMS scale to detect or measure effects of therapy with androgens and thereby discrimination between treatment responder / non-responder . To this end, many clinicians use the term "validity" and mean high utility for clinical studies or research. Two therapy studies were published yet that used the AMS as outcome.[20, 21]

It is well established that men with androgen deficiency react with a marked improvement of the HRQoL after testosterone treatment. It was the aim to demonstrate that the AMS scale is able to mirror changes of the HRQoL. The scores of the AMS scale improved after 3 months of testosterone substitution (application as gel or injection in the two studies). The magnitude of improvement of the scores (total and domain scores) reached almost 30% compared with the baseline score, respectively in both studies.
The study results confirmed the expectation: The higher the severity of complaints at baseline (before therapy) the greater was the relative improvement. The AMS scale showed a convincing ability to measure treatment effects on quality of life across the full range of subjectively perceived severity of complaints. An effect was even detected if the complaints were not severe what seems to be an important feature for clinical studies with PROs as outcome measure.

Effect modification by other variables at baseline (age, BMI, and testosterone level) on the "treatment associated effect on AMS score" was not observed in one of the two studies.[21] However, this is an issue for debate because of other observations that showed an impact of age and BMI on AMS scores in surveys without intervention. Therefore, it is still recommendable to consider at least these two variables as possible confounder in clinical studies.

Another issue of validation was to what extend the AMS scale could "predict" the "success" of the androgen therapy independently assessed by the treating physicians (blinded for AMS). Sensitivity (correct prediction of an improvement of complaints) and specificity (correct prediction of a negative outcome or no improvement) are important characteristics for a test that intends to "diagnose" successful therapy or to discriminate between responder and non-responder of the therapy. We have chosen the total score for this purpose, because we got the impression during the process of validation that the domain scores might be not stable enough for clinical therapy studies. As long as not other methodological information (e.g. from RCTs) is available we would generally prefer for the AMS total score as outcome parameter for clinical studies.

When plotting the sensitivity and specificity in a kind of ROC analysis against the degree of improvement of complaints found under androgen treatment (difference between the pre- and post-treatment score on AMS as percent of pre- treatment total score) we observed the highest sensitivity associated with a cut-off point of more than 5% relative score improvement (sens. = 94%) and a steep decline toward a cut-off-point of over 40% relative improvement (sens. = 31%). In contrast, the specificity increased from 24% (cut-off 5% improvement) towards 94% (cut-off 40% improvement). An optimal balance was reached when 22% and more improvement was chosen as cut-off: both sensitivity and specificity were about 70% and thereby sufficiently high.[21]

In other words, the AMS scale is able to assess the treatment success with sufficient good test characteristics, including good validity as utility to measure outcome of therapeutic intervention. It cannot be recommended to use the absolute change in scoring points as outcome, i.e. without calibrating for the baseline value. The underlying ROC curve (sensitivity / specificity) is here very flat. One reason might be that the sensitivity to perceive complaints very much varies among individuals, but we have not done a specific study to confirm or refute this hypothesis..

In absence of other information, we suggest 10-15% improvement of the AMS total score as minimum important difference (MID), i.e. deriving from our currently still limited experience with clinical studies on androgen treatment with the AMS scale. This, although the mean change after treatment was about 30% improvement of the AMS total score and the data showing a high sensitivity but low specificity for an improvement over 10% compared with baseline. It is not likely to miss positive treatment effects using this MID, but the specificity is low. Currently, explorative analyses of study data are warranted.

It is a methodological advantage to measure the "treatment success" with the comprehensively validated AMS scale because it is directly based on patients' reports of their symptoms and HRQoL, whereas the physician's evaluation could be rather a varying mixture of theoretical expectation/ experience and patient's report and therefore not recommendable.

3.4. Choice of method for interpretation

Theoretically, there are two options to assist interpretation of findings with the AMS 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 AMS scale as outcome.

Currently, there are norm values published for the German and the French AMS version (less than 40 years, and equal/over 40 years of age). Population-based surveys were also done in other countries (see AMS website) but population reference values were not published or are not easily accessible.

- More important, particularly for interpretation of clinical outcome studies, such as RCTs, is the evaluation of the relative improvement of AMS scores (compared with scores before treatment) as described above (cf. 3.3.).

4 Modification of the instrument

Based on the experience gathered during the validation process of the current scale, a modification of the AMS scale is planned but no time schedule discussed yet. This is not only due to lacking sources but mainly caused by the overall good results of the validation studies of the current version. If it comes to planning of a revision of the AMS scale, reasonable points could be – beside others: A more comprehensive introduction of the scale (with a defined recall period), shortening of the scale by elimination of a few items. But then a complete re-validation would be needed that might lead to further changes. In addition, new validated language versions would be needed.

References

1 Heinemann LAJ, Thiel Ch, Assmann A, Zimmermann T, Hummel W, Vermeulen A. Sex differences of „climacteric symptoms“ with increasing age? A pooled analysis of cross-sectional population-based surveys. *The Aging Male* 2000; 3:124-131

2 Daig I, Heinemann LAJ, Kim S, Leungwattanakij S, Badia X, Myon E, Moore C, Saad F, Potthoff P, Do Minh T.The Aging Males' Symptoms (AMS) scale: Review of its methodological characteristics. Health and Quality of Life Outcomes 2003; 1:77 (December 2003). <http://www.hqlo.com/content/1/1/77>

3 Myon E, Martin N, Taïeb C, Heinemann LAJ. Experiences with the French Aging Males' Symptoms (AMS) scale. Aging Male 2005;8(3/4):184-189.

4 Werner AA. The male climacteric: report of 273 cases. *JAMA* 1946; 132:188-94.

5 Bauer J. The male climacteric – a misnomer. *JAMA* 1944; 126: 914-18.

6 Degenhardt A. Wechseljahre beim Mann, gibt e sie? In: Fischer S, Streb-Lieder & Vogt, eds. Wechseljahre für Fortgeschrittene. Frankfurt: Verlag für akademische Schriften. 1995: 104-18.

7 McKinlay JB, Longscope CH, Gray A. The questionable physiological and epidemiologic basis for a male climacteric syndrome: preliminary results from the Massachusetts Male Aging Study. Maturitas 1989; 11:103-5.

8 Heinemann K, Saad F. Sweating attacks: Key Symptom in Menopausal Transition only for Women? Eur Urology 2003;44 :583-7

9 Heinemann LAJ, Zimmermann T, Vermeulen A, Thiel C, Hummel W. A new, aging males' symptoms' rating scale. The Aging Male 1999;2:105-114.

10 Heinemann LAJ, Saad F, Pöllänen P. Measurement of Quality of Life Specific for Aging Males. In: Schneider HPG (Ed.) Hormone Replacement Therapy and Quality of Life. Parthenon Publishing Group. London, New York, Washington. 2002: 63-83.

11 Degenhardt A, Schmidt H. Physische Leistungsvariablen als Indikatoren für die Diagnose 'Klimakterium Virile'. Sexuologie 1994; 3: 131-41.

12 Potthoff P, Heinemann LAJ, Schneider HPG, Rosemeier HP, Hauser GA. Menopause-Rating Skala (MRS II): Methodische Standardisierung in der deutschen Bevoelkerung. Zentralbl Gynakol 2000; 122: 280-286.

13 Heinemann LAJ, Saad F. The aging male's symptoms scale. In: Schneider HPG (Ed.) Menopause. The state of the art – in research and management. The Parthenon Publishing Group. Boca Raton, London, New York, Washington. 2003: 319-324.

14 Heinemann LAJ, Saad F, Zimmermann T, Novak A, Myon E, Badia X, Potthoff P, T'Sjoen G, Pöllänen P, Goncharow NP, Kim S, Giroudet C. The Aging Males' Symptoms (AMS) scale: update and compilation of international versions. Health and Quality of Life Outcomes 2003;1:15 (1 May 2003). http://www.hqlo.com/articles/browse.asp

15 www.aging-males-symptoms-scale.info (official AMS website; regularly updated)

16 Heinemann LAJ. Aging males' symptoms: AMS scale – a standardized instrument for the practice. J Endocrinol Invest 2005; 28: 34-38.

17 Myon E, Martin N, Taïeb C, Heinemann LAJ. Experiences with the French Aging Males' Symptoms (AMS) scale. Aging Male 2005;8(3/4):184-189.

[18] Kratzik CW, Reiter WJ, Riedl AM, Lunglmayr G, Brandstätter N, Rücklinger E, Metka M, Huber J. Hormone profiles, Body Maß Index and Aging Male Symptoms – results of the Androx Vienna Municipality study. *Aging Male* 2004;7:188-96.

[19] Kratzik C, Heinemann LAJ, Saad F, DoMinh T, Rücklinger E. Composite screener for androgen deficiency related to the Aging Males' Symptoms scale. Aging Male 2005;8 (3/4):157-161.

[20] Moore C, Huebler D, Zimmermann T, Heinemann LAJ., Saad F, Do Minh T. The AMS scale as outcome measure for treatment of androgen deficiency. Eur Urology 2004;46:80-7.

[21] Heinemann LAJ, Moore C, Dinger J, Stoehr D. Sensitivity as outcome measure of androgen replacement: the AMS scale. Health and Quality of Life Outcomes 2006;4:23. www.hqlo.com/content/4/1/23.

[22] Morley JE, Charlton E, Patrick P, Kaiser FE, Cadeau P, McCready D, Perry III HM. Validation of a screening questionnaire for androgen deficiency in aging males. *Metabolism* 2000; 49:1239-42.

[23] Smith KW, Feldman HA, McKinlay JB. Construction and field validation of a self-administered screener for testosterone deficiency (hypogonadism) in ageing men. *Clinical Endocrinology* 2000; 53:703-11.

[24] Heinemann LAJ, Saad F, Heinemann K, DoMinh Thai. Can results of the AMS scale predict those of screening scales for androgen deficiency? Aging Male 2004;7:211-18.

[25] Morley JE, Perry III HM, Kevorkian RT, Patrick P. Comparison of screening questionnaires for the diagnosis of Hypogonadism. Maturitas 20005, www.sciencedirect.com 2 September 2005.

[26] Heinemann LAJ, Saad F, Heinemann K, DoMinh Thai. Can results of the AMS scale predict those of screening scales for androgen deficiency? *Aging Male* 2004;7:211-18.

[27] Itoh N, Hisasue S, Kato R, Tanaka T, Takahashi A, Masomori N Tsukamoto T (abstract). Comaparison of Morley's ADAM questionnaire and Heinemann's Aging Male Symptoms (AMS) rating scale to screen andropause symptoms in Japanese males. *Aging Male* 2004;7:49.

[28] Soh J, Ishida Y, Naito Y, Ochiai A, Naya Y, Mizutani Y, Fujuito A, Kawauchi A, Fujiwara T, Tschida H, Fukui K, Miki T. (abstract). Correlations of AMS score, depression score and hormonal levels with the manifestation of partial androgen decline in the aging male (PADAM). *Aging Male* 2004;7:83.

[29] Jankowska EJ, Szklarska A, Lopuszanska M, Medras M (abstract). Hormonal determinants of andropausal symptoms in Polish men. *Aging Male* 2004;7:21.

[30] T'Sjoen G, Feyen E, Kuyper P, Comhaire F, Kaufman JM. Self –referred patients in an aging male clinic - much more than androgen deficiency alone. *Aging Male* 2003;6:157-165.

[31] Dunbar N, Gruman C, Reisine S, Kenny A. Comparison of two health status measures and their associations with testosterone levels in older men. *Aging Male* 2001;4:1-7.

[32] Myon E, Martin N, Taieb Ch. The French Anging Males Symptoms (AMS) scale: Methodological review. Health and Quality of Life Outcomes 2005; 3:20. www.hqlo.com/content/3/1/20

[33] Yoshida NM, Kumano H, Kuboki T. Does Aging Males' Symptoms scale assess major depressive disorder? A pilot study.

[34] Nieschlag E, Swerdloff R, Behre HM, Gooren LJ, Kaufman JM, Legros JJ, Lunenfeld B, Morley JE, Schulman C, Wang C, Weidner W, Wu FCW. Investigation, treatment and monitoring of late-onset hypogonadism in males. ISA, ISSAM, and EAU recommendations. European Urology 2005;48:1-4.