Menopause: The Journal of The North American Menopause Society
Vol. 26, No. 11, pp. 000-000
DOI: 10.1097/GME.0000000000001390

Review Article

Measurement properties of patient-reported outcome measures (PROMs) for women with genitourinary syndrome of menopause: a systematic review

Michaela Gabes, MSc,[1] Helge Knüttel, PhD,[2] Petra Stute, MD,[3] and Christian J. Apfelbacher, PhD[1]

Abstract

Objective: Genitourinary syndrome of menopause affects up to 50% of postmenopausal women and has negative impacts on the women's quality of life. In this systematic review, we aimed to identify and assess the measurement properties of all existing patient-reported outcome measures (PROMs) specific for genitourinary symptoms that were developed and/or validated for measuring patient-reported outcomes in postmenopausal women.

Methods: Studies which evaluated, described, or compared measurement properties of PROMs were considered as eligible. We performed a systematic literature search in MEDLINE, EMBASE, and Web of Science. The methodological quality of each study was assessed using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias checklist. Furthermore, predefined quality criteria for good measurement properties were applied and the quality of the evidence was graded.

Results: Nine articles reporting on four PROMs were included. Two instruments, the Vulvovaginal Symptoms Questionnaire and the Day-to-Day Impact of Vaginal Aging Questionnaire, can be further recommended for use. Both showed moderate to high quality of evidence for sufficient structural validity, internal consistency, and construct validity. The two other instruments, urogenital atrophy quality of life (UGAQoL) and the Urogenital Symptom Scale, cannot be recommended for use, whereby the UGAQoL still has the opportunity to be recommended if the authors gave access to the instrument and further validation studies were conducted.

Conclusions: Both Vulvovaginal Symptoms Questionnaire and Day-to-Day Impact of Vaginal Aging Questionnaire can be recommended for use and results obtained with these two instruments can be seen as trustworthy. Future validation studies should focus on those two instruments.

Key Words: Genitourinary syndrome of menopause – Measurement properties – Patient-reported outcome measures – Reliability – Responsiveness – Validity.

Genitourinary syndrome of menopause (GSM) is a recently agreed term for all vulvovaginal and urological symptoms and signs associated with a decrease in estrogen and other sex steroids in peri- and postmenopausal women.[1] At a terminology consensus conference in May 2013, two societies, the Board of Directors of the International Society for the Study of Women's Sexual Health and the Board of Trustees of The North American Menopause Society, concluded that previously used terms such as vulvovaginal atrophy or atrophic vaginitis are inadequate. GSM is a more comprehensive and neutral term that is not limited to genital symptoms of dryness, irritation, burning, and itching of vulva or vagina. It includes urinary problems of frequency and urgency, and recurrent urinary tract infections which can be associated with menopause and systematic aging. Urinary symptoms as part the postmenopausal period were often overlooked and the inclusion in GSM should counteract this fact.[2] Furthermore, many women report sexual symptoms as well, that is, decreased lubrication, arousal and desire, discomfort or pain with sexual activity leading to postcoital bleeding, or impaired function.[1,3] The sexual symptom complex can be seen as a consequence of the genitourinary components since a lack of lubrication is associated with vaginal dryness and decreased elasticity and results in pain

Received March 25, 2019; revised and accepted June 6, 2019.

From the [1]Medical Sociology, Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; [2]University Library, University of Regensburg, Regensburg, Germany; and [3]Department of Obstetrics and Gynecology, Inselspital Bern, Bern, Switzerland.

Funding/support: None reported.

Financial disclosure/conflicts of interest: C.J.A. has received institutional funding and consultancy fees from Dr. Wolff GmbH, Bielefeld, Germany and has received funding from Sanofi Genzyme.

Systematic review registration: PROSPERO CRD42018092384.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website (www.menopause.org).

Address correspondence to: Michaela Gabes, MSc, Department of Epidemiology and Preventive Medicine, University of Regensburg, Dr.-Gessler-Str. 17, 93051 Regensburg, Germany.
E-mail: michaela.gabes@klinik.uni-regensburg.de

and discomfort. Thus, affected women experience less desire and arousal in sexual activities. Urinary urgency and incontinence are often perceived as embarrassing and foster sexual reluctance even more.[4] We decided to cluster the components of GSM in two main components, the genital and the urinary component, and propose the sexual component as a consequence of the genital and the urinary component (Fig. 1). Up to 50% of midlife and older women worldwide suffer from menopause-related genitourinary symptoms.[5] Due to a ''vaginal taboo'' in our society, genitourinary symptoms are often under-reported. Affected women are not aware of their chronic and progressive character.[6,7] Negative impacts on quality of life (QoL) and sexual health are often reported. A higher prevalence of female sexual dysfunction and genitourinary conditions is associated with vulvovaginal symptoms and negative impacts on QoL and sexual health have been reported.[5,8-10]

To foster involvement of patients in both clinical research and routine health care, the use of patient-reported outcome measures (PROMs) has steadily increased in the past decades. These instruments reflect the patient's perspective of how they perceive their health status and whether health care interventions have been effective. PROMs are self-completed questionnaires measuring, for example, health-related QoL or health status.[11]

Several PROMs that cover diverse constructs have been developed and reported in the literature for women with GSM and vulvovaginal atrophy, respectively, for instance: the Vulvovaginal Symptoms Questionnaire (VSQ)[12] or the Day-to-Day Impact of Vaginal Aging (DIVA) Questionnaire.[13] The VSQ is a quality-of-life questionnaire with four scales: symptoms, emotions, life-impact, and sexual impact of vulvovaginal symptoms.[12] The DIVA measures the impact of vaginal symptoms on postmenopausal women's activities of daily living, emotional well-being, sexual function, and self-concept and body image.[13]

In clinical research, it is important to select measurement instruments which are reliable, valid, responsive, and feasible. The selection of instruments should be based on complete information regarding these measurement properties and the quality of the underlying research.

However, a systematic comparison of the existing PROMs for women with GSM and a judgment of the quality of these has not been undertaken.

AIM AND OBJECTIVES

Our overall aim was to critically appraise, compare, and summarize the quality of all existing PROMs in women with GSM.

More specifically, our objectives were:

1. To systematically assess the measurement properties of PROMs for women with GSM

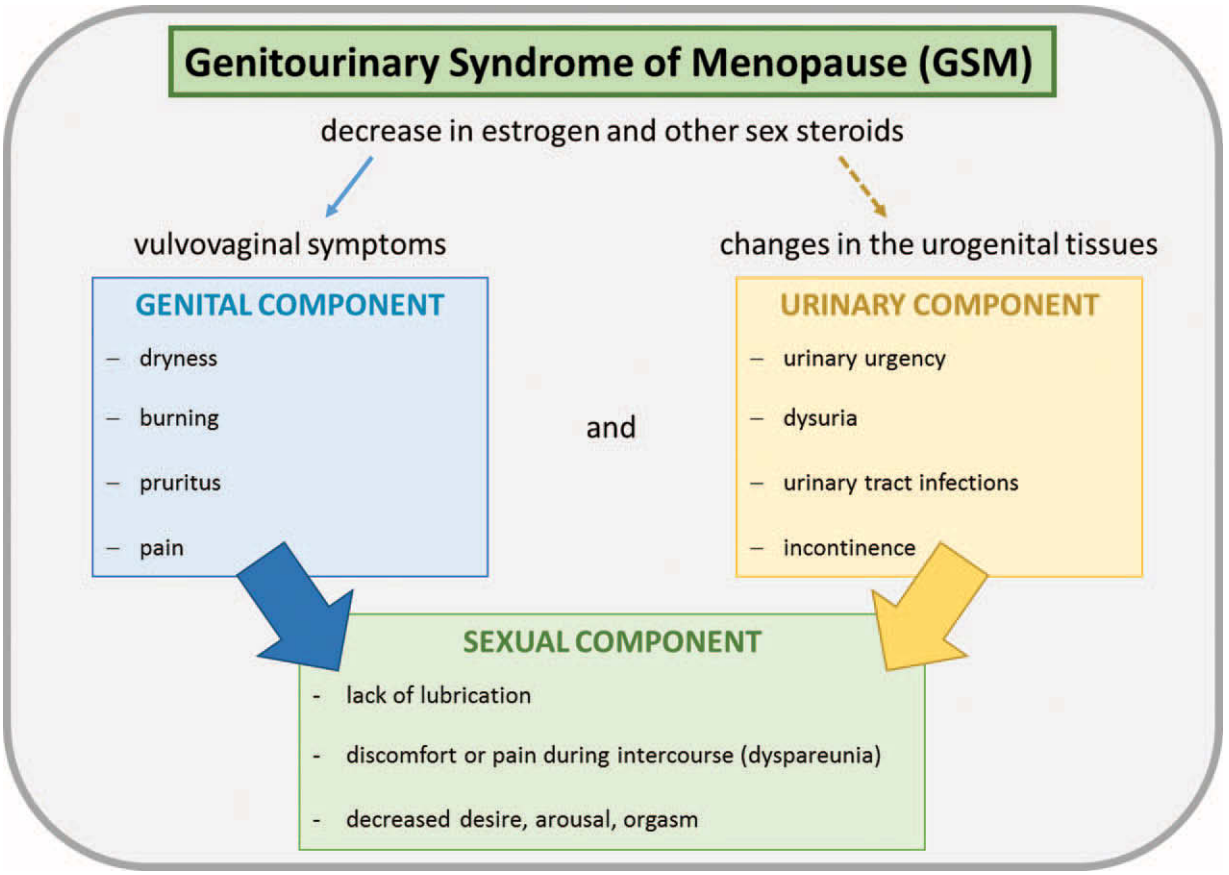

FIG. 1. The genitourinary syndrome of menopause with its two main components (genital and urinary component) and the sexual component as a consequence of them.

2. To identify PROMs for women with GSM
   a. That meet the predefined criteria to be recommended in future GSM trials
   b. That have the potential to be recommended in the future depending on the results of further validation studies
   c. That do not meet the predefined criteria to be recommended and therefore should not be used anymore
   d. We performed a systematic review of the measurement properties of all PROMs in GSM.

MATERIALS AND METHODS

Protocol and registration

The methods of this systematic review were developed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols statement.[14] The corresponding study protocol was registered in the International Prospective Register of Systematic Reviews: CRD42018092384.

Literature search

A systematic, librarian-assisted literature search was performed in the bibliographic databases MEDLINE (via Ovid, 1946-present, database code ''ppezy''), EMBASE (via Ovid, 1974-present, database code ''oemezd''), Science Citation Index Expanded (Web of Science, 1965-present), and Social Sciences Citation Index (Web of Science, 1990-present) with a last update on December 7, 2018. The search strategy was composed of the following search elements[15]:

1. Target population: Women with GSM. In order to reach maximal sensitivity a broad compilation of controlled vocabulary and free text terms was used. Terminology of GSM is variable and the full diagnosis is not always mentioned in publications for a variety of reasons.[1] Therefore, we also searched with terms for important symptoms and signs of GSM.
2. Construct of interest: All PROMs regardless of the underlying construct. For optimal sensitivity the search strategy of this search element was based on a combination of the PubMed filter ''QoL'' of Vissers and de Vries,[16] the PubMed filter '' PROMs'' of Jansma and de Vries,[17] and additional search terms from the ''PROM group construct and instrument type filter'' of Mackintosh et al.[18] PROMs is a broad term and it includes measures of QoL or health status.[11,19]
3. Measurement properties: The validated and sensitive search filter (recommended by the COSMIN group[20]) for finding studies on measurement properties developed by Terwee et al[21] was used. We employed the translation of the original PubMed filter to Ovid MEDLINE by Alberta University.[22]
4. Feasibility of PROMs: The search strategy for this element was based on the search terms for the concept ''feasibility'' of Heinl et al[23] (included in their search statement 1, additional file 2).
5. Individual PROMs: A list of known relevant PROMs.
6. Exclusion filter: This was the exclusion filter from Terwee et al[21] for a number of irrelevant publication types and for animal-only studies.

The search elements were combined as follows; to identify all articles on the measurement properties or the feasibility of PROMs in women with GSM or articles mentioning the names of GSM-specific PROMs. From these records the exclusion filter removed irrelevant publication types as well as animal-only studies: ((A AND B AND (C OR D)) OR (C AND E)) NOT F, or in words: ((population AND construct AND (measurement properties OR feasibility)) OR (individual PROMs AND measurement properties)) NOT (exclusion filter).

In addition, databases specific for PROMs were searched for records relevant to the target population: PROQOLID (https://eprovide.mapi-trust.org/about/about-proqolid), the COSMIN database of systematic reviews of outcome measurement instruments (http://www.cosmin.nl/database-of-systematic-reviews.html), the Test Archive of Leibniz Institute for Psychology Information (https://www.testarchiv.eu/), and the PubPsych search engine (https://pubpsych.zpid.de/pubpsych/). In addition to the electronic search, hand searching of the reference lists of the studies included and key articles on this topic were searched.

Search strategies for MEDLINE, EMBASE, and Web of Science were developed (see Supplemental Digital Content 1, which demonstrates the search strings and records for each database, http://links.lww.com/MENO/A442). The initially developed MEDLINE search strategy was translated to the other databases choosing appropriate syntax and index terms.

Subsequently, the bibliographic databases and the databases specifically on PROMs were searched again with the names of GSM-specific PROMs found during the initial search.

There were no restrictions regarding publication date. Only articles in English, German, French, or Italian were included. We deduplicated records in Endote X6 following the method of Bramer et al.[24] Records were then uploaded to the Covidence systematic review software (https://www.covidence.org/) for further processing, that is, title/abstract screening, full text review, and data extraction.

The literature search was re-run on December 7, 2018 to capture further relevant studies which have been published since the initial literature search.

Eligible studies

The eligibility criteria are in agreement with the COSMIN guideline for systematic reviews of PROMs.[20] The population of interest were postmenopausal women since up to 50% of postmenopausal women suffer from genitourinary symptoms. The included studies should concern PROMs specific for at least one main component (genital and urinary) of GSM, otherwise menopause-specific PROMs which are irrelevant for the syndrome would not have dropped out. The evaluation of measurement properties, the development of a PROM, or the evaluation of the interpretability of the PROMs of interest should be the principal aim of selected studies. Studies that only use the PROM to measure the outcome or in which the PROM is used for the validation of another instrument were excluded. Only full text articles were included because abstracts provide quite often very limited information on

TABLE 1. *Inclusion and exclusion criteria*

	Inclusion criteria	Exclusion criteria
Population	Postmenopausal women	All other
Study design	PROM development study, validation study	All other study designs
Outcome	All patient-reported outcomes	Non–patient-reported outcomes, such as biomarkers or physiology of the skin
Type of measurement instrument	Patient-reported outcome measures specific for at least one main component (genital and urinary) of GSM	All others
Publication type	Articles with available full-text	Abstracts

GSM, genitourinary syndrome of menopause; PROM, patient-reported outcome measure.

the design of a study. Studies that concern the development (''development paper'') and/or the evaluation of the measurement properties (''validation paper'') of PROMs were included as well (Table 1).

Study selection

Titles and abstracts found in the literature search were independently judged by two reviewers. For the remaining titles and abstracts, full-text articles were searched and judged for eligibility also by two reviewers independently. If any disagreement occurred, consensus was reached by consulting a third reviewer. If at least one reviewer considered a study as relevant based on the abstract, or in case of doubt, the full-text article needed to be screened.

Data extraction

Assessment of measurement properties and adequacy of the PROMs

Measurement properties were evaluated in the following order:

1. Evaluation of the content validity
2. Evaluation of internal structure including structural validity, internal consistency, and cross-cultural validity/measurement invariance
3. Evaluation of remaining measurement properties including reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness

All measurement properties were evaluated following three sub steps, except for the measurement property ''criterion validity'' since there exists no criterion standard for QoL.[12] For construct validity and responsiveness, we formulated hypotheses to evaluate the results against.

First, the methodological quality of the included studies was evaluated by two independent reviewers using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias checklist, which was developed exclusively for systematic reviews of PROMs.[25] The COSMIN Risk of Bias checklist consists of 10 Boxes, each for 1 measurement property (Table 2). Only those boxes for the measurement properties that are assessed in one article were completed. The COSMIN taxonomy was used to decide which measurement property has been evaluated. The standards include both preferred statistical methods based on classical test theory and item response theory or Rasch analyses.

All measurement properties of COSMIN Risk of Bias checklist are clearly defined.[26] Content validity was seen as the most important measurement property, because the items of a PROM have to be relevant, comprehensive, and comprehensible regarding the population and construct of interest.[27] If there is high quality evidence for insufficient content validity, the PROM was not further assessed and directly categorized as C, that is, the PROM should not be recommended for use. Each study was rated on a four-point rating scale (ie, ''inadequate,'' ''doubtful,'' ''adequate,'' ''very good''). The overall quality of a study was determined by the lowest rating of any standard in the box, that is, ''the worst score counts'' principle.[25] Each study on a measurement property was assessed separately and all measurement properties of each study were rated as either very good, adequate, doubtful, or inadequate.[15]

Second, we extracted relevant data on characteristics of the included PROMs and the included study populations and summarized them in evidence tables.[15] Interpretability and feasibility which are also important for a recommendation were described after the evaluation of the measurement properties. Interpretability means the degree to which qualitative meaning can be assigned to a PROM's quantitative score. Feasibility contains aspects of the ease of application (eg, costs, length, ease of administration).[15]

Furthermore, we applied criteria for good measurement properties (quality criteria). The updated criteria for good measurement properties recommended by the COSMIN group[20] are presented in Table 3. The result of each single

TABLE 2. *Boxes of the COSMIN Risk of Bias Checklist[25] (permission for publication given by Wieneke Mokkink on behalf of the COSMIN team)*

Box 1	PROM development	Content validity
Box 2	Content validity	
Box 3	Structural validity	
Box 4	Internal consistency	Internal structure
Box 5	Crosscultural validity\ measurement invariance	
Box 6	Reliability	
Box 7	Measurement error	
Box 8	Criterion validity	Remaining measurement properties
Box 9	Hypotheses testing for construct validity	
Box 10	Responsiveness	

COSMIN, COnsensus-based Standards for the selection of health measurement instruments; PROM, patient-reported outcome measure.

TABLE 3. *Updated criteria for good measurement properties[20] (permission for publication given by Wieneke Mokkink on behalf of the COSMIN team)*

Measurement property	Rating	Criteria
Structural validity	+	CTT CFA: CFI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR <0.08[a] IRT/Rasch No violation of *unidimensionality*[b]: CFI or TLI or comparable measure >0.95 OR RMSEA < 0.06 OR SMRM < 0.08 *AND* No violation of *local independence*: residual correlations among the items after controlling for the dominant factor <0.20 OR Q3's <0.37 *AND* no violation of *monotonicity*: adequate looking graphs OR item scalability > 0.30 *AND* adequate *model fit* IRT: χ^2 >0.001 Rasch: infit and outfit mean squares ≥0.5 and ≤1.5 OR Z-standardized values >−2 and <2 CTT: not all information for ''+'' reported
	?	IRT/Rasch: model fit not reported
	−	Criteria for ''+'' not met
Internal consistency	+	At least low evidence[c] for sufficient structural validity[d] AND Cronbach alpha(s) ≥0.70 for each unidimensional scale or subscale[d]
	?	Criteria for ''At least low evidence[c] for sufficient structural validity[e]'' not met
	−	At least low evidence[c] for sufficient structural validity[e] and Cronbach alpha(s) <0.70 for each unidimensional scale or subscale[d]
Reliability	+	ICC or weighted Kappa ≥0.70
	?	ICC or weighted Kappa not reported
	−	ICC or weighted Kappa <0.70
Measurement error	+	SDC or LoA < MIC[e]
	?	MIC not defined
	−	SDC or LoA > MIC
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis[f]
	?	No hypothesis defined (by the review team)
	−	The result is not in accordance with the hypothesis[f]
Crosscultural validity/measurement invariance	+	No important differences found between group factors (such as age, sex, language) in multiple group factor analysis OR no important DIF for group factors (McFadden R^2 <0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	−	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with criterion standard ≥0.70 OR AUC ≥0.70
	?	Not all information for ''+'' reported
	−	Correlation with criterion standard <0.70 OR AUC <0.70
Responsiveness	+	The result is in accordance with the hypothesis[f] OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	−	The result is not in accordance with the hypothesis[f] OR AUC <0.70

''+'' = sufficient; ''−'' = insufficient; ''?'' = indeterminate. The criteria are based on Terwee et al[28] and Prinsen et al.[29]
AUC, area under the curve; CFA, confirmatory factor analysis; CFI, comparative fit index; CTT, classical test theory; DIF, differential item functioning; ICC, intraclass correlation coefficient; IRT, item response theory; LoA, limits of agreement; MIC, minimal important change; RMSEA, root mean square error of approximation; SEM, standard error of measurement; SDC, smallest detectable change; SRMR, standardized root mean residuals; TLI, Tucker-Lewis index.
[a]To rate the quality of the summary score, the factor structure should be equal across studies.
[b]Unidimensionality refers to a factor analysis per subscale, whereas structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure.
[c]As defined by grading the evidence according to the GRADE approach.
[d]The criteria ''Cronbach's alpha < 0.95'' was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM.
[e]This evidence may come from different studies.
[f]The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses.

study was rated as either sufficient (+), insufficient (−), or indeterminate (?).[15]

Third, we aimed to summarize the evidence per measurement property per PROM, rate the overall result against criteria for good measurement properties, and grade the quality of the evidence by using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. Here we focused on the PROM and not as in the previous steps on the single studies.[15]

The third substep included several further substeps:

First, we had to decide if the results of all studies per measurement property are consistent or not.[15]

If they were consistent, they could be pooled or summarized and an overall rating as either sufficient (+), insufficient (−), or indeterminate (?) could be provided after the comparison

against the quality criteria. Finally, their quality of the evidence was graded.[15]

If the results were inconsistent, we looked for explanations for inconsistency.

1. If an explanation was found, the different results would be summarized (eg, per subgroup of consistent results) followed by an overall rating for the specific measurement property. We considered that high-quality studies provided more evidence than low-quality studies when determining the overall rating.[15]
2. If no explanation for inconsistency was found, the overall rating could be either inconsistent (±) or based on the majority of the results and therefore downgraded for inconsistency (see GRADE approach explained below).[15]

Second, we pooled the results quantitatively or summarized them qualitatively in Summary of Findings tables, each measurement property per PROM in one table.[15]

Third, each pooled or summarized result was again rated against the quality criteria (Table 3) to obtain an overall rating for the pooled or summarized result as either sufficient (+), insufficient (−), inconsistent (±), or indeterminate (?). This rating was added to the Summary of Findings Tables.[15]

Fourth, the quality of the evidence was graded to define whether the pooled or summarized result was trustworthy. It is important to consider the quality of evidence because insufficient attention to quality of evidence can lead to inappropriate recommendations that may have negative impacts for the patients. The recognition of the quality of evidence can help to prevent misguided recommendations.[30] Using the GRADE approach, we determined whether confidence in estimates of true measurement properties is given. For this systematic review we used a GRADE approach with four GRADE factors: risk of bias, inconsistency, imprecision, and indirectness. Those depend on four levels of quality evidence (ie, high, moderate, low, or very low) which are specified by the GRADE approach (Table 4). If the results did not seem trustworthy, the quality of evidence was downgraded. Each PROM was graded separately.[20] If the overall rating for a measurement property was indeterminate (?), the quality of the PROMs could not be judged and therefore the quality of evidence was not graded.[20] All results are added to the Summary of Findings tables as well.[15]

Generating recommendations for the use of PROMs for women with GSM

Each assessed instrument was assigned to a recommendation category according to its methodological quality and adequacy. Three categories of recommendation were proposed by the COSMIN group[20]:

1. PROMs with evidence for sufficient content validity (any level) AND at least low quality evidence for sufficient internal consistency
2. PROMs categorized not in A or C
3. PROMs with high-quality evidence for an insufficient measurement property

PROMs of category A can be recommended for use and results obtained with these PROMs can be seen as trustworthy. For PROMs of category B, further validation is needed; however, they still have the opportunity to be recommended for use. PROMs of category C should not be recommended for use. If only PROMs of category B are found, the PROM with the best evidence for content validity can be preliminarily recommended for use, until further evidence is given.[15]

Our aim was to identify the best (currently available) PROMs in GSM.

RESULTS

Searching the bibliographic databases yielded 9,883 records of which 6,077 remained after deduplication and moved into the screening. After the second literature search, a further 393 records were identified and screened. Eight studies were included after the full-text screening (Fig. 2).[12,13,32-37] One further relevant article was found in the reference lists of those eight studies. It contained data on the content validity of the DIVA, but did not formally meet the inclusion criteria.[39] It was therefore excluded. Nevertheless, supplementary information on content validity was extracted to assess the methodological quality of the PROM

TABLE 4. *GRADE approach for grading the quality of evidence*[20]

Quality of evidence	Lower if
High (We are very confident that the true measurement property lies close to that of the estimate of the measurement property)	Risk of bias - 1 Serious - 2 Very serious - 3 Extremely serious
Moderate (We are moderately confident that the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different)	Inconsistency - 1 Serious - 2 Very serious
Low (Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property)	Imprecision - 1 total n = 50-100 - 2 total n < 50
Very Low (We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property)	Indirectness - 1 Serious - 2 Very serious

Starting point: assumption that the evidence is of high quality.
Information on how to downgrade is described in the COSMIN user manual.[15]
Definitions were adapted from the GRADE approach.[31]
n = sample size.

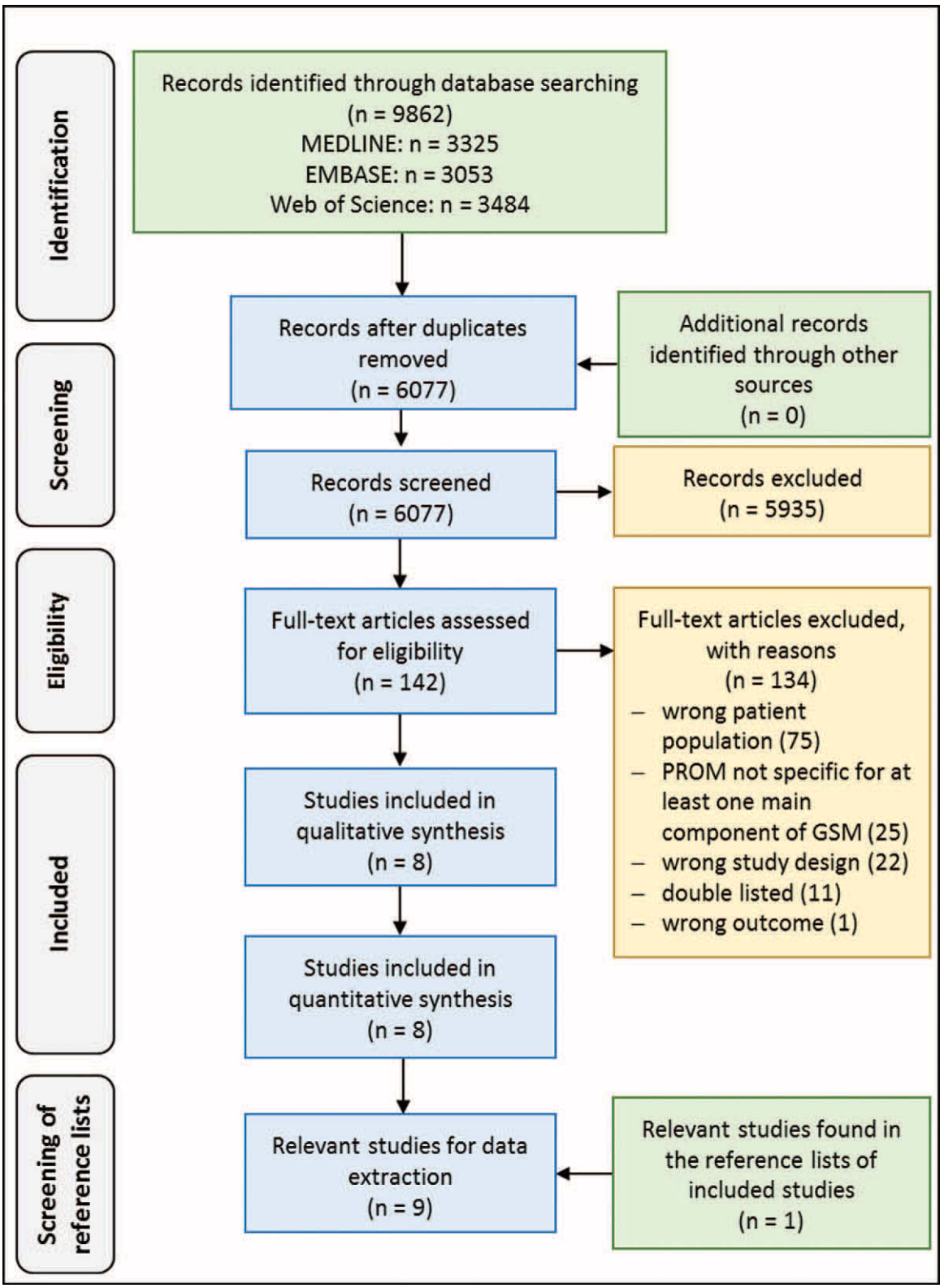

FIG. 2. Adapted Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols (PRISMA) 2009 flow diagram.[31] For more information, visit www.prisma-statement.org. GSM, genitourinary syndrome of menopause; PROM, patient-reported outcome measure.

TABLE 5. *COSMIN risk of bias overall ratings for content validity*

	VSQ	DIVA	UGAQoL	Urogenital symptom scale
Box 1. PROM development	Inadequate	Inadequate	Inadequate	Inadequate
Box 2. Content validity	Doubtful	Doubtful	Doubtful	Doubtful

COSMIN, COnsensus-based Standards for the selection of health Measurement Instrument; DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; PROM, patient-reported outcome measure; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.

development. Three included studies reported on the VSQ,[12,34,35] three on the DIVA,[13,36,38] one on the urogenital symptom scale,[33] and one on the urogenital atrophy quality of life (UGAQoL).[37] One of the three studies reporting on the DIVA was found during the second literature screening.[38]

Data extraction

Evaluation of content validity

The PROM development rating of the DIVA was inadequate since the PROM development study was performed in a sample not exactly representing the target population for which the PROM was developed. The PROM development study was performed in women with moderate to severe vulvovaginal symptoms, whereas the DIVA was developed for women with all kinds of vulvovaginal severity levels, including women with mild vulvovaginal symptoms. The ''inadequate'' PROM development rating of the UGAQoL was due to the description of the construct to be measured since the needs of the needs-based model were not further specified. The reason for the ''inadequate'' PROM development rating of the VSQ and the urogenital symptom scale is that the target population was not involved in the elicitation phase of relevant items. All content validity studies were of doubtful quality due to a lack of detailed information about different aspects of the procedure (Table 5).

The quality of evidence of the VSQ and DIVA was moderate since at least one content validity study of doubtful quality was available. A copy of the UGAQoL was not available; thus, it was not possible for the reviewers to rate relevance, comprehensiveness, and comprehensibility. Only the first page of the questionnaire with four questions was available; thus, it was partly possible to assess the response options. An overall rating of ''?'' is basically not possible since the reviewer's rating is usually always available. Because of a lack of information, we had no other choice than to rate the content validity as ''?.'' In case of an indeterminate overall rating, the quality of the PROM cannot be judged and therefore, there was no grading of the quality of the evidence. The quality of evidence of the urogenital symptom scale could not be graded because the results were rated as ''inconsistent'' (Table 6).

We could not find high-quality evidence that the content validity of any PROM was insufficient; thus, every PROM was further assessed.

Evaluation of the remaining measurement properties (structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness)

In total, the methodological quality of 24 measurement properties was rated. Fourteen measurement properties (58%) had very good, three (13%) had adequate, three (13%) had doubtful, and four (16%) had inadequate methodological quality (Table 7).

Characteristics of the included PROMs and study populations

An overview of the included PROMs is presented in Table 8. Characteristics of the included study populations are shown in Table 9. The lowest number of items in a questionnaire is 3, and the highest is 23. Two questionnaires use a dichotomous response format; the others apply a 4- or 5-point Likert scale. Sample sizes ranged from 104 to 757 patients.

Information on interpretability and feasibility

The fact that <1% of all DIVA item responses were missing[13] is an aspect of interpretability of the DIVA. In one study regarding the VSQ,[35] there was evidence that the data were normally distributed. No information on floor and ceiling effects, minimal important change or difference

TABLE 6. *Content validity rating of the included patient-reported outcome measures*

		Relevance	Comprehensiveness	Comprehensibility	Content validity rating
VSQ	Overall rating	+	+	+	Sufficient (+)
	Quality of evidence		Moderate (due to risk of bias)		
DIVA	Overall rating	+	+	+	Sufficient (+)
	Quality of evidence		Moderate (due to risk of bias)		
UGAQoL	Overall rating	?	?	+	Indeterminate (?)
	Quality of evidence		No grading if overall rating is indeterminate		
Urogenital symptom scale	Overall rating	+	−	+	Inconsistent (±)
	Quality of evidence		No grading if overall rating is inconsistent		

DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; PROM, patient-reported outcome measure; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.

TABLE 7. *COSMIN risk of bias overall ratings for the remaining measurement properties*

	VSQ	DIVA	UGAQoL	Urogenital symptom scale
Structural validity	Very good[a] Inadequate[a] Very good[c] Very good[c]	Very good[b] Very good[b]		
Internal consistency	Very good[a] Very good[f]	Very good[b]	Doubtful[d]	Very good[e]
Crosscultural validity/measurement error				
Reliability	Adequate[a]	Adequate[b]	Doubtful[d]	Very good[e]
Measurement error				
Criterion validity				
Hypotheses testing for construct validity	Very good[a] Very good[f]	Inadequate[b] Adequate[b] Very good[g] Inadequate[h] Very good[h]	Doubtful[d]	
Responsiveness				Inadequate[e]

COSMIN, COnsensus-based Standards for the selection of health Measurement Instrument; DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; PROM, patient-reported outcome measure; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.
[a]Erekson et al, 2013.
[b]Huang et al, 2015.
[c]Erekson et al, 2016.
[d]McKenna et al, 1999.
[e]Chen et al, 2010.
[f]Fernandez-Alonso et al, 2017.
[g]Hunter et al, 2016.
[h]Nappi et al, 2019.

values, and information on response shift could be extracted of the included studies.

Regarding feasibility aspects, all four PROMs are self-reported and neither a high mental ability level nor a physical activity level is required. The score calculation for the VSQ, DIVA, and UGAQoL is a simply summing up of the single items, for the urogenital symptom scale no information about the scoring is given. Only one article regarding the DIVA[13] includes a statement about copyright. The DIVA is copyrighted; however, no charge and no written permission for its use are required from the authors. Only for the UGAQoL, a time interval of 4 to 15 minutes is stated for its completion. The UGAQoL is, however, not accessible since the authors are not willing to disseminate this PROM.

Summary of findings tables and recommendation

The summarized results per measurement property per PROM are presented in Table 10. The overall rating for

TABLE 8. *Characteristics of the included patient-reported outcome measures*

	PROM			
Characteristic	VSQ	DIVA	UGAQoL	Urogenital symptom scale
Construct	Symptoms, emotions, life impact of vulvovaginal symptoms, sexual impact of vulvovaginal symptoms	Activities of daily living, emotional well-being, sexual functioning, self-concept, and body image	Quality of life (extent to which individuals are able to satisfy their needs)	Urogenital atrophy
Target population	Postmenopausal women	Symptomatic postmenopausal women	Women with UGA	Women aged 40 to 60 years
Mode of administration	Self-reported	Self-reported	Self-reported	Self-reported
Recall period	1 wk	4 wk	At the moment[a]	*Supposed*: at the moment (see GCS)
(Sub)scales (number of items)	3 or 4 Scales, 17 items for the full sample, 21 items for sexually active women	5 Scales, 19 items for the full sample, 23 items for sexually active women	20 Items	3 Items
Response options	Yes (1), no (0)	5-Point Likert scale (0-4)	Yes/no (2007[a]: yes, a lot; yes, a little bit; no, not at all)	No information given GCS: 4-point Likert scale (0-3)
Range of scores/scoring	0-20	0-76 (92)	0-20	No information given
Original language	English	English	English, Swedish	Chinese
Available translations	Spanish	Spanish, Italian	–	

DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; GCS, Greene Climacteric Scale; PROM, patient-reported outcome measure; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.
[a]http://www.galen-research.com/measures-database/.

TABLE 9. *Characteristics of the included study populations*

PROM	Ref.	Population		PROM administration			Response rate
		N	Age mean (SD)	Setting	Country	Language	
VSQ	Erekson et al, 2013	120	66.3 (10.9)	General gynecology practice serving women seeking treatment for pelvic floor disorders	USA	English	99%
	Erekson et al, 2016	358	62.0 (9.9); With vulvar symptoms: 60.9 (9.2); without vulvar symptoms: 63.1 (10.4)	Internal medicine practices and adult senior centers	USA	English	Senior centers: 96% primary care practices: 85.7%
	Fernandez-Alonso et al, 2017	150	59.5 (4.9); With vulvovaginal symptoms: 59.9 (5.1); without vulvovaginal symptoms: 58.2 (4.0)	Outpatient clinic	Spain	Spanish	81.3%
DIVA	Huang et al, 2015	757	56.2 (8.5)	Parent RRISK3 cohort (long-time enrollees in the Kaiser Permanente Northern California integrated health care delivery system) see Huang et al, 2015	USA	English	98.4%
	Hunter et al, 2016	2,160	58.9 (6.7)	46 outpatient menopause centers and gynecology centers	Spain and Italy	Spanish and Italian	90%
	Nappi et al, 2019	50	55.9	Postal survey	UK	English	100%
UGAQoL	McKenna et al, 1999	54	67.7	Postal survey	Sweden	Swedish	100%
Urogenital symptom scale	Chen et al, 2010	611	Seminar: 49.7 (5.0) Postal survey: 48.2 (5.6)	Health seminar and postal survey (respondents to a newspaper article)	China	Chinese	Seminar: 100% Postal survey: 78.5%

DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; N, number of participants per study; PROM, patient-reported outcome measure; Ref, reference; SD, standard deviation; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.

internal consistency of the VSQ was inconsistent since not all studies reported Cronbach alpha values $\geq$0.7. We decided to base the overall rating on the majority of the results and therefore downgraded the quality of evidence for one level due to inconsistency. Structural validity, internal consistency, and reliability of the DIVA were all downgraded due to indirectness since the relevant study[13] was partly performed in another population of interest.

The results of the summary of findings tables were used to recommend the most appropriate PROM. The final recommendations according to the COSMIN guidelines[15] for all four PROMs are presented in Table 11.

DISCUSSION

This systematic review assessed the measurement properties of four different PROMs for postmenopausal women with genitourinary complaints. Two PROMs, the VSQ, and the DIVA can be further recommended for use and results obtained with these PROMs can be seen as trustworthy. In our opinion, the DIVA should be preferred over the VSQ since affected women were involved in the item generation phase and a widely recognized qualitative data collection method, focus groups, was used. Furthermore, the sample size of the included studies was considerably higher for the DIVA than for the VSQ which supports the trustworthiness of the results regarding the DIVA. Nevertheless, more validation research on the DIVA and the VSQ is desirable. Especially test-retest reliability of both PROMs should be further assessed since this measurement property had an insufficient overall rating, albeit on a low quality of evidence level. A potential weakness of both PROMs, VSQ and DIVA, is that they do not cover the whole construct of GSM because the urinary component was not taken into account. Even the UGAQoL still has the opportunity to be recommended for use, but due to aspects on feasibility, it is not issued by the authors and therefore, cannot be recommended for use. The Urogenital symptom scale could not be recommended for use for several reasons. First, there was high-quality evidence for an insufficient measurement property (internal consistency). Second, for the PROM development no qualitative method was used. The scale was developed by a group of experts and affected women were not involved. Third, the three-item scale was developed as an additional scale to the Greene Climacteric Scale (GCS); however, this additional scale diminished the model fit of the standard GCS. Even the authors did not support the inclusion of this scale to the standard GCS.[33] Future validation studies should also look at interpretability of PROMs since only little information was available for the currently included PROMs.

All included PROMs were developed before the name changed from vulvovaginal atrophy to GSM. However, two PROMs, UGAQoL and the urogenital symptom scale, have already considered urinary aspects since they referred to urogenital atrophy. Our two preferred instruments, VSQ and DIVA, did not take into account the urinary component. It is important to mention that urinary symptoms as part of GSM are still less studied and understood with respect to

TABLE 10. *Summary of Findings tables*

Structural validity	Summary or pooled result	Overall rating	Quality of evidence
VSQ	CFI: 0.97-0.99	Sufficient	High
DIVA	CFI: 0.978-0.979	Sufficient	Moderate (due to indirectness)
Internal consistency	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
VSQ	0.623-0.87; Sample size: 131-244	Inconsistent → overall rating based on the majority of the results: 5+, 3− → sufficient	Moderate (due to inconsistency)
DIVA	0.82-0.94, Consistent results; sample size: 462-745	Sufficient	Moderate (due to indirectness)
UGAQoL	0.89-0-90, No evidence for sufficient structural validity; sample size: 104 women	Indeterminate	—
Urogenital symptom scale	0.43; Sample size: 290	Insufficient	High
Reliability	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
VSQ	0.55-0.75; Sample size: 91	Insufficient	Low (due to risk of bias and imprecision)
DIVA	0.47-0.72; Sample size: 462-745	Insufficient	Low (due to risk of bias and indirectness)
UGAQoL	0.85-0.92, ICC or Kappa not reported; sample size: 104	Indeterminate	—
Urogenital symptom scale	0.81; Sample size: 52	Sufficient	Moderate (due to imprecision)
Hypotheses testing	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
VSQ	3 Out of 3 hypotheses confirmed	Sufficient	High
DIVA	5 Out of 6 hypotheses confirmed, expectations toward known-groups validity confirmed	Sufficient	High
UGAQoL	2 Out of 2 hypotheses not confirmed	Insufficient	Low (due to risk of bias)
Responsiveness	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
Urogenital symptom scale	Effect size: 0.46; sample size: 19	Indeterminate	—

CFI, Comparative Fit Index; DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; UGAQoL, urogenital atrophy quality of life; VSQ, Vulvovaginal Symptoms Questionnaire.

postmenopausal estrogen deficiency and further elicitation of this component is needed.

Strengths and limitations of this review

This systematic review has several strengths: the protocol was registered; a comprehensive and sensitive search filter was applied; three big databases (MEDLINE, EMBASE, and Web of Science), several small databases, and reference lists of the included studies were searched; predefined eligibility criteria were applied; and the COSMIN risk of bias checklist was used to assess the methodological quality of the included studies. The fact that no additional studies were found in all searched small databases and that only one relevant study was found in the reference lists of the included studies, supports the quality of the search filter developed by our academic librarian. In every step of the review process, at least two independent reviewers were involved. One reviewer (M.G.) carried out every step to ensure consistency during the review process. Discrepancies were frequently discussed and resolved within the research team. A potential limitation of this systematic review is that not

TABLE 11. *Recommendations for use in future GSM trials*

	Category A		Category C	
PROM	Sufficient content validity (any level)	At least low quality evidence for sufficient internal consistency	High quality evidence for an insufficient measurement property	Recommendation
VSQ	✓	✓	×	A
DIVA	✓	✓	×	A
UGAQoL	×	×	×	B
Urogenital symptom scale	×	×	✓	C

DIVA, Day-to-Day Impact of Vaginal Aging Questionnaire; GSM, genitourinary syndrome of menopause; PROM, patient-reported outcome measures; UGAQoL, urogenital atrophy quality of life; VSQ, vulvovaginal symptoms questionnaire.

all reference lists of relevant full-texts were searched for further eligible articles.

CONCLUSIONS

This systematic review suggests that currently two PROMs, VSQ and DIVA, can be recommended for use. Both PROMs cover the genital and sexual component of GSM. Future validation studies should focus on those PROMs. It would be desirable to extend these PROMs to a urinary component to depict the whole construct of GSM.

Acknowledgments: The authors are grateful to Peter Werkmann (Regensburg, Germany) for the screening of all titles, abstracts, and full-texts as a second reviewer.

REFERENCES

1. Portman DJ, Gass ML. Genitourinary syndrome of menopause: new terminology for vulvovaginal atrophy from the International Society for the Study of Women's Sexual Health and the North American Menopause Society. *Menopause* 2014;21:1063-1068.
2. Vieira-Baptista P, Marchitelli C, Haefner HK, Donders G, Perez-Lopez F. Deconstructing the genitourinary syndrome of menopause. *Int Urogynecol J* 2017;28:675-679.
3. Faubion SS, Sood R, Kapoor E. Genitourinary syndrome of menopause: management strategies for the clinician. *Mayo Clin Proc* 2017;92:1842-1849.
4. Lee DM, Tetley J, Pendleton N. Urinary incontinence and sexual health in a population sample of older people. *BJU Int* 2018;122:300-308.
5. Parish SJ, Nappi RE, Krychman ML, et al. Impact of vulvovaginal health on postmenopausal women: a review of surveys on symptoms of vulvovaginal atrophy. *Int J Womens Health* 2013;5:437-447.
6. Gandhi J, Chen A, Dagur G, et al. Genitourinary syndrome of menopause: an overview of clinical manifestations, pathophysiology, etiology, evaluation, and management. *Am J Obstet Gynecol* 2016;215:704-711.
7. Farrell AME. Genitourinary syndrome of menopause. *Aust Fam Physician* 2017;46:481-484.
8. Nappi RE, Palacios S. Impact of vulvovaginal atrophy on sexual health and quality of life at postmenopause. *Climacteric* 2014;17:3-9.
9. Management of symptomatic vulvovaginal atrophy: 2013 position statement of The North American Menopause Society. *Menopause* 2013;20:888-902.
10. Nappi R E, Palacios S, Panay N, Particco M, Krychman M L. Vulvar and vaginal atrophy in four European countries: evidence from the European REVIVE Survey. *Climacteric* 2016;19:188-197.
11. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract* 2006;12:559-568.
12. Erekson EA, Yip SO, Wedderburn TS, et al. The Vulvovaginal Symptoms Questionnaire: a questionnaire for measuring vulvovaginal symptoms in postmenopausal women. *Menopause* 2013;20:973-979.
13. Huang AJ, Gregorich SE, Kuppermann M, et al. Day-to-Day Impact of Vaginal Aging questionnaire: a multidimensional measure of the impact of vaginal symptoms on functioning and well-being in postmenopausal women. *Menopause* 2015;22:144-154.
14. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;350:g7647.
15. Mokkink LB, Prinsen CA, Patrick DL, et al. *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs)—user manual* 2018; Available at: https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/. Accessed June 17, 2019.
16. Vissers T, Vries Rd. Quality of life (QoL) search blocks. Available at: https://blocks.bmi-online.nl/catalog/294. Accessed March 16, 2018.
17. Jansma EP, Vries Rd. Patient reported outcome measures (PROMs) search blocks. Available at: https://blocks.bmi-online.nl/catalog/248. Accessed March 16, 2018.
18. Mackintosh A, Comabella CC, Hadi M, Gibbons E, Fitzpatrick R, Roberts N. PROM group construct & instrument type filters. Available at: http://www.cosmin.nl/images/upload/files/PROM%20Gp%20filtersOCTOBER%202010FINAL.pdf. Accessed March 16, 2018.
19. Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med* 2005;60:833-843.
20. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147-1157.
21. Terwee CB, Jansma EP, Riphagen II, De Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115-1123.
22. Canada AU. Search filter for finding studies on measurement properties for OVID (Medline). Available at: https://www.cosmin.nl/tools/pubmed-search-filters/. Accessed June 17, 2019.
23. Heinl D, Prinsen CA, Drucker AM, et al. Measurement properties of quality of life measurement instruments for infants, children and adolescents with eczema: protocol for a systematic review. *Syst Rev* 2016;5:25.
24. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. *J Med Libr Assoc* 2016;104:240-243.
25. Mokkink LB, De Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* 2018;27:1171-1179.
26. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737-745.
27. Terwee CB, Prinsen CA, Chiarotto A, et al. COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs). User manual; 2017. Available at: https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/. Accessed June 17, 2019.
28. Terwee CB, Bot SD, De Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34-42.
29. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a ''Core Outcome Set''—a practical guideline. *Trials* 2016;17:449.
30. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.
31. Schünemann H, Brozek J, Guyatt G, Oxman A. GRADE Handbook—Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach. 2013. Available at: https://gdt.gradepro.org/app/handbook/handbook.html. Accessed June 17, 2019.
32. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *Open Med* 2009;3:e123-e130.
33. Chen RQ, Davis SR, Wong CM, Lam TH. Validity and cultural equivalence of the standard Greene Climacteric Scale in Hong Kong. *Menopause* 2010;17:630-635.
34. Erekson EA, Li FY, Martin DK, Fried TR. Vulvovaginal symptoms prevalence in postmenopausal women and relationship to other menopausal symptoms and pelvic floor disorders. *Menopause* 2016;23:368-375.
35. Fernandez-Alonso AM, Alcaide-Torres J, Fernandez-Alonso IM, Chedraui P, Perez-Lopez FR. Application of the 21-item Vulvovaginal Symptoms Questionnaire in postmenopausal Spanish women. *Menopause* 2017;24:1295-1301.
36. Hunter MM, Nakagawa S, Van Den Eeden SK, Kuppermann M, Huang AJ. Predictors of impact of vaginal symptoms in postmenopausal women. *Menopause* 2016;23:40-46.
37. McKenna SP, Whalley D, Renck-Hooper U, Carlin S, Doward LC. The development of a quality of life instrument for use with post-menopausal women with urogenital atrophy in the UK and Sweden. *Qual Life Res* 1999;8:393-398.
38. Nappi RE, Palacios S, Bruyniks N, Particco M, Panay N; EVES Study Investigators. The burden of vulvovaginal atrophy on women's daily living: implications on quality of life from a face-to-face real-life survey. *Menopause* 2019;26:485-491.
39. Huang AJ, Luft J, Grady D, Kuppermann M. The day-to-day impact of urogenital aging: perspectives from racially/ethnically diverse women. *J Gen Intern Med* 2010;25:45-51.

