## SUPPLEMENTARY METHODS

### User experience survey

Rationale behind the selection of survey questions

To assess for usability and acceptability of *Rheumatic?,* we designed a survey to retrieve information on users' general impression of the questionnaire regarding usefulness (question 3), as well as users' specific impression on the number of questions (question 1), clarity of questions (question 2) and coverage of questions (question 3). We also included the most classical Net Promotor Score (NPS) question (1): *Would you recommend Rheumatic? to a friend or other patient* (question 5), with the follow-up open-ended question: *Do you have suggestions for improving Rheumatic?*. The rationale behind the selection of survey questions was to gather information on if and how *Rheumatic?* can be improved.

Rationale behind the evaluation of survey response scores

Responses to survey questions 2 to 5 were recorded on an 11-point Likert scale (0-10), where 0 is the most negative response and 10 is the most positive response. Despite using NPS-style questions, we concluded that calculation of the Net Promotor Scores by re-categorisation of scores into *detractors* (score 0-6), *passives* (score 7-8) and *promoters* (score 9-10), followed by subtracting the percentage of *detractors* from the percentage of *promoters* – as suggested by the NPS methodology (1) – was not optimal for our dataset, since leaving out the *passives* (scoring 7 or 8) would discard 46-52% of our data.

We agree with the criticism of the NPS system put forward by Kristensen and Eskildsen (2) among others, in particular the interpretation of a score 6 as being negative (*i.e.* classified as *detractors*) and the loss of data (*i.e.* by leaving out the *passives*). Notably, Kristensen and Eskildsen show that using other cut-off points for re-categorisation (score 0-4 for *detractors*, score 5-7 for *passives*, and score 8-10 for *promoters*) is significantly more accurate than using the NPS categories suggested by Reichheld (1). Still, they argue that using the original 0-10 rating scale would lead to higher precision and better predictive power than using collapsed categories.

Hence, we choose the original 0-10 rating scale for our analysis, and present results (per item) as percentage of users per score category and mean scores, as well as the proportion of users that scored 6-10 (interpreted as being positive) and the proportion of users that scored 0-4 (interpreted as being negative), leaving out the midpoint score 5 (interpreted as being neutral).

The rating scale for survey question 1, *How appropriate did you find the number of questions*, was also 0-10, but here 0 = too few questions, 5 = good number of questions, and 10 = too many questions. We present results as percentage of users per score category, as well as the proportion of users that scored 4-6 (interpreted as being positive to the number of questions), the proportion that scored 0-3 (interpreted as being negative to the number of questions, *i.e.* too few questions), and the proportion that scored 7-10 (interpreted as being negative to the number of questions, *i.e.* too many questions).

Analysis of the free text from the open-ended question

After excluding responses where participants had written “no” or “no comment”, we randomly selected n=500 responses and manually went over the comments. Twenty seven percent of these contained more complex phrases indicating “no comments”, or positive feedback. The remaining responses were categorized based on subject (notably, a response could be flagged with multiple categories). In the current study, we report categories identified in ≥5% of the responses.

To control for sample selection bias, we repeated the random selection process, with n=300 and n=400 responses, respectively. Each sample selection gave a similar outcome, indicating that the conclusions based on our selection of n=500 responses can be extended to the whole study population.

## REFERENCES

1. Reichheld FF. The one number you need to grow. Harv Bus Rev. 2003;46–54:124.

2. Kai Kristensen and Jacob Eskildsen. Is the NPS a trustworthy performance measure? The TQM Journal, 2014, Vol. 26 No. 2, pp. 202-214 ©Emerald Group Publishing Limited 1754-2731. DOI 10.1108/TQM-03-2011-0021.