Statistical Analysis Plan:

The below describes the statistical considerations for analysing the data collected in the study, including the specific analytical steps and methods used at each step, and the assumptions tested and verified. Best practice recommendations for the development and validation of new patient-reported outcome measures (PROMs) by the COSMIN initiative were followed throughout (Mokkink et al., 2018; Prinsen et al., 2018).

Stage 1: Expert screening

Best practice for Delphi methodology analysis was followed (McMillan et al., 2016). After Round 1, median relevance and clarity ratings were computed and presented in Round 2 alongside anonymised qualitative feedback. Following Round 2, new median ratings were calculated and analysed in conjunction with all the qualitative feedback. Content validity indices for items (I-CVI) were computed by dividing the number of expert clinicians who scored an item's relevance/clarity as at least four out of six by the total number of expert clinicians (Yusoff, 2019). It was specified a priori that a minimum I-CVI of 0.75 would be required to indicate acceptable consensus of content validity, with a minimum median score of 4 (McMillan et al., 2016; Polit & Beck, 2006). Refinements were made to create second versions of the RUTISS and the RUTIIQ, ready for patient testing in Stage 2.

Stage 2: Cognitive interviews

Guidance by Willis on conducting and analysing cognitive interviews for questionnaire design and appraisal was followed (Willis, 2005; Willis, 2015). Anonymised verbatim transcripts of the interviews were created using speech-to-text software (Otter; https://otter.ai) with transcription errors manually corrected. The corresponding portions from every interview transcript for each PROM item were collated to create individual documents for each item containing all feedback pertaining to that question from across the entire sample. Top-down question feature coding was applied to these summaries using the Question Appraisal System (QAS-99) (Willis & Lessler, 1999). The systematic evaluation of every response to each PROM item for both the RUTISS and the RUTIIQ was conducted to identify potential problems, which were corrected based on verbatim quotes and interviewer field notes through collaborative analysis with the entire research team. If there were any uncertainties found in the first round of interviews about whether to include or exclude an item, the item was retained in the second round to gain further feedback. Further versions of the RUTISS and the RUTIIQ were created at this stage and used in the second interview round. The same process was undertaken after this, with an amended version created for pilot testing.

Stage 3: Online pilot testing

Data preparation and assumption testing:

Total scores were calculated for each of the Recurrent Urinary Tract Infection Symptom Score (RUTISS) subscales at Baseline and Test-Retest Assessments. Total scores for the RUTISS as a whole measure were not computed to avoid misrepresenting multidimensional constructs as a single score (Strauss & Smith, 2009). Data collected using the Urinary Tract Infection Symptom Assessment (UTISA) and Numerical Pain Rating Scale (NPRS) were scored according to their standard scoring systems.

The following assumptions were assessed: (1) normality – Shapiro-Wilk $p > .05$, so non-parametric analyses were conducted where applicable; (2) linearity – examination of scatter plots confirmed no violation; (3) multicollinearity – Variance Inflation Factor scores for the target label data as well as possible demographic characteristics were sufficiently close to 1; (4) independence of observations – Durbin-Watson statistic was approximately 2 for each subscale

score; and (5) homoscedasticity – (inspection of the standardised residuals plotted against standardised predicted values confirmed no violation.

Case-wise diagnostics of the subscale scores indicated five possible outliers outside 3*SD*; however, since these were not measurement errors and could feasibly reflect true variation within the population, they were not removed. Analyses conducted with and without these cases indicated that they had no statistically significant effect on the outcomes ($p > .05$).

Statistical analyses:

Exploratory factor analysis (EFA) was conducted for both the RUTISS and RUTIIQ data. Suitability for EFA was checked through bivariate correlation matrices with coefficients greater than .80 examined for potential multicollinearity (Field, 2013). Preliminary data analyses were undertaken to assess the factorability of the dataset using Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (Yong & Pearce, 2013). Principal Axis Factoring was selected for application above Principal Components Analysis and Maximum Likelihood Analysis to facilitate determination of a latent factor structure (Fabrigar et al., 1999; Yong & Pearce, 2013). Varimax rotation with Kaiser Normalisation was applied to minimise cross-loading of items on factors (Fabrigar et al., 1999). Visual inspection of scree plots determined the retained number of factors, combined with evaluation of eigenvalues exceeding Kaiser's criterion of 1 (Kaiser, 1960; Watkins, 2018). Items with extracted communalities less than .40 were removed, and a minimum factor loading of .40 was defined as acceptable a priori (Fabrigar et al., 1999; Watkins, 2018).

Internal consistency was measured for each finalised post-EFA RUTISS subscale and the RUTISS as a whole measure using Cronbach's alpha (Cronbach, 1951). To assess test-retest reliability, intraclass correlation coefficients (ICC) and their 95% confidence intervals were computed based on a single-rating, absolute-agreement, two-way mixed effects model, as recommended for test-retest reliability analysis of measures intended to demonstrate outcomes at an individual (single-rater) level (Koo & Li, 2016). Construct validity was assessed by computing the level of Spearman's correlation between each RUTISS subscale and the observed UTISA and NPRS scores (Schober et al., 2018).

Linear regression analyses were also conducted to examine whether there was a statistically significant effect of any demographic variables on the RUTISS scores (i.e., measurement invariance), and to test the predictive value of the global rating of change (GRC) scale.

An Automated Readability Index, known to be especially useful when applied to technical, non-narrative text (Kincaid & Delionback, 1973), was computed to estimate the literacy level required for comprehension of the RUTISS.

References:

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Field, A. (2013). *Discovering statistics using SPSS* (4 ed.). SAGE Publications.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141-151. https://doi.org/10.1177/001316446002000116

Kincaid, J., & Delionback, L. (1973). Validation of the Automated Readability Index: A follow-up. *Human Factors*, *15*(1), 17-20. https://doi.org/10.1177/001872087301500103

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

McMillan, S. S., King, M., & Tully, M. P. (2016). How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy*. https://doi.org/10.1007/s11096-016-0257-x

Mokkink, L. B., De Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, *27*(5), 1171-1179. https://doi.org/10.1007/s11136-017-1765-4

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147-1157. https://doi.org/10.1007/s11136-018-1798-3

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ane.0000000000002864

Strauss, M. E., & Smith, G. T. (2009). Construct validity: advances in theory and methodology. *Annu Rev Clin Psychol*, *5*, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219-246. https://doi.org/10.1177/0095798418771807

Willis, G. (2005). *Cognitive Interviewing: A Tool For Improving Questionnaire Design*. SAGE Publications.

Willis, G., & Lessler, J. (1999). *Question Appraisal System QAS-99*. Research Triangle Institute.

Willis, G. B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford University Press. http://ebookcentral.proquest.com/lib/reading/detail.action?docID=1987880

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79-94. https://doi.org/10.20982/tqmp.09.2.p079

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal*, *11*(2), 49-54. https://doi.org/10.21315/eimj2019.11.2.6