PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Psychometric properties of a self-assessment questionnaire concerning symptoms and impairment in urinary tract infections: the UTI-SIQ-8
AUTHORS	Gagyor, Ildiko; Rentzsch, Katrin; Strube-Plaschke, Stephanie; Himmel, Wolfgang

VERSION 1 – REVIEW

REVIEWER	Anne Holm University of Copenhagen
REVIEW RETURNED	07-Sep-2020

GENERAL COMMENTS	Thank you for the opportunity to read your validation study. Useful, validated, PROMs for UTI are scarce and especially your sensitivity to change analysis and network analysis are most relevant. Your four-item PROM is a valuable contribution to the field. There are two studies you may want to mention in your introduction:[1,2] [1] Holm A, Cordoba G, Siersma V, et al. Development and validation of a condition-specific diary to measure severity, bothersomeness and impact on daily activities for patients with acute urinary tract infection in primary care. Health Qual. Life Outcomes. 2017;15:57. [2] Colgan R, Keating K, Dougouih M. Survey of symptom burden in women with uncomplicated urinary tract infections. Clin. Drug Investig. 2004;24:55–60. I am an author on the former, so the following remarks may be influenced by the choices we made in our own validation process. You do not mention how many of existing diaries have included a content validation procedure. Since this is one of two essential steps in a PROM-validation, you should elaborate on that. You also seem to have skipped this step in your own study. Could you state in your introduction why you did not find it relevant and in your discussion how this may have affected your final PROM? You state that 14-18 items are too large a burden to patients. Do you have a reference on that? For example patient interviews? Prior studies on UTI with a low response rate due to a large item pool? And could you elaborate on the impact on relevance and coverage of the included items when the item pool was narroved down? You include gynaecologists in your practice population. You may have to shortly describe how these are organized in your country in order to judge if your results can be transferred to primary care in other countries.

	You have based your items and scores on previous PROMs. Different PROMs use different items and scores. Could you explain how you ended up with these specific four items and five response categories? You seem to have performed a thorough psychometric validation using several different validation methods. Why did you not test for unidimentionality? This step is quite important in order to make sum-scores. I do not know enough about network analysis to know if this is performed correctly, but you use a sum-score from your PROM. How have you ensured, your items could actually be summed, ie. That they are part of the same dimension? You analysis and previous research points to the opposite. What was your response rate? Again, in your discussion you should include the above study by Holm et al since it also includes a day-to day diary and measurement on impact on daily activity. However, that study had only a 80% response rate which may have been due to the numerous items. Was yours better? Also, we found low abdominal pain to be two different entities (uncomfortable pressure and actual pain) where “uncomfortable pressure” was linked to dysuria but pain was not. This may be relevant to compare to your results since you also find abdominal pain to be a different entity. Suddenly in your discussion, you mention low back pain as an item in your PROM. Have I misunderstood something? Was your PROM longer but you only validated some of the items? It is interesting you find good use of a mobile app. Did you find any groups not able to use this? The elderly? We decided against it for that reason but the elderly may be completely fine with it now.

REVIEWER	Plata, Mauricio Fundacion Santa Fe de Bogota Universidad de los Andes school of medicine
REVIEW RETURNED	08-Sep-2020

GENERAL COMMENTS	Thank you for the opportunity to review this article. Some comments outlined below: 1. Sample size was described to be 200 subjects but only 131 were included. Please be more specific on how this will affect the results of the study 2. Describe the cultural adaptation done into German with the proper process of translation from English and back translation 3. Please explain how content validity was assessed in the original questionnaire and how the questions included were assembled together. Were experts in the topic who determined which questions should be included? Please elaborate more on the process. 4. Construct validity were assessed by convergent and discriminant validity however it is highly recommended to explain broadly the findings in the result section. 5. I suggest that the network analysis will be reviewed by a biostatistician 6. Discussion is well supported and documented 7. Despite how well conducted the trial can be, the main research question will not only be the validation process of the UTI scale, it is its main usefulness itself. Will I change my practice with this new questionnaire? If I apply the questionnaire to my patients the treatment options will change or will remain being the same? That

	will be the main issue to be addressed by the authors that I do not see stated anywhere in the document. In the other hand, there is no categorization of the final score to determine level risk which make more difficult for a clinician to interpret the final score or the score change after any therapy.

REVIEWER	Prof. Dr. Abdülbari Bener Istanbul University, Dept. of Biostatistics & Medical Informatics
REVIEW RETURNED	10-Oct-2020

GENERAL COMMENTS	Overall, this study addresses an important psychosocial and public health issue well written and presented. However, the methods section is grossly deficient from epidemiological and statistical point view. It seems subjctes particioants are non-randomly selected which subject and conclusion might be considered as a bias. The authors reported that the survey was conducted among 120 women. 1. Title: The title is adequate. 2. Abstract: The abstract is adequately addressed study; but conclusion NOT stated 3. Introduction: The introduction is NOT adequate. Most of the text parts are belong to the discussion Methods: a. Study design: Cross-sectional study was carried out .on 120 women b .Sampling technique: It has NOT been adequately described c. Recruitment of Subjects: It has NOT been described adequately (Convenient sampling???) hap hazard sampling method? d. Study duration: The study was performed from November 2016 to May 2018. e. Setting: It has NOT been described adequately f. Eligibility criteria; It has been reported adequately g. Data collection and measurements: Data collection tool and analysis described in very briefly. h- Statistical analysis: Simple descriptive statistical analysis has been performed. h. Ethical consideration: Ethics approval was obtained from the ethics committee of Göttingen Medical School (17/4/16). 5.Results: These daily changes in an individual's symptoms moved parallel with daily changes in the EQ-5D-5L (b = 1.68, SE = 0.12, $p < .001$) and the VAS (b = 0.03, SE = 0.003, $p <$ .001), also highlighting convergent validity with respect to daily changes in symptom severity 6. Discussion: The discussion is NOT well and concise written. 7. Contribution to the literature: The authors should provide key points and the contribution of current study to literature and what messages are provided with the present study? 8. Limitations: The author reported some of the key limitations of this study, which not mentioned randomization and eligibility criteria. I think the major concern of this submission is it lacks sufficient novelty and or original study, as many of such similar studies were seen published in currently documented literature.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1

1. here are two studies you may want to mention in your introduction:[1,2]

[1] Holm A, Cordoba G, Siersma V, et al. Development and validation of a conditionspecific diary to measure severity, bothersomeness and impact on daily activities for patients with acute urinary tract infection in primary care. Health Qual. Life Outcomes. 2017;15:57.

[2] Colgan R, Keating K, Dougouih M. Survey of symptom burden in women with uncomplicated urinary tract infections. Clin. Drug Investig. 2004;24:55–60.

I am an author on the former, so the following remarks may be influenced by the choices we made in our own validation process.

Authors' comment

We thank you for these suggestions. Both references fit well in the Introduction as a further proof for the important dimensions and items that can also be found in our questionnaire (new ref 17 and 24). We also discuss the specific role of (abdominal) pain in your and our study (see p 17, second para) as you also suggested in a later concern (see also our comment to # 9).

The Colgan et al. study supports our decision to choose frequency, urgency, dysuria and abdominal pain as the most important and most commonly reported symptoms (cited now on p 6, second para, p 17, second para, lines 51 ff.).

2. You do not mention how many of existing diaries have included a content validation procedure. Since this is one of two essential steps in a PROM-validation, you should elaborate on that.

You also seem to have skipped this step in your own study. Could you state in your introduction why you did not find it relevant and in your discussion how this may have affected your final PROM?

Authors' comment

Thank you for your suggestion to make the process of development of the questionnaire more transparent. We now shortly report our decision for the two dimensions and 8 items of our questionnaire in the Introduction (p 4, lines 16 ff. and p 5, lines 17 ff.) Moreover, we introduced a new subchapter "Development of the UTI-SIQ-8 Questionnaire" in the Method section (second para on p 6) where we describe the development process of the questionnaire in detail (see also our comment to # 6).

2

3. You state that 14-18 items are too large a burden to patients. Do you have a reference on that? For example, patient interviews? Prior studies on UTI with a low response rate due to a large item pool?

And could you elaborate on the impact on relevance and coverage of the included items when the item pool was narroved down?

Authors' comment

It is not our intention to regard 14-18 items as a burden to patients. However, when it comes to daily diary studies, it is important to lower the number of items in order to ensure the least possible burden to participants. According to Rolstad's et al. review on response burden (https://pubmed.ncbi.nlm.nih.gov/22152180/), there is a greater chance of response when patients were presented with a comparatively shorter questionnaire. However, the authors warned, at the same time, that factors other than length may be at least as important and argue that the quality of the content matters more from the patient's point of view than the length per se. So, we think that the length of a survey should neither be overestimated nor underestimated. Moreover, we feel an obligation, and should do everything, not to put patients under any kind of (additional) stress. More items than absolutely necessary, daily or even more frequently asked, may be a stress so that we consider the 8-item-survey an

advantage of our study. We now discuss the length of our questionnaire in a very balanced way so that it becomes clear that a short survey may meet a woman's interest not to be overly burdened, but that it is not a sign of quality per se (see p 15, last para).
4. You include gynaecologists in your practice population. You may have to shortly describe how these are organized in your country in order to judge if your results can be transferred to primary care in other countries.
Authors' comment
In Germany, gynaecologists are often involved in primary care of women. Therefore, many women contact their gynaecologists first with UTI symptoms. We added this information in the Methods section (see p 7, paragraph 2, including the new ref 28)
5. You have based your items and scores on previous PROMs. Different PROMs use different items and scores. Could you explain how you ended up with these specific four items and five response categories?
Authors' comment
The symptom score we used in our previous clinical trials (ref 5 and 23) was developed based on Ferry's et al. study (ref 11) and similar studies which showed that urgency, frequency, dysuria and low abdominal pain were the most common symptoms reported by affected women and are best suited to measure the success of clinical trials. We used a rating scale with five response categories in order to allow sufficiently differentiated responses from patients regarding their symptom severity and bothersomeness and at the same time an adequate verbal labelling of all single points of the response scale. Response scales with more scale points often face the problem of an inadequate verbal labelling of every response option. We, therefore, decided to use a unipolar rating scale with five response options for psychometric reasons. We added more information on this development in the revised version (see p 6 f., second para).
3
6. You seem to have performed a thorough psychometric validation using several different validation methods.
Why did you not test for unidimentionality? This step is quite important in order to make sum-scores. I do not know enough about network analysis to know if this is performed correctly, but you use a sum-score from your PROM. How have you ensured, your items could actually be summed, ie. That they are part of the same dimension? Your analysis and previous research points to the opposite.
Authors' comment
We thank Reviewer 1 for raising this important point. As described in our comment to # 2 and # 5, urgency, dysuria, frequency and low abdominal pain are, according to many international studies, the most frequent symptoms reported by affected women and build the backbone of our questionnaire. Results from a principal component analysis indicated either a one-factor solution or a two-factor solution. In the two-factor solution, 6 of the items, referring to urgency, dysuria, and frequency loaded high on the first factor, while both pain items loaded on a second factor. The 2 pain items were not sufficient to build a psychometrically sound scale of its own, so we decided to accept the one-factor solution including all symptoms. A main rationale to do this was that the reliability of this single scale, expressed in Cronbach's Alpha, was excellent. Moreover, the corrected item-total correlation of the 2 pain items, with .44 and .39 on day 1 and .58 each on day 2, was sufficient to consider these results as an indicator for the unidimensionality of the scale.
Stimulated by your comment, we now additionally performed a confirmatory factor analysis to test the one-factor solution more properly (see p 9, last line and p 10, first lines; p 13, third para, lines 37 ff.). By including correlated residuals between the items referring to "severity of dysuria" and "impairment due to dysuria" and the items on "severity of low abdominal pain" and "impairment due to low abdominal pain", the analysis revealed a satisfactory fit of the model, with a CFI of .92 (comparative fit index, which should be > .90) and a SRMR of .07

(standardized root mean square, which should be < .11); only the RMSEA (root mean square error of approximation) lies with a value of .16 above the recommended threshold value of .08). Based on our results from principal component analysis, corrected item-total correlation and confirmatory factor analysis, we are confident that the scale score consisting of the sum of the item responses reflect severity of and impairment by symptoms as one single dimension. We think this seems logical from the patient perspective: Whatever the symptom: severity influences perceived impairment and vice versa. So, we decided to follow no longer the usual distinction between severity and impairment (or bothersomeness) but made clear that the both pain items have a special status (p 17, last para).

7. What was your response rate?
Authors' comment
We used screening lists in the participating practices, to determine the participation rate. The practice nurses were responsible to carefully filling out these lists. However, due to the time pressure in day-to-day business, the lists did not prove to be reliable. Therefore, we did not feel safe enough to report a response rate (now mentioned on p 8, lines 13 ff. in the new version). We would like to add that a selection bias, even if it happened, is rather unlikely to corrupt the results of a study for validation of a questionnaire.

8. Again, in your discussion you should include the above study by Holm et al since it also includes a day-to day diary and measurement on impact on daily activity. However, that study had only a 80% response rate which may have been due to the numerous items. Was yours better?
Authors' comment
The Holm et al. study is considered in the revised version of the manuscript (see also our comment to # 1). It was not possible to determine a participation rate because our efforts were ruined by the practical conditions in everyday business (see our comment to # 7). Therefore, we cannot judge whether our participation rate was better or worse than in other studies. Overall, we can state that of 131 patients included, complete or almost complete data sets of 120 were accessible for analysis. In other words, the dropout rate was low.
9. Also, we found low abdominal pain to be two different entities (uncomfortable pressure and actual pain) where "uncomfortable pressure" was linked to dysuria but pain was not. This may be relevant to compare to your results since you also find abdominal pain to be a different entity.
Authors' comment
Yes, an interesting parallel – and also an interesting difference. We now discuss this issue in more detail by comparing our findings with your study results, also referring to Colgan et al. who showed that urgency, frequency and dysuria was experienced by nearly all women, abdominal pain only by about 60% (p 17, last para).
10. Suddenly in your discussion, you mention low back pain as an item in your PROM. Have I misunderstood something? Was your PROM longer but you only validated some of the items?
Authors' comment
Thank you very much for carefully reading our paper. It was simply a typing error, perhaps caused by the Leydon et al. study which we cited in this paragraph. Indeed, these authors talked about LOW BACK pain as a symptom that at least some women experience. To make things no longer confusing, we deleted the whole paragraph (see p 18, first lines). We now emphasize that there may be symptoms such as abdominal pain that are not relevant/burdensome for all women with a UTI (this may explain its special status in the factor analysis) but, at least, important for a group of women (now clarified on p 17, last para, with reference to the Colgan et al. study, ref 24).

11. It is interesting you find good use of a mobile app. Did you find any groups not able to use this? The elderly? We decided against it for that reason but the elderly may be completely fine with it now.

Authors' comment

We share your view that the use of a mobile app can be detrimental to the inclusion of elderly patients. However, the 'mobile app group' had an average age of 43 years, in other words, it also included many 'older' women. Looking only at women > 65 years, still half (6/12) of them

5

used the app (we now report this result in the new version on p 12, second para). We think this is a positive sign for the use of apps in the future. Since we did not ask the participants about their experience with the app, it is difficult to say whether the elderly felt completely comfortable with mobile apps, but the study indicates that the elderly are willing, and are able, to use electronic data collection methods. Perhaps, we sometimes underestimate them in terms of their technical skills (see p 18, lines 32 f.).

Reviewer #2

12. Sample size was described to be 200 subjects but only 131 were included. Please be more specific on how this will affect the results of the study

Authors' comment

A total of 131 women participated in this study. After exclusion of 11 patients due to screening failure or technical reasons, the valid sample size was 120 women (see p. 12, first para and our comments to # 7 and # 8). The number of participants is sufficient for a validation study and a selection bias is rather unlikely (see our comments to # 20 and # 26).

13. Describe the cultural adaptation done into German with the proper process of translation from English and back translation

Authors' comment

We developed the questionnaire based on the items used by Ferry et al. and others (see our comments to # 2 and # 5) and involved a multi-professional team of GPs, young doctors, psychiatrists, social scientists and study nurses in the development process before using the questionnaire in the trial. The team gave feedback to the content, understanding, wording, order of the questions and the answer scores (we have added this information in the new version, see p 6, second para). A back translation was not part of the development process because it was not directly translated from an English questionnaire.

14. Please explain how content validity was assessed in the original questionnaire and how the questions included were assembled together. Were experts in the topic who determined which questions should be included? Please elaborate more on the process

Authors' comment

The new subchapter "Development of the UTI-SIQ-8 Questionnaire" in the Methods section (second para on p 6) describes how we ensured content validity (please, see also our comments to # 2 and # 5).

15. Construct validity were assessed by convergent and discriminant validity however it is highly recommended to explain broadly the findings in the result section..

6

Authors' comment

We now explain the findings in more detail in the Results section (see pp. 13, last para and p 14, first para of the revised manuscript and our comments to # 6).

16. Despite how well conducted the trial can be, the main research question will not only be the validation process of the UTI scale, it is its main usefulness itself. Will I change my practice with this new questionnaire? If I apply the questionnaire to my patients the treatment options will change or will remain being the same? That will be the main issue to be addressed by the authors that I do not see stated anywhere in the

document. In the other hand, there is no categorization of the final score to determine level risk which make more difficult for a clinician to interpret the final score or the score change after any therapy.

Authors' comment

There are several implications for the future practice. The questionnaire could be used for the baseline assessment of symptom severity in women with uncomplicated UTI and may help to decide whether to start with symptomatic treatment by using, for example, a delayed prescription approach, as described by ref 8 and mentioned in our paper (p 18, lines 18 f.). It can also be used to follow-up the course of the symptoms and their duration on a day by day basis as described in the section 'Implications for practice and research' (p 18, lines 20 f.). We are aware of the need of more research on identifying the cut-off values for clinically relevant changes of the symptom course and duration before implementing it in the practice (see p 18, lines 28 ff.).

Reviewer #3

17. Overall, this study addresses an important psychosocial and public health issue well written and presented. However, the methods section is grossly deficient from epidemiological and statistical point view. It seems subjctes particioants are nonrandomly selected which subject and conclusion might be considered as a bias. The authors reported that the survey was conducted among 120 women.

Authors' comment

We thank the Reviewer for the overall positive assessment of our study. Regarding the methods, we would like to emphasize that this study did neither collect epidemiological data nor was it the aim of the present study to decide whether pain management alone is equivalent to antibiotic therapy. The aim of the present study was to develop and validate a new patient reported outcome measure for uncomplicated urinary tract infections and to analyse its psychometric properties. We are, therefore, confident that sampling bias due to non-randomization does not corrupt the results of our study (see also our comment to # 7 and # 20). Indeed, we analysed responses from 120 participants who provided more than 700 data entries across this survey-based diary study. This sample size allowed a proper analysis of the psychometric properties of the questionnaire (see p 16, first lines and also our comments to # 12).

7

18. Abstract: The abstract is adequately addressed study; but conclusion NOT stated

Authors' comment

We inserted further parts of the conclusion from the main text in the Abstract (see p 2, lines 41 ff.) so that the conclusion is now better supported from the results of our study.

19. Introduction: The introduction is NOT adequate. Most of the text parts are belong to the discussion

Authors' comment

With respect, we disagree with the Reviewer. We start the Introduction with a statement about the relevance of giving women a voice ("From the women's perspective, particularly concerning symptom severity, the impairment and the impact on daily activities are of high relevance"; see end of first para). This is followed by the presentation of the state of the art in patient reported outcome measures, mentioning the most relevant symptom questionnaires, including possible shortcomings (see second para). We then mention a new questionnaire of our working group, which has been pretested in two of our studies but not validated before (see third para). Consequently, the last para introduces the aim of the present study, i.e., "to validate this questionnaire" and, as an additional objective, "to prove the applicability of a mobile application among participating women".

We believe this is an adequate introduction of a validation study and not a discussion. Instead, not until the Discussion section, we interpret the findings from the present study and discuss them with respect to previous research and findings.

20. Methods: Sampling technique: It has NOT been adequately described. Recruitment of Subjects: It has NOT been described adequately (Convenient sampling???) hap hazard sampling method?

Authors' comment

Our results are based on a convenient sampling. The selection was made arbitrarily by the responsible practice nurses. Female patients (18 years and older) with typical symptoms of a UTI were included in the study by practice nurses and physicians according to the exclusion criteria (e.g. signs of a complicated infection), described in the new version of the manuscript on p 8, lines 25 ff. Despite a convenient sampling strategy, the consistency of our data with the results of other studies – for example with regard to the E. coli detection rate and antibiotics prescription rate – suggests that the sample is representative for women with uncomplicated UTIs (Holm A et al. Scand J Prim Health Care 2019; 37: 83–89; Schmiemann G et al. BMC Urol 2012; 12: 33).

21. Setting: It has NOT been described adequately

Authors' comment

We added the information that gynaecologists, too, may be the first point of contact in Germany for women with a UTI, the most relevant difference of gatekeeper systems in the management of UTIs (see p 7, lines 47 ff. with new ref 28 and our comment to # 4).

8

22. Data collection and measurements: Data collection tool and analysis described in very briefly.

Authors' comment

We tried to balance detailed information against a short and non-lengthy information for readers in the previous version. So, we strictly concentrated on those information which were essential for the validation process. We now provide some more details about data collection (p 8, lines 25 ff.). As already in the previous version, we add two Appendices, one with the questionnaire, and the other with details of the electronic version of the questionnaire. We now explain that the electronically collected data were transferred directly into the database while the non-electronic data (paper-and-pencil version of the questionnaire, urine test results, telephone interviews) had to be entered into the database by the study team (p 9, lines 10 ff.).

23. Results: These daily changes in an individual's symptoms moved parallel with daily changes in the EQ-5D-5L ($b = 1.68$, $SE = 0.12$, $p < .001$) and the VAS ($b = 0.03$, $SE = 0.003$, $p < .001$), also highlighting convergent validity with respect to daily changes in symptom severity

Authors' comment

We agree with the Reviewer's assessment of our results on sensitivity to change.

24. Discussion: The discussion is NOT well and concise written.

Authors' comment

We made changes in the Discussion section according to several suggestions of the reviewers (see also our comments to # 9, 10 and 16) and hope the discussion is more concise now.

25. Contribution to the literature: The authors should provide key points and the contribution of current study to literature and what messages are provided with the present study?

Authors' comment

We provided key points and emphasized our contribution to the literature in the Journal's mandatory category 'Strengths and limitations of this study" (p 3 of the manuscript). Moreover, you can also find our contribution to the literature in the subchapter 'Comparison with existing literature' (p 16 ff.) and our key message in the 'Conclusions' on p 18.

26. Limitations: The author reported some of the key limitations of this study, which

not mentioned randomization and eligibility criteria.
9
Authors' comment
The study was designed for validation purposes in female patients suffering from UTI so that randomization was not necessary. We are aware of a small selection bias towards those with a stronger affinity to enter their data electronically (see also our comment to # 12). Due to this, older women were at higher risk to be excluded from the study. But this did not corrupt the result of our study.
27. I think the major concern of this submission is it lacks sufficient novelty and or original study, as many of such similar studies were seen published in currently documented literature.
Authors' comment
As we stated in our comment to # 24, the novel aspects of the study are presented, besides others, in the Journal's mandatory category 'Strengths and limitations of this study" (p 3 of the manuscript) and in our key message in the 'Conclusion' on p 17. We are therefore confident that the present study provides an important and novel contribution to the existing literature.

VERSION 2 – REVIEW

REVIEWER	Anne Holm University of Copenhagen, Denmark
REVIEW RETURNED	17-Dec-2020

GENERAL COMMENTS	Thank you for the revised version. It has much improved. This PROM will be a beneficial addition to existing PROMs. I hope you will consider using patient interviews in development of future PROMs to ensure relevance and coverage.