Health and Quality of Life Outcomes

BioMed Central

Review

Open Access

The Aging Males' Symptoms (AMS) scale: review of its methodological characteristics

Isolde Daig[1], Lothar AJ Heinemann*[2], Sehyun Kim[3], Somboon Leungwattanakij[4], Xavier Badia[5], Eric Myon[6], Claudia Moore[7], Farid Saad[8], Peter Potthoff[9] and Do Minh Thai[2]

Address: [1]Institute Medical Psychology, University Centre for Human & Health Research, Berlin, Germany, [2]Center for Epidemiology & Health Research Berlin, Invalidenstr. 115, 10115 Berlin, Germany, [3]Department of Preventive Medicine, College of Medicine, Pochon CHA University, Korea, [4]Section of Male Sexual Dysfunction, Division Urology, Ramathibodi Hospital, Bangkok, Thailand, [5]Health Outcomes Research Europe, Barcelona, Spain, [6]PharmacoEconomics Programmes, Pierre Fabre S.A., Boulogne-Billancourt, France, [7]Medical Affairs Andrology, Jenapharm Jena, Germany, [8]Fertility Control/Hormone Therapy, Corporate Strategic Marketing Male Health Care, Schering AG, Berlin, Germany and [9]NFO Health Europe, Munich, Germany

Email: Isolde Daig - isolde.daig@charite.de; Lothar AJ Heinemann* - Heinemann@zeg-berlin.de; Sehyun Kim - skim@cha.ac.kr; Somboon Leungwattanakij - ptuaja@hotmail.com; Xavier Badia - xbadia@hor-europe.com; Eric Myon - eric.myon@pierre-fabre.com; Claudia Moore - Claudia.Moore@jenapharm.de; Farid Saad - farid.saad@schering.de; Peter Potthoff - peter.potthoff@nfoeurope.com; Do Minh Thai - dominhthai@zeg-berlin.de

* Corresponding author

Published: 15 December 2003

Health and Quality of Life Outcomes 2003, **1**:77

This article is available from: http://www.hqlo.com/content/1/1/77

Received: 09 October 2003
Accepted: 15 December 2003

Abstract

Background: The current paper reviews data from different sources to get a closer impression on the psychometric and other methodological characteristics of the Aging Males' Symptoms (AMS) scale gathered recently. The scale was designed and standardized as self-administered scale to (a) to assess symptoms of aging (independent from those which are disease-related) between groups of males under different conditions, (b) to evaluate the severity of symptoms over time, and (c) to measure changes pre- and post androgen replacement therapy. The scale is in widespread use (14 languages).

Method: Original data from different studies in many countries were centrally analysed to evaluate reliability and validity of the AMS.

Results: *Reliability* measures (consistency and test-retest stability) were found to be good across countries, although the sample size was sometimes small.

Validity: The internal structure of the AMS in healthy and androgen deficient males, and across countries was sufficiently similar to conclude that the scale really measures the same phenomenon. The sub-scores and total score correlations were high (0.8–0.9) but lower among the sub-scales (0.5–0.7). This however suggests that the subscales are not fully independent.

The comparison with other scales for aging males or screening instruments for androgen deficiency showed sufficiently good correlations, illustrating a good criterion-oriented validity. The same is true for the comparison with the generic quality-of-life scale SF36 where also high correlation coefficients have been shown.

Methodological analyses of a treatment study of symptomatic males with testosterone demonstrated the ability of the AMS scale to measure treatment effect, irrespective of the severity of complaints before therapy. It was also shown that the AMS result can predict the independently generated (physician's) opinion about the individual treatment effect.

Conclusion: The currently available methodological evidence points towards a high quality of the AMS scale to measure and to compare HRQoL of aging males over time or before/after treatment, it suggests a high reliability and high validity as far as the process of construct validation could be pressed ahead yet. But certainly more data will become available, particularly from ongoing clinical studies.

Background

The interest of clinical research in aging males increased in recent years and thereby the interest to measure health-related quality of life and symptoms. Aging men, however, are not as much aware of the fact as women that they, too, undergo some kind of "transition" at the time when women experience their menopausal transition – but males overlook their symptoms usually. For example, the various types of sweating do not show differences in frequency between males and females in the course of aging [1]. This applies more or less also for other symptoms – at least in frequency [2].

The Aging Males' Symptoms (AMS) scale is a health-related quality of life scale (HRQoL) and was originally developed in Germany in 1999 [3]. The scale was designed as self-administered scale to (a) to assess symptoms of aging (independent from those which are disease-related) between groups of males under different conditions, (b) to evaluate the severity of symptoms over time, and (c) to measure changes pre- and post androgen therapy [3]. 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 [4,5].

The AMS scale was internationally well accepted. Many translations were done following international methodological recommendations (English, Dutch, French, Spanish, Portuguese, Italian, Swedish, Korean,, Thai, and Japanese language), some are available only as simple forward translation (Finnish, Flemish, Russian). These versions are available in a published form [4,6,7], on their way to be published is a translations into Indonesian language.

The evaluation of the AMS scale is simple and have been published recently again [6] and there are norm values to compare with [3,5]. Norm values of the standardized scores (total score and three domain scores) however were only published for Germany until now. In so far it is important to compare the internal structure of the scale among countries to get an impression whether one should worry about compatibility if one intends to pool results of clinical studies across countries.

Like in other QoL scales, it is a challenge to satisfy the demands of a clinical utility and outcomes sensitivity, and this in addition to the conventional psychometric requirements of test reliability and validity.

The aim of this paper is to present psychometric and clinical data to discuss the methodologically relevant characteristics of the AMS scale.

Methods

The development of the scale, instrument characteristics (item selection, scaling), and norms and standardized scores have been published elsewhere [3,5]. This applies also for a few data that have been published on test-retest stability and criterion-dependent validity [5].

During the last two years a number of smaller and larger investigations were made from different groups to further check methodological features of the scale. In other words, this paper deals with a secondary analysis of data gathered elsewhere and partly for other purposes than this paper, however, we cannot see reason for a particular bias. The original data were collected by the authors to set up a comprehensive database suitable for the purposes of this paper.

The majority are descriptive studies of community samples of aging males (all covering the age range 40 to about 70 years), some were planned as test-retest investigation with a time interval of about two weeks, but there is only one intervention study (before and after testosterone treatment) completed yet. The latter study is not published yet in detail but some methodologically relevant results will be presented here.

A short description of the study groups follows:

Descriptive studies of aging males in Germany(aged 40–70 years)

- 1996: 116 males with banal/minor health issues such as common cold were recruited via GP's. This was the initial study group for the development of the AMS scale

- 1997: A representative community sample of 958 males was drawn for an opinion poll (EMNID) the AMS scale was also applied.

- 2003: 4633 aging males of a representative community sample participated in another opinion research (KDA) that included the AMS scale.

Test-Retest (pilot) studies with community samples (aged 40–70 years): time interval between tests about two weeks in all studies with a variation of a few days. The reason for this time interval is: acceptable balance between the possibility that the respondent could still remember how he answered the previous questionnaire and the other possibility that the complaints changed in between.

- United Kingdom: A group of 96 healthy aging males sampled from the community (2000)

- France: Convenience community sample; 21 males for an orienting pilot study on test-retest stability (2002). Convenience sample means that a group of repassing or known males was investigated without efforts to formally draw a random sample from a defined community.

- Spain: Convenience community sample; 33 males for an orienting pilot study on test-retest stability (2002)

- Portugal: Convenience community sample; 20 males for an orienting pilot study on test-retest stability (2002)

- Italy: Convenience community sample; 20 males for an orienting pilot study on test-retest stability (2002)

- Sweden: Convenience community sample; 24 males for an orienting pilot study on test-retest stability (2002)

- Thailand: Convenience community sample; 20 males for an orienting pilot study on test-retest stability (2002)

- Korea: Convenience community sample; 25 males for an orienting pilot study on test-retest stability (2002)

Intervention study (age range 17–83 years)

- Germany: 943 urology patients with diagnosed androgen deficiency that were treated with testosterone depot over 12 weeks; AMS measurement before and at the end of treatment.

Using this database from different sources, we were able to scrutinize many methodological characteristics of the AMS scale to review most fundamental psychometric characteristics as well validity in a clinical study setting.

Results and Discussion

Reliability

For all scientific measurements it is required to give evidence of replicability (consistency) and test-retest reliability. In contrast to systematic and random variation, reliability gives an estimate of method-related measurement error which should be low not to cover systematic changes – due to treatment for example.

Table 1 shows the internal consistency measured with Cronbach's Alpha. The consistency coefficients fell between 0.7 and 0.9 across countries, time periods, and total score as well the three subscales. This is indicative for a very acceptable consistency of the AMS scale in our opinion. Moreover, there is no evidence that the scale works different in so many different countries in Europe and Asia from this angle.

Table 1: Internal consistency coefficients (α) for the AMS scale across countries: total score and scores for the psychological, somatic, and sexual sub-scale. Community samples.

	EUROPE						ASIA		
	Overall	Germany			UK	Rest Europe*	Overall	Thailand	Korea
		1996	1997	2003	2000	2002		2002	2002
N	5809	116	958	4633	96	118	45	20	25
Total score	0.90	0.88	0.86	0.90	0.90	0.88	0.92	0.79	0.94
Psychological score	0.83	0.82	0.76	0.83	0.84	0.76	0.87	0.58	0.89
Somatic score	0.82	0.84	0.79	0.82	0.85	0.77	0.83	0.70	0.87
Sexual score	0.81	0.80	0.72	0.81	0.80	0.80	0.77	0.62	0.76

* France, Spain,, Portugal, Italy, Sweden (about n = 20 each)

Table 2: Test-retest coefficients (Pearson's correlation coefficient r) for the AMS scale across countries: total score and scores for the psychological, somatic, and sexual sub-scale. Community samples.

	EUROPE								ASIA		
	Overall	Germany	UK	France	Spain	Portugal	Sweden	Italy	Overall	Korea	Thailand
N	316	102	96	21	33	20	24	20	45	25	20
Total score	0.86	0.86	0.91	0.85	0.80	0.74	0.88	0.79	0.91	0.90	0.87
Psychological score	0.77	0.81	0.81	0.82	0.50	0.67	0.69	0.54	0.87	0.85	0.78
Somatic score	0.83	0.85	0.91	0.71	0.89	0.45	0.94	0.76	0.87	0.87	0.81
Sexual score	0.82	0.80	0.81	0.88	0.74	0.67	0.84	0.98	0.80	0.77	0.68

* France, Spain,, Portugal, Italy, Sweden (about n = 20 each)

The test-retest correlation coefficients (Pearson's correlation) support the suggestion of a good temporal stability of the total scale and its three sub-scales (table 2), although most of the assessments across countries are based on very small numbers. The intention of these pilot studies was to get a preliminary idea about retest stability. Larger sample sizes are required to permit final conclusions for individual countries / languages.

The test-retest coefficients of the total score range between 0.8 and 0.9 across Europe and Asia. When it comes to the subscales with much fewer items, the variation increased and some of the coefficients went down to 0.5 (psychological scale: Spain, Italy; somatic scale: Portugal). Altogether, the test-retest stability over a time period of two weeks supports the notion of an acceptable reliability of the total scale and their three sub-scales.

Although there is an impressive set of information currently available concerning the reliability of the AMS scale, there are also limitations: Small sample sizes prevent a final conclusion regarding test-retest reliability in some of the languages the scale has been translated in. In addition, further studies on consistency of the scale would be welcome in some of the languages.

Validity

Similar to reliability that assesses the consistency of measurement, the validity estimates if a QoL scale measures what it intends to measure. 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.

Internal structure of the AMS across countries

The first step of validation is usually to demonstrate multivariately the internal structure ("dimensions") of a given scale through factor analysis.

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 (table 3). Since then, two large community samples of aging males (also in Germany) were performed and analysed for this paper, i.e. in 1997 and 2003. The loadings of the 17 items on the 3 factors are astonishingly similar with those of the initial factor analysis. This suggests that the scale measures constantly the same phenomenon over time in Germany. However, there are two items that are not particularly helpful in forming the sexual factor: The statement that the beard growth decreased and the feeling to have passed the peak of life. The latter had already a lower loading at the very first analysis and the previous was intentionally included into the scale on clinical considerations although there was almost no loading. This indicates that item 12 ("Feeling that you have passed your peak")and 14 ("Decrease of beard growth") could be eliminated if a new standardization is planned, currently however these items have to be kept – otherwise standardization and norms will not be applicable anymore.

An interesting piece of evidence for a good test characteristics is the internal structure of the scale in *males with androgenic dysfunctions* before treatment. The data came from the baseline examination of the testosterone intervention study by urologists in Germany (see methods).

The internal structure of the scale seems to be identical with the "normal population", but the three factors explain here more of the total variance (65% as compared with around 55% in the normal population). This is

Table 3: Germany: Internal structure of the AMS over time. Comparison of community samples with a sample of males with dysfunctions requiring treatment. Factor loadings only depicted if 0.5 or more in the sub-scales: psychological (psych), somatic (somat), and sexual sub-scale (sex).

		COMMUNITY SAMPLES									SAMPLE WITH DYSFUNCTION		
	Year	1996			1997			2003			2002		
	N	116			958			4633			943		
	Item Nr.	Psych	Somat	Sex	Psych	Somat	Sex	Psych	Somat	Sex	Psych	Somat	Sex
Burned out	13	0.8			0.6			0.7			0.7		
Depressive, more	11	0.8			0.6			0.8			0.8		
Irritability, increased	6	0.7			0.5			0.6			0.8		
Anxious, more	8	0.7			0.7			0.7			0.8		
Nervousness, more	7	0.6			0.6			0.7			0.8		
Joint complaints, more	2		0.8			0.7			0.8			0.8	
Sweating, increased	3		0.7			0.5			0.5			0.7	
Sleep, need for more	5		0.6			0.6			0.6			0.6	
Well-being, impaired	1		0.6			0.7			0.7			0.6	
Sleep disturbances, more	4		0.6			0.6			0.5			0.6	
Muscular weakness	10		0.5			0.5			0.6			0.5	
Physical exhaustion	9		0.5			0.7			0.6			0.5	
Sexual potency, impaired	15			0.9			0.8			0.9			0.9
Morning erections, less	16			0.9			0.8			0.9			0.9
Libido, disturbed	17			0.8			0.8			0.9			0.8
Passed peak	12			0.6				0.5			0.5		
Decrease of beard growth	14			0.2		0.6							

indicative for a particular sensitivity to measure androgenic dysfunction.

The comparison of the internal structure between the large German sample and the other countries is affected by the small sample size in the latter. This may have introduced much random error. Anyway, the structure found in Germany (4 samples lumped together), UK, "Rest of Europe" (France, Spain, Portugal, Italy, Sweden pooled), and the two small samples from Asia (Thailand and Korea) is – in average – in the same ballpark (table 4). However, beside similarities there are also differences: Three out of the five items originally belonging to the psychological factor seem to be associated more with the somatic factor in the Asian samples. There seems to be also some evidence that the factors are not really independent in the two Asian samples – whether this is real or not cannot be decided with such small samples. In UK, the somatic factor seems to have a slightly different pattern: 3 out of 7 items seem to be more associated with the psychological factor than the somatic. The item "increased sweating" may have particular difficulties. The question is, whether this is a problem of small numbers and random error or reflection of remaining cultural differences. Research with larger samples is suggested, on this occasion also local norms of standardized score levels should be determined which are currently available only for Germany.

The possibly slight differences in the internal structure of the AMS scale across country groups seen in table 4 suggest further research but may 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. However, it cannot be excluded 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 provide a confirmation of the internal structure of the AMS scale across countries, some less some more convincing.

Sub-scores and total score correlations

The relations among the sub-scales and the aggregate total scale are patterns that are important in the methodological assessment of a scale. In an ideal world, the correlations between subscales (supposed to be independent) would be closer to 0 than the correlations with the con-

Table 4: International comparison of community samples: Internal structure of the AMS. Factor loadings only depicted if 0.5 or more in the sub-scales: psychological (psych), somatic (somat), and sexual sub-scale (sex). Germany: aggregate of samples 1996, 1997, and 2003.

		Germany			United Kingdom			Rest Europe			Asia		
	N	5809			96			118			45		
	Item Nr.	Psych	Somat	Sex	Psych	Somat	Sex	Psych	Somat	Sex	Psych	Somat	Sex
Burned out	13	0.7						0.5			0.8		
Depressive, more	11	0.8			0.7			0.7			0.6	*0.6*	
Irritability, increased	6	0.6			0.8			0.7			0.7		
Anxious, more	8	0.7			0.8			0.7				*0.6*	
Nervousness, more	7	0.7			0.8			0.7				*0.5*	
Joint complaints, more	2		0.8			0.5			0.5			0.8	
Sweating, increased	3		0.5		*0.7*			*0.6*			*0.8*		
Sleep, need for more	5		0.6			0.7			0.8			0.6	
Well-being, impaired	1		0.7			0.6			0.5			0.6	
Sleep disturbances, more	4		0.5		*0.8*				0.7			0.8	
Muscular weakness	10		0.6			0.7			0.5		*0.6*	0.5	
Physical exhaustion	9		0.6		*0.7*				0.6			0.8	
Sexual potency, impaired	15			0.9			0.9			0.9			0.9
Morning erections, less	16			0.9			0.8			0.8			0.9
Libido, disturbed	17			0.9			0.9			0.8			0.8
Passed peak	12	*0.5*				*0.6*		*0.5*			*0.6*		
Decrease of beard growth	14					*0.6*			*0.5*			*0.8*	

Table 5: Domain score – total score correlations of the AMS scale across countries. Community samples.

	DOMAINS		
	Psychological score	Somatic score	Sexual score
Germany (n = 5809)			
Total score	0.8	0.9	0.8
Psychological score	--	0.7	0.5
Somatic score	--	--	0.5
UK (n = 96)			
Total score	0.9	0.9	0.8
Psychological score	--	0.7	0.5
Somatic score	--	--	0.6
Other Europe* (n = 118)			
Total score	0.8	0.9	0.8
Psychological score	--	0.6	0.5
Somatic score	--	--	0.6
Asia (n = 45)			
Total score	0.9	0.9	0.8
Psychological score	--	0.8	0.7
Somatic score	--	--	0.6

* France, Spain,, Portugal, Italy, Sweden (about n = 20 each)

struct of the aggregate total score to which all sub-scales should significantly contribute. But that is theory; table 5 shows only a somewhat lower correlation among sub-scales (0.5–0.7) as compared with correlation of sub-scales with the total score (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

Table 6: Domain score – total score correlations of the AMS scale for a sample of community individuals and a sample of males with dysfunctions requiring treatment. Examples from Germany.

	DOMAINS		
	Psychological score	Somatic score	Sexual score
Germany (n = 4633)		***Community sample***	
Total score	0.8	0.9	0.8
Psychological score	--	0.7	0.4
Somatic score	--	--	0.5
Germany (n = 943)		**Dysfunctional sample**	
Total score	0.9	0.9	0.8
Psychological score	--	0.7	0.6
Somatic score	--	--	0.6

with orthogonal factors. The situation was similar in Germany, UK, Rest of Europe, and Asia. It is important to realize how similar these correlation coefficients are among countries/aggregates. This is suggestive of pretty similar features of the AMS scale across the countries of this review. It is even more important to underline that this pattern is true both for the situation in the normal population (community sample) and in patients with apparent androgenic dysfunctions prior to treatment (table 6).

Criterion-oriented validity: correlation with other scales

In fact, the comparison with other scales of similar purpose is important. It is known from other quality of life scales that comparisons with scales with similar purposes are much more important than comparisons with so-called objective parameters such as exercise tests, physiological or chemical parameters – in our case with hormones.

Health related quality of life should be validated against quality of life measured with other generic QoL scales (e.g., SF36), validated against specific instruments to measure symptoms in aging males (e.g. Finnish Turku scales), or with scales to screen for androgen deficiency (e.g., ADAM, Smith's scale).

Finnish scales

Regarding validity of the AMS scale, a Finnish research group observed a strong and statistically significant correlation with their Turku 14-items scale for aging males (r = 0.8; n = 95), and similar promising results when comparing with their own "3-Item-Scale" [5]. The two scales can be regarded as measuring the same phenomena and this speaks in favour of good test characteristics of the AMS scale.

Androgen deficiency screening scales

In an investigation in 2003 we compared the AMS scale with two screening instruments: The ADAM scale of Morley at al [8] and the Screener of Smith et al [9]. These two scales were developed to screen males for a possible androgen deficiency, i.e. to select persons for a lab test of testosterone, for example.

A convenience sample of 150 German males aged 40–70 years was drawn from a population panel (not patients). The three scales were administered for completion from all participants. This study will be published in more detail elsewhere.

To describe the ability of the AMS total score to predict the results of each of the two other scales, a simple 2 × 2 table was constructed: AMS negative (<27) or positive (27 and more scoring points; the cut-off was arbitrarily chosen), ADAM negative or positive (see [8]), and Smith's scale "negative" (0–4 points; see [9]) or "positive" (two groups of increasing "suspicion of hypogonadism" together: 5–10+ points). The associations between the AMS categories and the (ADAM) or (SMITH's) categories are significant. The Cramer's V coefficients are as follows: AMS / ADAM 0.33, AMS / Smith's screener 0.31.

Using the above mentioned cut-off points of the scales, 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%, and sensitivity = 29%. Thus, the AMS predicts well a positive screening result of the ADAM scale, but less good negative screening results of the ADAM scale. Similarly concerning 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% (sens.). The values for the prediction of ADAM results regarding the Smith' screener results are somewhat lower: 57% (pPV), 50% (nPV), 60% (sens.), and 46% (spec.), respectively. Just for comparison, Morley [8] reported for a positive ADAM result and a low bioavailable testosterone level (<70 ng/ml) a sensitivity of 88% and a specificity of 60%.

These results showed again, that the AMS scale has a good criterion-oriented validity, although the results can be discussed with a bit of reservation because both scales (ADAM, SMITH's) have not be compared in the original language. Just a simple German translation (no full cultural adaptation was done) was the basis for this investigation. Moreover, the AMS data could have been combined with age and body mass index to match the approach of the Smith's screener better. This and other details will be published elsewhere.

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 subscales of the SF36 [5]. 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). The correlation is inverse due to the fact that the sum-scores of the AMS increase with numbers (intensity) of symptoms/complaints and 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 [5].

Discriminative validity: detection of treatment effects

In this section, we summarize what information became recently available regarding predictive or criterion-oriented validity, i.e. the ability of the AMS scale to detect or predict therapeutic effects or subjective judgments of this effect by physicians. To this end, many clinicians use the term "validity" and mean high utility for clinical work or research. In so far it is important to address also this issue.

In the meantime the first androgen treatment study with the AMS scale as outcome measure have been completed (other clinical studies are under way elsewhere). It was a simple open treatment study with testosterone depot performed by the company Jenapharm in Germany. Specific details will be published elsewhere. It is by no means the intention to discuss the efficacy of testosterone treatment or of a specific type. The aim is to check the evidence that the AMS can detect "improvement of symptoms" and mirror the subjective judgment of the treating physicians about the effectiveness (not efficacy) of their treatment. Therefore, methodologically relevant information will be described here.

It is well established that men with androgen deficiency react with a marked improvement of the health-related quality of life (HRQoL) after testosterone treatment. It is important to demonstrate that the AMS scale is able to mirror changes of the HRQoL, i.e. that it can detect an improvement of complaints after therapeutic intervention. In the above mentioned study over 1000 men with androgen-deficiency relevant complaints were followed-up by urologists in their routine medical practice. After some diagnostics (e.g. testosterone) they were treated with testosterone over 12 weeks. The AMS scale was applied before treatment and after 3 months of treatment. Data of 711/700 patients before /after therapy were available for our analysis (50 years or older)

Figure 1 demonstrates that the increased mean AMS total score at baseline (before treatment) decreased after 12 weeks under treatment, i.e. indicating an improvement of complaints & HRQoL. This is also the case for the three domains (data not shown). The absolute improvement of symptoms during treatment was 15 scoring points of the AMS in average. This is equivalent to 32% of the baseline score. This is similar also for all three sub-scales (data not shown). In other words, the AMS scale was fully able to detect treatment effects (predictive value see further down).

To answer the question whether the sensitivity of the AMS scale is sufficient to detect even treatment-related changes in patients with only mild or moderate symptoms as compared with severe ones, the analysis was stratified by the severity at baseline (Figure 2). An improvement of complaints/QoL was seen in an increasing degree in patients with mild, moderate and severe symptoms at baseline. The relative improvement increases with the degree of severity of symptoms at baseline, what fits the general expectation. And there was still a positive treatment effect in men with moderate or even mild symptoms.

Figure 3 shows the capacity of the AMS scale to detect therapeutic efficiency from another angle: the comparison with norm values of the population. One can see that the level of complaints in elderly patients before therapy is very much shifted toward higher degree of severity (higher AMS total score). After 12 weeks of testosterone treatment the frequency distribution of patients with a certain severity of complaints became similar to the distribution in the general population of aging males. This is re-assuring and indicates that comparisons with norm values could be

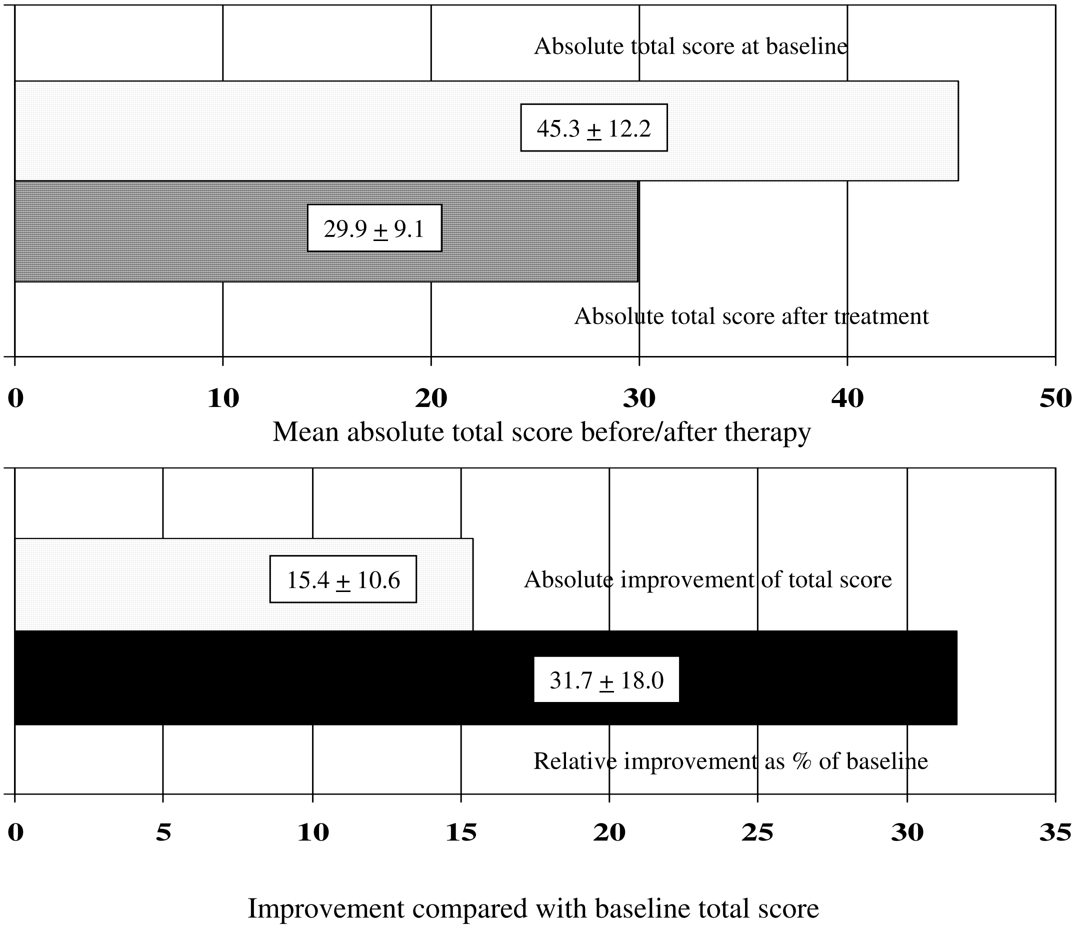

Figure 1
Comparison of the mean AMS total scores before and after testosterone therapy in patients 50 years and older: mean (SD) absolute scores (upper graph) and mean (SD) improvement (lower graph) compared with the baseline score: absolute and (upper column) as percent (%) of the AMS score before therapy (lower column).

helpful for interpreting results of intervention studies. It is another way to look at therapeutic efficiency with the assistance of the AMS scale.

The AMS scale was also able to predict the independent judgement of the urologists regarding the therapeutic effect. The treating urologist assessed individually the effectiveness of the hormone treatment in the above mentioned intervention study – without knowledge of the results of the self-administered AMS questionnaire which was only later analysed. Grouping the urologist's expert opinion regarding treatment efficiency into two categories: *effective* (very effective and effective) and *not effective* (little, no or negative effect) this alternative variable can be used for a comparison with the AMS result (total score only). We used two definitions for " treatment efficiency" based on the change of the AMS total score between baseline and treatment (as percent of the baseline total score): no effectiveness = percent change (improvement of complaints/quality of life up to +5% vs. up to +20% compared with baseline. For the first cut-off off point (5% improvement of total score) the positive predictive value was 89%, the negative predictive value 59%, sensitivity (correct prediction of a positive judgment of the physician concerning therapeutic effectiveness) 96%, and specificity (correct prediction of a negative assessment by the physician) 30%. In other words, the change of the AMS score fits well with a positive judgement of the physician concerning therapy efficiency, however predicts not as good a negative therapy assessment of the physician. The respective data for the cut-off point "20% improvement of total score" were 92%, 35%, 81%, and 61%. With higher cut-

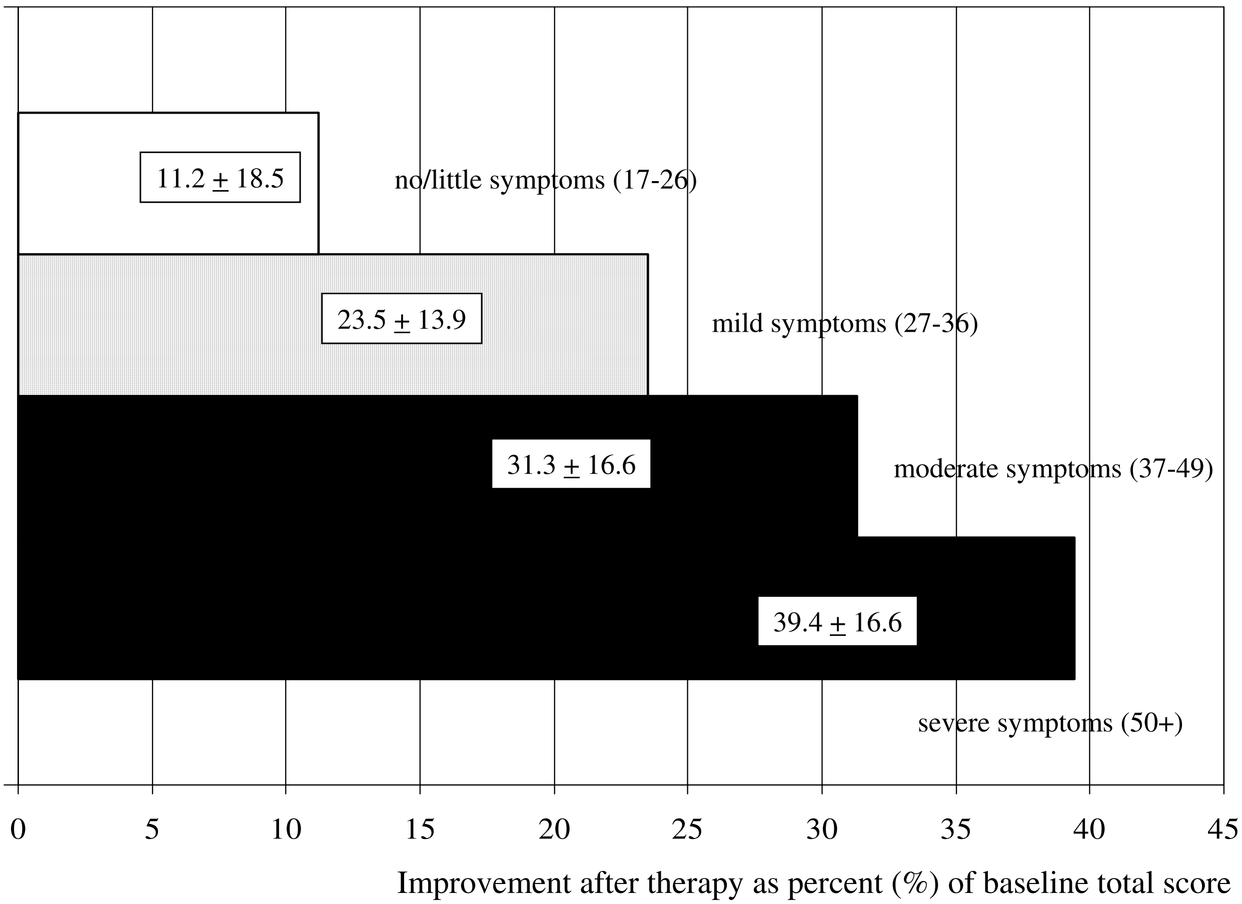

Figure 2
Average improvement of symptoms after therapy in four categories of severity at baseline. Improvement of the total score after therapy – expressed as percent(%) of the value before therapy (baseline). Means and standard deviation of the relative improvement are depicted. The severer the symptoms at entry the higher the improvement of complaints/quality of life.

points the positive predictive value remains stable at a level of about 92%. This means, using 5% relative change of the total AMS score as criterion, it would result in a high positive predictive value but a lower negative predictive value, very high sensitivity but low specificity. Still, the results of the AMS scale and the judgment of the physician regarding androgen treatment are in good agreement according to our opinion.

In any case, it was not the aim of this exercise to suggest '*5% relative improvement*' as a criterion for a new "diagnostic test". The aim was only to demonstrate that the AMS score can very well predict the clinical efficiency of an androgen therapy in aging males with apparent androgen deficiency. More details will be published elsewhere, but it was our aim to present the methodological essentials in this review paper. It should be just an example to illustrate the capacity of the test and certainly not a suggestion to use the scale as a substitute for the rather complex medical judgement of the treating physician in a clinical setting. However, it might be useful to apply the standardized "objective" scale of the test in clinical studies instead of a subjective judgement of a physician.

Conclusions

The AMS scale is a standardized HRQoL scale with good psychometric characteristics. The use in many countries offered the possibility to compare the test characteristics across countries. Reliability measures (consistency and test-retest stability) were found to be good in all countries where data were obtained – however, some samples were very small and therefore lumped together.

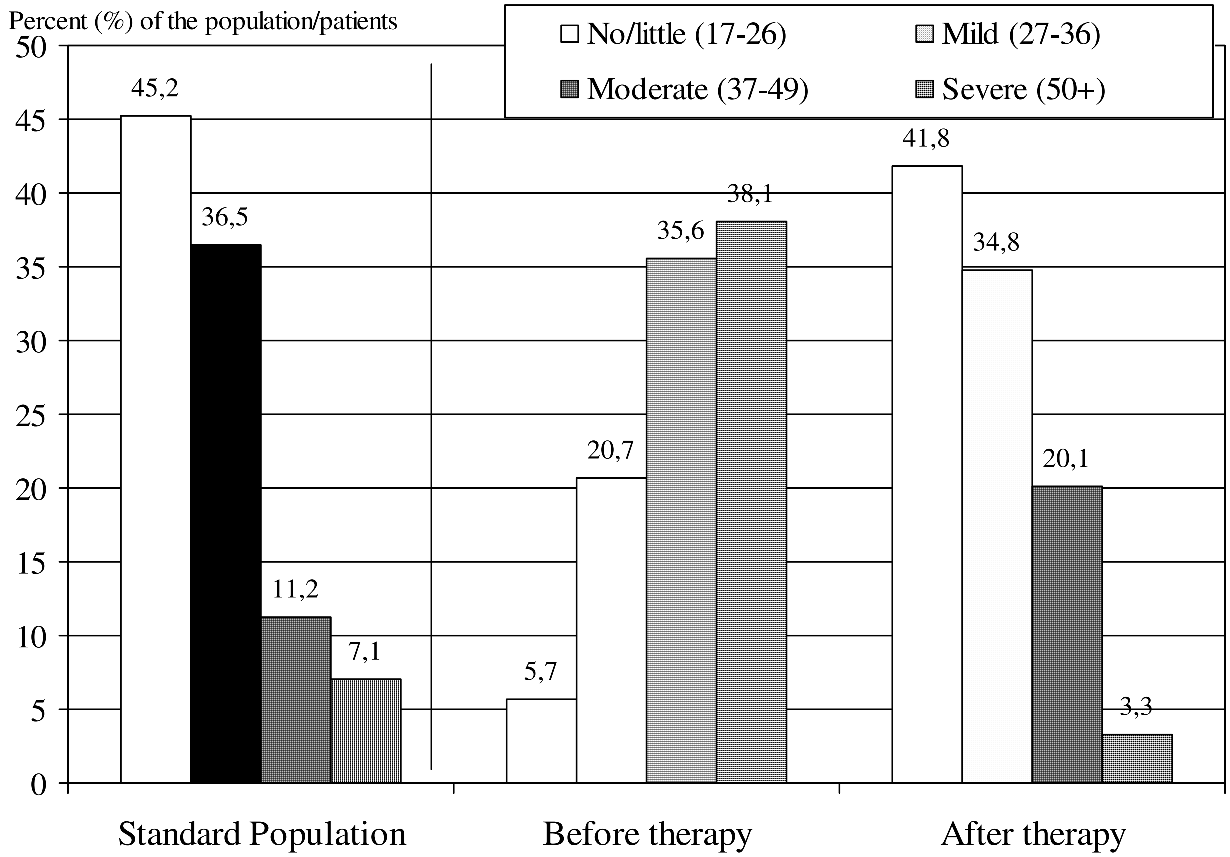

Figure 3
Percentage (%) of patients in the four categories of severity of complaints according to AMS results at two points in time: before therapy with testosterone and 12 weeks later after testosterone treatment. The disturbed frequency distribution of severity before therapy (compared to the standard-left part) went back to the distribution in the "normal" standard population after therapy (right part). The population data came from the standardization of the AMS test [3,5]

The validity was measured in its various forms: The internal structure in healthy as well androgen deficient males across countries was sufficiently similar to conclude that the scale really measures the same phenomenon. The subsores and total score correlations showed high coefficients with the total score and less among the sub-scales. This however indicates that the subscales are not fully independent in practice.

The comparison with other scales for aging males or screeners for androgen deficiency showed high correlation coefficients, i.e. illustrating a good criterion-oriented validity. The same is true for the comparison with the generic QoL scale SF36 where also high correlation coefficients have been shown.

Methodological analyses of a testosterone treatment study of symptomatic males demonstrated the ability of the AMS scale to detect a treatment effect, irrespective of the severity of complaints before therapy. It was also shown that the AMS result can predict the independent (physician's) opinion about the individual treatment effect.

Thus, the currently available methodological evidence points towards a high quality of the scale to measure and to compare HRQoL of aging males over time or intervention. It suggests a high reliability and high validity as far as the process of construct validation could be pressed ahead yet. But certainly more data will become available, particularly from ongoing clinical studies. The latter is an particular aim in the process of construct validation. It would be also interesting to invest more into the ability of the scale to measure treatment effects regarding sexual dysfunction specifically.

Competing interests

None declared.

Authors' contributions

ID: involved in writing the manuscript, responsible for computation of reliability and parts of the validity assessment. **LAJH**: developer of the AMS scale, responsible for the collection of all data, and involved in writing of the paper. **SK**: responsible for studies in Korea, provided the data, and contributed to writing of the paper. **SL**: responsible for studies in Thailand, provided the data, and contributed to writing of the paper. XB: provided the data from Spain, and contributed to writing of the paper. **EM**: provided the data from France, contributed to writing of the paper. CM: responsible for the androgen therapy study, provided the data, and contributed to writing of the paper. **FS**: involved in the co-ordination of the study, contributed to writing of the paper. **PP**: involved in the co-ordination of studies in Sweden, Portugal, and Italy, provided data from these countries, and contributed to writing of the paper. **DMT**: responsible for setting up and checking the integrated database, responsible for several analyses regarding validity, and contributed to writing of the paper.

Acknowledgements

The market research company EMNID, Bielefeld performed the fieldwork and we are grateful to having received the database for our evaluation. We thank KDA Marktforschungs GmbH, Frankfurt a. Main for performing the field work and to provide us with the database for evaluation.

References

1. Heinemann K, Saad F: **Sweating attacks – Key Symptom in Menopausal Transition only for Women?** *Eur J Urology* 2003 in press.
2. 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.
3. Heinemann LAJ, Zimmermann T, Vermeulen A, Thiel C: **A New 'Aging Male's Symptoms' (AMS) Rating Scale.** *The Aging Male* 1999, **2:**105-114.
4. Heinemann LAJ, Saad F, Thiele K, Wood-Dauphinee S: **The Aging Males' Symptoms (AMS) rating scale. Cultural and linguistic validation into English.** *The Aging Male* 2001, **3:**14-22.
5. Heinemann LAJ, Saad F, Pöllänen P: **Measurement of Quality of Life Specific for Aging Males.** In: *Hormone Replacement Therapy and Quality of Life. Parthenon Publishing Group.* Edited by: *Schneider HPG. London, New York, Washington*:63-83.
6. 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
7. Convay K, Heinemann LAJ, Giroudet C, Johannes EJ, Myon E, Taieb C, Raynaud JP: **Harmonized French version of the Aging Males' Symptoms Scale.** *Aging Male* 2003, **6:**106-109.
8. Morley JE, Charlton E, Patrick P, Kaiser FE, Cadeau P, McCready D, Perry HM III: **Validation of a screening questionnaire for androgen deficiency in aging males.** *Metabolism* 2000, **49:**1239-42.
9. 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.