May 2016

This is an author's draft of an accepted article submitted and published in Multiple Sclerosis Journal DOI: http://msj.sagepub.com/content/early/2016/07/14/1352458516657441

Measuring treatment satisfaction in MS: Is the Treatment Satisfaction Questionnaire for Medication (TSQM) fit for purpose?

Vermersch P,[1] Hobart J,[2] Dive-Pouletty C,[3] Bozzi S,[3] Hass S,[4] Coyle PK[5]

[1]Department of Neurology, University of Lille, Lille, France; [2]Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK; [3]Sanofi Genzyme, Chilly-Mazarin, France; [4]Sanofi Genzyme, Cambridge, MA, USA; [5]SUNY at Stony Brook, Stony Brook, NY, USA

Corresponding author: Patrick Vermersch, Department of Neurology, University of Lille, Rue Emile Laine 59037, Lille cedex, France; Email: patrick.vermersch@univ-lille2.fr; Tel: +33 3 20 44 57 65; Fax: +33 3 20 44 44 84.

Target journal: *Multiple Sclerosis Journal*

Keywords:

Multiple Sclerosis, Teriflunomide, Psychometrics, Treatment Satisfaction, Outcomes Assessment, Disease-Modifying Therapy, Relapsing–Remitting

Running Title:

Evaluation of the TSQM in Patients with MS

Word count: 3,000 (journal limit 3,000)

Abstract word count: 200 (journal limit 200)

Abstract

Background

The Treatment Satisfaction Questionnaire for Medication (TSQM) was designed to assess patient treatment satisfaction in chronic diseases. Its performance has not been examined in MS. The 14 items of the TSQM cover 4 domains: Effectiveness; Side Effects; Convenience; and Global Satisfaction.

Objective

To evaluate performance of the TSQM in patients with relapsing MS, using data collected from the TENERE study (NCT00883337), in which 324 patients received oral teriflunomide or subcutaneous interferon beta-1a for ≥48 weeks.

Methods

Five measurement properties were examined using traditional psychometric methods: data completeness, scale-to-sample targeting, scaling assumptions, reliability (including test-retest), and construct validity (internal: item-level scaling success, confirmatory factor analysis, and exploratory factor analysis; external: convergence, discrimination, and group differences).

Results

There were few (<2%) missing item data; domain scores could be computed for all patients. Score distributions were skewed towards higher satisfaction; 2 domains had marked ceiling effects. Scaling assumptions were supported. Internal consistency reliability was high (Cronbach's α >0.90). Internal validity tests supported item groupings. Correlations supported convergent and discriminant construct validity; hypothesis testing supported group differences validity.

Conclusion

This investigation found the TSQM to be a useful tool, exhibiting good psychometric measurement properties in patients with relapsing MS in the TENERE study.

Introduction

Patient satisfaction with medication, resulting from factors such as the effectiveness, convenience (e.g. route of administration, dosing frequency), or side effects of the medication, is associated with better adherence to, and persistence with, treatment.[1, 2] These findings, consistent across many diseases and clinical settings,[2] highlight the ongoing need to evaluate and improve patients' treatment experience.

Many scales have been used to measure treatment satisfaction. Frequently, they are applied inconsistently and/or have not been evaluated in the specific disease setting being assessed.[1] In their roadmap to patient-focused outcome measurement in clinical trials (**Figure 1**), the US Food and Drug Administration highlight the importance of examining clinical outcome assessments (COAs) in their context of use.[3] This is because COA suitability, as a measure of the concept of interest, is dependent upon the context of use. There is, therefore, no such thing as a "validated instrument". EU and US guidelines recommend that, if a measurement instrument is applied in a new disease setting, it is confirmed as fit for purpose in that context.[4, 5]

The Treatment Satisfaction Questionnaire for Medication (TSQM) was designed as a general measure of treatment satisfaction with medication. An initial pool of 55 candidate items was developed from focus groups of a panel of 500 patients with chronic disease (migraine, arthritis, hypertension, asthma, diabetes, psoriasis, hypercholesterolemia, and depression) and refined to 31 test items. Via a multistep iterative process, these were reduced to 14 final items (**Supplementary Appendix 1**) covering the majority of the variance in the test population.[6]

The TSQM has been examined, using standard psychometric methods, in several settings,[6, 7] though not yet in MS. A study using data from 400 patients with cystic fibrosis treated with inhaled antibiotics concluded that the TSQM had good measurement properties in patients with this condition.[7] Using data from patients with various chronic diseases (see above) Atkinson and colleagues applied psychometric tests to examine the performance of the TSQM, and also concluded that it possessed good psychometric properties.[6] Some of their findings are noteworthy for patients with MS; they reported significant differences across the TSQM between different methods of treatment administration, with individuals using injectable therapies reporting low satisfaction and convenience.[6]

In patients with RMS, longer treatment duration has been linked with improved long-term outcomes,[8] so it is important to ensure patient treatment satisfaction in order to maximize persistence with treatment over the long term.[2] Teriflunomide, a once-daily oral immunomodulator approved for the treatment of relapsing-remitting MS, demonstrated consistent efficacy with a well-documented safety profile in randomized, placebo-controlled monotherapy studies in patients with relapsing forms of MS (RMS)[9-11] and in patients with a first clinical episode suggestive of MS.[12] The phase 3 TENERE study (NCT00883337) compared teriflunomide with subcutaneous interferon beta-1a (scIFNβ-1a) in patients with RMS, and included the 14-item TSQM to measure patient satisfaction with either intervention.[13] The TSQM has been used in many studies of patient satisfaction in MS (reviewed by Ting and colleagues[14]), but to our knowledge, its measurement performance has yet to be examined comprehensively in the MS context of use.

Here, we examine the performance of the TSQM in patients with RMS using traditional psychometric methods to determine its fitness for purpose in the TENERE sample of patients with RMS.

Methods

Study design and participants

Details of the TENERE study are published elsewhere.[13] Briefly, patients aged 18 years or older with a diagnosis of RMS, an Expanded Disability Status Scale (EDSS) score of ≤5.5, and no relapse(s) within the prior 30 days were randomized (1:1:1) to receive once-daily teriflunomide 14 mg or 7 mg, or scIFNβ-1a 44 µg thrice weekly. The study was designed to end 48 weeks after the last patient was randomized.[13] Patient satisfaction with treatment was assessed using the TSQM version 1.4.[13]

TSQM structure

The TSQM (version 1.4) comprises 14 items across 4 domains focusing on effectiveness (3 items), side effects (5 items), convenience (3 items), and global satisfaction (3 items) of the medication over the previous 2–3 weeks, or since the patient's last use.[6] With the exception of item 4 (presence of side effects; yes or no), all items have 5 or 7 responses, scored from 1 (least satisfied) to 5 or 7 (most satisfied). The 7-item scales had a non-neutral midpoint, such that there were more positive response options than negative response options, to allow for precise information to be obtained at the upper end of the score distribution. Item scores are summed to give 4 domain scores, which are in turn transformed to a scale of 0–100. Item 4 was not included for scoring. If an item score is missing and half of the items in the domain are complete, domain scores may be imputed from the person-specific mean score of completed items.[15]

TSQM administration

The TSQM was administered every 12 weeks from Week 12 to Week 48, and every 24 weeks thereafter up to Week 96.[13] The TSQM was administered in patients' local languages, usingtranslations of the original questionnaire certified by translation agencies as linguistically equivalent (Supplementary Appendix 2).

TSQM evaluation

Five TSQM measurement properties were evaluated using TENERE data. Week 48 data were used unless otherwise indicated, as Week 48 was the timepoint used for the primary analysis of TENERE.[13]

Data completeness

To assess the extent to which the TSQM could be used successfully in TENERE (i.e. how acceptable the questionnaire is to test subjects), we computed item-level missing data for randomized patients, and the proportions of patients for whom domain scores could be computed. Fewer missing data indicate greater acceptability.[16]

Scaling assumptions

We assessed the legitimacy of summing TSQM item scores from TENERE, without weighting or standardization, to generate domain scores. Summing is considered legitimate when items of a domain are broadly parallel and contribute similarly to the construct being measured. These requirements are considered satisfied when items have similar means and variances[17], and item-to-domain score correlations, corrected for overlap, exceed 0.30.[18]

Scale-to-sample targeting

To examine the match between the potential range measured by the TSQM and the observed range measured in TENERE, we examined domain score distributions to ascertain the extent to which these met the recommended criteria of: spanning the available scale range,[19] mean scores located near the scale mid-point,[20] not being excessively skewed (skewness <1.0),[16] and floor and ceiling effects (proportions of patients with minimum and maximum scores, respectively) <20%.[21]

Reliability

Multiple reliability indicators are available to evaluate the extent to which scale scores are free from random error. We examined internal consistency (corrected item-total correlations, Cronbach's α, and homogeneity coefficients [mean item-item correlations for each domain]), test-retest reproducibility (agreement between scores at separate time points), and standard errors of measurement. Reliability is considered adequate for group

comparisons when corrected item-total correlations are >0.30,[22] Cronbach's α >0.80,[23] and homogeneity coefficients >0.30.[20]

The relatively long measurement interval in TENERE (≥12 weeks) could allow change over time to confound interpretation of test-retest estimates. Therefore, a conservative estimate of test-retest reproducibility was approximated by comparing TSQM values at Weeks 24 and 48 for patients with stable disease, defined as patients without relapses for the duration of treatment. A random effects model intra-class correlation coefficient was calculated using values generated by a repeated measures ANOVA, and a score >0.80 was considered acceptable.[23]

Standard errors of measurement, computed as standard deviation*√[1-reliability coefficient]) were used to interpret reliability estimates as confidence intervals (CIs) around scores (95% CI = score ±1.96*standard error of measurement), using Cronbach's α as the reliability coefficient. Low standard errors of measurement demonstrate low measurement error.[24]

Validity

To assess the extent to which the TSQM measures the constructs it purports to measure, we first tested internal construct validity (the extent to which items of the TSQM are grouped correctly into domains) as a prerequisite for interpretation of external construct validity tests (which provide more direct information on the constructs measured). Three examinations of internal construct validity were undertaken. Item-level convergent and discriminant validity was tested by computing scaling success rates. A definite scaling success was scored when an item's correlation with its own domain (corrected for overlap) was significantly higher (>2*standard error) than its correlations with another domain. Exploratory factor analysis [EFA], performed as a maximum likelihood factor analysis, was used to identify factors that explain the maximum amount of variance. Confirmatory factor analysis [CFA] was performed as an hypothesis-driven approach to further understand shared variance between variables due to factors. Goodness-of-fit indices were assessed against predefined criteria for good fit: Root Mean Square Error of Approximation <0.08, Normed Fit Index >0.9, Goodness of Fit Index >0.9, Adjusted Goodness of Fit Index >0.9, and standardized Root Mean Square Residual <0.05.

Two examinations of external construct validity of the TSQM were undertaken. Firstly, scale-level convergent and discriminant construct validity was tested by examining the extent to which the direction, magnitude, and pattern of correlations between variables were consistent with expectation. We examined correlations between TSQM domains and baseline patient characteristics (age, gender, EDSS, and Fatigue Impact Scale [FIS] scores), hypothesizing that that these correlations would be lower than the TSQM between-domain correlations. Secondly, group differences construct validity was tested using score differences between responders and non-responders on a range of clinical outcomes. The outcomes were selected based on measured parameters that we hypothesized would be likely to explain a clinical difference, and are detailed in **Supplementary Table 1**. Group mean score differences were expressed in terms of statistical (*p* value from independent samples analysis of variance [ANOVA]) and clinical significance (Cohen's d; effect size, ES). ES was interpreted using Cohen's criteria: ≥0.2–<0.5 for a small difference; ≥0.5–≤0.8 for a moderate difference; and >0.8 for a large difference.[25]

Results

Study participants

Patient characteristics in the TENERE study (**Table 1**) [13] were generally similar to those of patients in other Phase 3 studies of teriflunomide[9, 10] and other oral disease-modifying treatments for RMS,[26-29] albeit with a slightly lower mean EDSS score at baseline in TENERE.

Data completeness

TSQM data completeness in TENERE was good. Each item was missing a response in fewer than 2% of patients (*n*=324; range: 0.3–1.9%; **Table 2**). Domain scores could be computed for all participants (**Supplementary Table 2**).

Scaling assumptions

Scaling assumptions were satisfied for all 4 domains. Item mean scores and variances were similar (**Supplementary Table 2**) and all item total correlations (corrected for overlap) exceeded 0.30 (**Table 2**). This supports, for each domain, the summing of item scores to generate domain scores without standardization or weighting.

Scale-to-sample targeting

For all domains except Effectiveness, scores did not span the whole scale range, demonstrating skewing towards high scores (**Table 2**). Mean and median scores exceeded the scale midpoint (50). High mean scores accompanied by ceiling effects (defined as maximum scores in >20% of patients) were particularly marked for Side Effects (mean score 90.1, 72% of patients with maximum score), and Convenience (mean score 82.2, 38% of patients with maximum score). Both domains had notable higher ceiling effects with oral treatment (teriflunomide) than with injectable treatment (scIFNβ-1a). There were no notable floor effects, with small percentages of patients with minimum scores (minimal satisfaction) in each domain. Together, these high scores suggest good overall treatment satisfaction that was generally higher with teriflunomide than with scIFNβ-1a.[13, 30]

Reliability

Internal consistency reliability was high for all domains, with Cronbach's α >0.90 and all homogeneity coefficients >0.75; corresponding standard errors of measurement were thus relatively small. Test-retest reproducibility coefficients exceeded 0.70 for three domains (Side Effects, Convenience, Global Satisfaction) indicating adequate reproducibility given that these were likely conservative etimates.[23] The coefficient for Effectiveness was low (0.44).

Validity

Tests of internal construct validity supported the proposed item groupings. Definite scaling success rates for all 4 domains were 100% (**Table 2**). EFA grouped the 13 scoring items into 4 factors with item content equivalent to the 4 TSQM domains (**Supplementary Table 3**). CFA (**Figure 2**) also supported TSQM item groupings; at Week 48, Goodness of Fit indices met the predefined criteria. The largest contribution to Global Satisfaction came from Effectiveness (standardized estimate for association was 0.63), followed by Convenience (0.54) and Side Effects (0.32).

Tests of external construct validity supported the constructs measured by the domains. Correlations among TSQM domains were consistent with expectation, and supported the 4 domains as measures of related but different constructs (**Supplementary Table 2**). As in the CFA, perceived effectiveness was linked with Global Satisfaction (correlation coefficient,

0.69). Correlations between TSQM domains and age, gender, EDSS, and FIS were low (ranging from 0.01 to -0.31), indicating treatment satisfaction was not biased by these variables (**Supplementary Table 2**).

As hypothesized, there was a statistically significant ($p \leq 0.05$) and clinically meaningful (ES >0.3) relationship between each TSQM domain and the clinical outcomes tested (**Table 3**). For example, the minimal number of patients with adverse events leading to treatment discontinuation had a statistically ($p<0.0001$) and clinically (ES, 3.24) significantly (reduced Side Effects domain score (31.3; $n=2$) compared with patients who did not (90.6; $n=243$). There were also highly statistically significant ($p<0.0001$) relationships between the Convenience domain and relevant clinical outcomes. Treatment received (teriflunomide/scIFNβ-1a, used as a proxy for mode of administration) showed the strongest relationship (ES=1.74) with Convenience.

Discussion

This analysis provided a comprehensive evaluation, using traditional psychometric methods, of the extent to which the 14-item version of the TSQM is a fit-for-purpose measure of treatment satisfaction in the TENERE study of patients with RMS. Overall, we found that the TSQM exhibits good measurement properties and met the requirements of traditional psychometric tests. Specifically, we found that item scores could be summed without weighting or standardization to form total scores that were reliable, and for which evidence supported their validity as measures of different aspects of treatment satisfaction.

Analysis of scale-to-sample targeting identified a potential limitation of the TSQM for the relapsing MS context of use. Marked ceiling effects for the Side Effects and Convenience domains were observed in the teriflunomide-treated group. This may be a reflection of high levels of patient satisfaction with teriflunomide treatment, which is supported by the significant and clinically meaningful improvement in TSQM score for the teriflunomide 14-mg group versus the scIFNβ-1a 44-µg group on the Side Effects and Convenience domains in TENERE.[13, 30] Preliminary results from the Teri-PRO (**Teri**flunomide **P**atient-**R**eported **O**utcomes; NCT01895335) study of real-world teriflunomide use also indicate that patient satisfaction, as measured by the TSQM, increases when patients switch their disease-modifying therapy to teriflunomide.[31] Furthermore, an analysis of the TSQM in patients with chronic diseases found that injectable modes of administration were associated with lower

TSQM scores, which could again suggest that scores for teriflunomide-treated patients are expected to be higher than those of patients treated with scIFNβ-1a.[6] The skewed mean scores and high ceiling effects we observed may indicate that the TSQM limited the possible measurement of satisfaction in these patients, with the "true" satisfaction of the teriflunomide-treatment group likely to be higher than that actually measured; the differences between scIFNβ-1a and teriflunomide may, therefore, be larger than measured.

In this analysis, internal consistency indicators (Cronbach's α and homogeneity coefficients) were very high, particularly given the small numbers of items in each domain. This implies the items in each domain were closely related and may suggest possible item redundancy.[32] However, indicators of internal consistency may also be elevated spuriously by ceiling or floor effects, and we have noted skewed score distributions in our analysis. Reanalysis of reliability could help to determine if there is true item redundancy. Though traditional psychometric methods are widely used, they do have recognized limitations.[33] In this instance, reliability analyses using the person separation index generated by the more modern Rasch measurement theory analysis,[34] might be informative.

Though the intervals between TSQM data collection were too long to permit a robust evaluation of test-retest reproducibility, our conservative approximations implied that high reproducibility is to be expected for 3 domains (Global Satisfaction, Convenience, Side Effects). It is difficult to know how best to interpret the value of 0.44 for Effectiveness, and this merits further investigation.

CFA implied that Global Satisfaction with treatment within the TENERE study population was driven primarily by Effectiveness, followed by Convenience and Side Effects. This is consistent with studies of treatment adherence in patients with MS, which have identified treatment efficacy as important and lack of efficacy as a key reason for treatment discontinuation,[35, 36] and also with findings in other diseases, which showed Global Satisfaction was most strongly linked with Effectiveness.[6] It would be of interest to explore how relapses and disability progression are linked with changes in TSQM, and if these clinical changes in turn affect its measurement properties.

Though the patient-unblinded nature of TENERE may have influenced patient satisfaction ratings,[13] we do not expect it to influence the empirical measurement performance of the TSQM, as analyzed in this study.

An important next step would be to examine the item content of the TSQM, to optimize it for the RMS patient population. Qualitative research might identify new items that extend the measurement range of the TSQM, reduce ceiling effects, and advance measurement of treatment satisfaction in patients with RMS.

To our knowledge, this is the first time the performance of the TSQM has been evaluated in a sample of patients with RMS. While, as noted, evaluation in a single study population does not confirm measurement performance in all contexts, our comprehensive analysis supports the TSQM as a fit-for purpose measure of treatment satisfaction in TENERE. Based on this, it seems reasonable to conclude that TSQM is likely to be appropriate for use in studies of disease-modifying therapies for patients with RMS. Indeed, the tool is being used as an outcome measure to provide further understanding of patient experiences of teriflunomide treatment in routine clinical practice in ongoing phase 4 studies,[31] and it is our intention to use data from such studies to perform a follow-on evaluation of TSQM performance in the context of use of real-world patients with RMS. However, as with all instruments, detailed analysis demonstrates room for improvement. Here, the suboptimal scale-to-sample targeting implies that treatment satisfaction maybe underestimated by the TSQM in this context of use, and modification of the TSQM may overcome this limitation.

Acknowledgments

This manuscript was reviewed by Larisa Miller, PharmD, of Sanofi Genzyme. Editorial support was provided by Victoria Lawson, of Fishawack Communications, and was funded by Sanofi Genzyme.

Declaration of conflicting interests

Sanofi Genzyme was the sponsor.

PV: Honoraria, consulting fees (Almirall, Bayer, Biogen Idec, GSK, Merck Serono, Novartis, Sanofi Genzyme, Teva); research support (Bayer, Biogen Idec, Merck Serono, Sanofi Genzyme,). JH: Honoraria, consulting fees (Acorda, Biogen Idec, Critical Path Institute, LORA group, MAPI Research Institute, Sanofi Genzyme); license fee payments or royalty payments (Plymouth University receives fees for the use of rating scales developed as part of author's research); research support (Biogen Idec, Novartis, Merck Serono). CD-P: Employee of Sanofi

Genzyme at time of data analysis. SB and SH: Employees of Sanofi Genzyme. PKC: Consulting fees (AbbVie, Accordant, Acorda, Bayer, Biogen, Genentech/Roche, Genzyme/Sanofi, Mallinckrodt, Novartis, Serono, Teva); research support (Actelion, Genentech/Roche, Novartis, Opexa).

References

1. Barbosa CD, Balp MM, Kulich K, Germain N and Rofail D. A literature review to explore the link between treatment satisfaction and adherence, compliance, and persistence. *Patient Prefer Adherence*. 2012; 6: 39-48.

2. Atkinson MJ, Kumar R, Cappelleri JC and Hass SL. Hierarchical construct validity of the treatment satisfaction questionnaire for medication (TSQM version II) among outpatient pharmacy consumers. *Value Health*. 2005; 8 Suppl 1: S9-S24.

3. FDA. Roadmap to PATIENT-FOCUSED OUTCOME MEASUREMENT in Clinical Trials. 2013.

4. Chassany O, P. S, Marquis P, Fullerton S and Aaronson N. Patient-reported outcomes: the examples of health-related quality of life – A European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. *Drug Inf J*. 2002; 36: 209–38.

5. FDA. Qualification of CLINICAL OUTCOME ASSESSMENTS 2015.

6. Atkinson MJ, Sinha A, Hass SL, et al. Validation of a general measure of treatment satisfaction, the Treatment Satisfaction Questionnaire for Medication (TSQM), using a national panel study of chronic disease. *Health Qual Life Outcomes*. 2004; 2: 12.

7. Regnault A, Balp MM, Kulich K and Viala-Danten M. Validation of the Treatment Satisfaction Questionnaire for Medication in patients with cystic fibrosis. *J Cyst Fibros*. 2012; 11: 494-501.

8. Kappos L, Kuhle J, Multanen J, et al. Factors influencing long-term outcomes in relapsing-remitting multiple sclerosis: PRISMS-15. *Journal of neurology, neurosurgery, and psychiatry*. 2015; 86: 1202-7.

9. Confavreux C, O'Connor P, Comi G, et al. Oral teriflunomide for patients with relapsing multiple sclerosis (TOWER): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet Neurology*. 2014; 13: 247-56.

10. O'Connor P, Wolinsky JS, Confavreux C, et al. Randomized trial of oral teriflunomide for relapsing multiple sclerosis. *The New England journal of medicine*. 2011; 365: 1293-303.

11. O'Connor PW, Li D, Freedman MS, et al. A Phase II study of the safety and efficacy of teriflunomide in multiple sclerosis with relapses. *Neurology*. 2006; 66: 894-900.

12. Miller AE, Wolinsky JS, Kappos L, et al. Oral teriflunomide for patients with a first clinical episode suggestive of multiple sclerosis (TOPIC): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet Neurology*. 2014; 13: 977-86.

13. Vermersch P, Czlonkowska A, Grimaldi LM, et al. Teriflunomide versus subcutaneous interferon beta-1a in patients with relapsing multiple sclerosis: a randomised, controlled phase 3 trial. *Multiple sclerosis*. 2014; 20: 705-16.

14. Ting J, Liu Y, Petrillo J, Giannattasio G and Sabatella G. Treatment Satisfaction With Disease Modifying Therapies In Multiple Sclerosis: A Systematic Review of Studies Using The Treatment Satisfaction Questionnaire For Medication (Tsqm). *Value Health*. 2015; 18: A760-1.

15. Stewart A and Ware JE, Jr. Chapter 5. *Measuring functioning and well-being: the Medical Outcomes Study approach*. Durham, NC: Duke University Press, 1992.

16. McHorney CA, Ware JE, Jr., Lu JF and Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical care*. 1994; 32: 40-66.

17. Likert R. A technique for the development of attitudes. *Arch Psychol*. 1932; 140: 5-55.

18. Howard K and Forehand G. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and psychological measurement*. 1962; 22: 731-5.

19. Ware JE, Jr., Brook RH, Davies-Avery A, Williams K, Stewart A and Rogers W. Model of health and methodology. *Conceptualization and measurement of health for adults in the health insurance study*. 1980.

20. Eisen M, Ware JE, Jr., Donald CA and Brook RH. Measuring components of children's health status. *Medical care*. 1979; 17: 902-21.

21. Hobart JC, Lamping DL, Freeman JA, et al. Evidence-based measurement: which disability scale for neurologic rehabilitation? *Neurology*. 2001; 57: 639-44.

22. Ware JE, Jr., Harris W, Gandek B, Rogers B and Reese P. Chapter 2. *MAP-R for Windows: multitrait/multi-item analysis program - revised user's guide*. Boston, MA: Heath Assessment Laboratory, 1997.

23. Nunnally J and Bernstein I. Chapter 7. *Psychometric theory*. 3 ed. New York: McGraw-Hill, 1994.

24. Nunnally J. Chapter 5. *Introduction to psychological measurement*. New York: McGraw-Hill, 1979.

25. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edn ed.: Lawrence Erlbaum Associates, 1988, p.273-406.

26. Kappos L, Radue EW, O'Connor P, et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *The New England journal of medicine*. 2010; 362: 387-401.

27. Calabresi PA, Radue EW, Goodin D, et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *The Lancet Neurology*. 2014; 13: 545-56.

28. Gold R, Kappos L, Arnold DL, et al. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *The New England journal of medicine*. 2012; 367: 1098-107.

29. Fox RJ, Miller DH, Phillips JT, et al. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *The New England journal of medicine*. 2012; 367: 1087-97.

30. Mäurer MVW, Bart , de Seze J, Meca-Lallana J, Bozzi S and Vermersch P. Significant and Meaningful Improvement in Treatment Satisfaction With Teriflunomide vs Subcutaneous IFNβ-1a in Patients With Relapsing MS: Results From TENERE [PND73]. *ISPOR*. Amsterdam, The Netherlands2014.

31. Coyle PK, LaGanke C, Khatri B, et al. Improvements in Patient-Reported Outcomes With Teriflunomide: Week 24 Interim Results From the US Cohort of the Teri-PRO Phase 4 Study. *ECTRIMS*. 2015: P562.

32. Panayides P. Coefficient Alpha - Interpret With Caution. *Europe's Journal of Psychology*. 2013; 9: 687-96.

33. Hobart J and Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*. 2009; 13: iii, ix-x, 1-177.

34. Andrich D. An Index of Person Separation in Latent Trait Theory, the Traditional KR.20 Index, and the Guttman Scale Response Pattern. *Educational and Psychological Research*. 1982; 9: 95-104.

35. Rio J, Porcel J, Tellez N, et al. Factors related with treatment adherence to interferon beta and glatiramer acetate therapy in multiple sclerosis. *Multiple sclerosis*. 2005; 11: 306-9.

36. Mohr DC, Goodkin DE, Likosky W, et al. Therapeutic expectations of patients with multiple sclerosis upon initiating interferon beta-1b: relationship to adherence to treatment. *Multiple sclerosis*. 1996; 2: 222-6.

37. Randomised double-blind placebo-controlled study of interferon beta-1a in relapsing/remitting multiple sclerosis. PRISMS (Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis) Study Group. *Lancet*. 1998; 352: 1498-504.

38. Doyle C, Lennox L and Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ open*. 2013; 3.

Tables and Figures

Table 1. Baseline demographics and disease characteristics.

	sc IFN β-1a (*n*=104)	**Teriflunomide 7 mg (*n*=109)**	**Teriflunomide 14 mg (*n*=111)**
Age, years¸ mean (SD)	37.0 (10.6)	35.2 (9.2)	36.8 (10.3)
Female, n (%)	71 (68.3)	70 (64.2)	78 (70.3)
Caucasian, n (%)	104 (100)	109 (100)	111 (100)
Time since first symptoms of MS, years, mean (SD)	7.7 (7.6)	7.0 (6.9)	6.6 (7.6)
No. of relapses within previous year, mean (SD)	1.2 (1.0)	1.3 (0.8)	1.4 (0.8)
Relapsing–remitting MS, n (%)	104 (100)	109 (100)	108 (97.3)[a]
Use of DMT in previous 2 years, n (%)	25 (24.0)	23 (21.1)	13 (11.7)
Baseline EDSS score, mean (SD)	2.0 (1.2)	2.0 (1.2)	2.3 (1.4)
Baseline FIS score, mean (SD)	34.2 (32.7)	39.5 (34.8)	42.5 (37.8)

Randomized population, *n*=324.

[a]Secondary progressive MS, *n*=1; progressive relapsing MS, *n*=2.

DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; FIS: Fatigue Impact Scale; IFN: interferon; sc: subcutaneous; SD: standard deviation.

Table 2. Item-level analyses of TSQM.

Domain	Item		Response categories, *n*	Patients with missing data, *n* (%)[a]	Correlation with domain[b,c]				Scaling success rate, %[d]
					Effect-iveness	Side Effects	Conven-ience	Global Satis-faction	
Effect-iveness	Q1	Satisfaction with prevention/treatment	7	3 (0.9)	**0.90**	0.19	0.27	0.54	100
	Q2	Satisfaction with symptom relief	7	5 (1.5)	**0.88**	0.24	0.27	0.56	
	Q3	Satisfaction with time to start working	7	5 (1.5)	**0.89**	0.21	0.31	0.56	
Side Effects	Q4	Side effect presence[e]	2	6 (1.9)	NA	NA	NA	NA	100
	Q5	Bother from side effects	5	4 (1.2)	0.19	**0.76**	0.44	0.26	
	Q6	Side effects interference with physical function	5	4 (1.2)	0.25	**0.83**	0.46	0.26	
	Q7	Side effects interference with mental function	5	2 (0.6)	0.23	**0.66**	0.42	0.29	
	Q8	Impact of side effects on satisfaction	5	2 (0.6)	0.23	**0.71**	0.50	0.34	
Conven-ience	Q9	Treatment easy to use	7	1 (0.3)	0.24	0.44	**0.83**	0.39	100
	Q10	Easy planning of use	7	1 (0.3)	0.23	0.42	**0.82**	0.41	
	Q11	Intake convenience	7	2 (0.6)	0.36	0.41	**0.82**	0.52	
Global Satis-faction	Q12	Confidence in benefits	5	2 (0.6)	0.51	0.19	0.36	**0.81**	100
	Q13	Balance between good and bad things	5	2 (0.6)	0.49	0.23	0.38	**0.83**	
	Q14	Global satisfaction	7	3 (0.9)	0.59	0.33	0.55	**0.80**	

[a]Randomized population, *n*=324; [b]Patients from intent-to-treat population with complete TSQM domain information at Week 48, *n*=243–246; [c]Item-own domain correlations corrected for item overlap (bold); [d]Percentage of correlations where item-own domain correlation (corrected for overlap) exceeds item –other domain correlation by more than 2*SE (where SE = 1/√n); [e]Dichotomous item, not scored.

NA: not applicable; SE: standard error; TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4).[6]

Table 3. Relationships between clinical outcomes and TSQM domains at Week 48.

Domain	Clinical outcome	Patients with outcome		Patients without outcome		Effect size, Cohen's *d*	*p* value[a]
		n	Score, mean (SD)	*n*	Score, mean (SD)		
Effectiveness	Treatment failure[b]	51	61.2 (19.5)	192	68.8 (22.4)	0.35	0.028
	Confirmed relapse	50	61.6 (19.6)	193	68.7 (22.4)	0.33	0.041
Side Effects	AEs leading to treatment discontinuation	2	31.3 (17.7)	243	90.6 (18.3)	3.24	0.020
	Nervous system disorders	92	86.1 (21.5)	153	92.6 (17.0)	0.38	0.009
	General disorders or administration-site conditions[c]	73	82.0 (24.6)	172	93.6 (15.0)	0.63	<0.0001
Convenience	Treated with sc IFN β-1a[d]	74	63.2 (19.1)	176	89.8 (13.4)	1.74	<0.0001
	General disorders or administration-site conditions[c]	74	74.0 (22.1)	172	85.8 (17.0)	0.63	<0.0001
Global Satisfaction	Treatment failure[b]	52	63.2 (21.2)	193	72.2 (20.7)	0.43	0.006
	Confirmed relapse	51	63.6 (21.3)	194	72.1 (20.8)	0.41	0.011

Table shows all relationships with $p<0.05$ for patients with complete TSQM domain information at Week 48; [a]*P* value from ANOVA; [b]Confirmed relapse or permanent treatment discontinuation for any reason; [c]General disorders and administration-site conditions were mainly driven by influenza-like illness;

[d]Specific outcomes for convenience are difficult to identify in a randomized controlled trial, and we observed a relationship with AEs related to mode of administration (injectable sc IFN β-1a vs oral teriflunomide) using treatment received as a proxy.

AE: adverse event; ANOVA: analysis of variance; IFN: interferon; sc: subcutaneous; SD: standard deviation; TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4). [6]

Supplementary Table 1. Testing performed on relationships between clinical outcomes and TSQM domains.

Domain	Clinical outcome	Rationale for choice	Hypothesis tested
Effectiveness and **Global Satisfaction**	**Treatment failure** Confirmed relapse or permanent treatment discontinuation for any reason	Primary study endpoint	As clinical efficacy has been demonstrated for both teriflunomide and IFN β,[9, 10, 37] as both Effectiveness and Global Satisfaction improve following initiation of teriflunomide treatment,[31] and as the clinical effectiveness of a treatment has been linked to treatment satisfaction,[38] we hypothesize that these sets of measures would be linked
	Confirmed relapse	A commonly used efficacy measure in studies of DMTs in RMS (other efficacy measures such as disability or MRI outcomes were not recorded in TENERE)	
Side Effects	**AEs leading to treatment discontinuation**	To be representative of the relationship between AEs and treatment satisfaction	Since tolerability is linked with patient treatment satisfaction,[38] we would expect to see a relationship between the Side Effects domain, and these AE parameters as a clinical outcome
	Nervous system disorders **General disorders or administration-site conditions**	The AEs with the highest incidence in this study	
Convenience	**Treated with sc IFN β-1a** Proxy for mode of administration (injection vs oral) **General disorders or administration-site conditions**	Convenience has been shown to be linked to mode of administration,[6] and specific outcomes for convenience are hard to identify in a randomized controlled trial	We hypothesize that the improved convenience with teriflunomide vs IFN β seen in TENERE,[30] may be explained by the differing modes of administration

AE: adverse event; DMT: disease-modifying therapy; IFN: interferon; RMS: relapsing forms of MS; sc: subcutaneous; TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4).

Supplementary Table 2. Domain-level analyses of TSQM.

Variable	TSQM domain			
	Effectiveness	**Side Effects**	**Convenience**	**Global Satisfaction**
k	3	5	3	3
n	243	245	246	245
Scaling assumptions				
Item mean range	4.95–5.10	4.49–4.67	5.83–6.00	3.69–5.43
Item SD range	1.34–1.43	0.69–0.94	1.16–1.33	1.01–1.19
Item total correlation range[a]	0.74–0.96	0.69–0.80	0.72–0.83	0.75–0.86
Scale-to-sample targeting[b]				
Mean (SD)	67.2 (22.0)	90.1 (19.1)	82.2 (19.4)	70.3 (21.1)
Median (range)	66.7 (0–100)	100 (18.8–100)	83.3 (11.1–100)	71.4 (14.3–100)
Floor effects, *n*/*N* (%)				
All patients	4/243 (1.65)	4/245 (1.63)	1/246 (0.41)	1/245 (0.41)
sc IFN β-1a 44 µg	1/69 (1.45)	4/69 (5.80)	1/70 (1.43)	1/70 (1.43)
Teriflunomide[c]	3/174 (1.72)	0[d]/176	0[d]/176	0[d]/175
Ceiling effects, *n/N* (%)				
All patients	25/243 (10.3)	176/245 (71.8)	94/246 (38.2)	39/246 (15.9)
sc IFN β-1a 44 µg	5/69 (7.25)	26/69 (37.7)	4/70 (5.71)	22/175 (12.6)
Teriflunomide[c]	20/174 (11.5)	150/176 (85.2)	90/176 (51.1)	7/70 (10.0)

Reliability				
Cronbach's α	0.95	0.94	0.91	0.91
Homogeneity coefficient	0.82	0.76	0.77	0.79
Standard error of measurement	4.92	4.68	1.75	6.33
Test-retest reliability[e]	0.44	0.72	0.82	0.81
Validity				
Within-scale correlations				
Effectiveness	1	0.25	0.42	0.69
Side Effects		1	0.48	0.42
Convenience			1	0.46
Global Satisfaction				1
Correlations between TSQM domains[f] and other variables				
Age at baseline	-0.15	-0.05	0.01	-0.10
Gender[g]	-0.058	-0.157	0.007	-0.072
EDSS score at Week 12	-0.19	-0.05	-0.02	-0.15
FIS score at Week 12	-0.23	-0.31	-0.11	-0.17

Patients from intent-to-treat population with complete TSQM domain information at Week 48. [a]Corrected for item overlap; [b]Scores are transformed to give a range of 0–100; [c]14-mg and 7-mg dose groups combined; [d]No floor effect, for lowest score see range, above. Floor score (lowest patient satisfaction) = 0, ceiling score (greatest patient satisfaction) = 100; [e]Intra-class correlation coefficient between scores at Week 24 and Week 48 in patients without relapses; [f]At Week 12, Pearson correlations; [g]Biserial correlation.

EDSS: Expanded Disability Status Scale; FIS: Fatigue Impact Scale; IFN: interferon; sc: subcutaneous; SD: standard deviation; TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4).[6]

Supplementary Table 3. Exploratory factor analysis: rotated factor pattern by TSQM item.

TSQM item		Rotated factor pattern			
		Factor 1	**Factor 2**	**Factor 3**	**Factor 4**
Q1	Satisfaction with prevention/treatment	0.074	**0.907**	0.110	0.235
Q2	Satisfaction with symptom relief	0.133	**0.847**	0.084	0.280
Q3	Satisfaction with time to start working	0.097	**0.860**	0.144	0.259
Q5	Bother from side effects	**0.913**	0.061	0.186	0.090
Q6	Side effects interference with physical function	**0.918**	0.132	0.222	0.050
Q7	Side effects interference with mental function	**0.795**	0.107	0.189	0.103
Q8	Impact of side effects on satisfaction	**0.839**	0.084	0.258	0.158
Q9	Treatment easy to use	0.261	0.094	**0.844**	0.138
Q10	Easy planning of use	0.251	0.076	**0.824**	0.170
Q11	Intake convenience	0.219	0.189	**0.814**	0.252
Q12	Confidence in benefits	0.069	0.301	0.159	**0.795**
Q13	Balance between good and bad things	0.118	0.257	0.173	**0.836**
Q14	Global satisfaction	0.184	0.370	0.343	**0.705**
Factor coincides with TSQM domain		Side Effects	Effectiveness	Convenience	Global Satisfaction
Homogeneity coefficient		0.87	0.87	0.83	0.78

TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4).

Figure 1. FDA roadmap to patient-focused outcome measurement in clinical trials.[3].

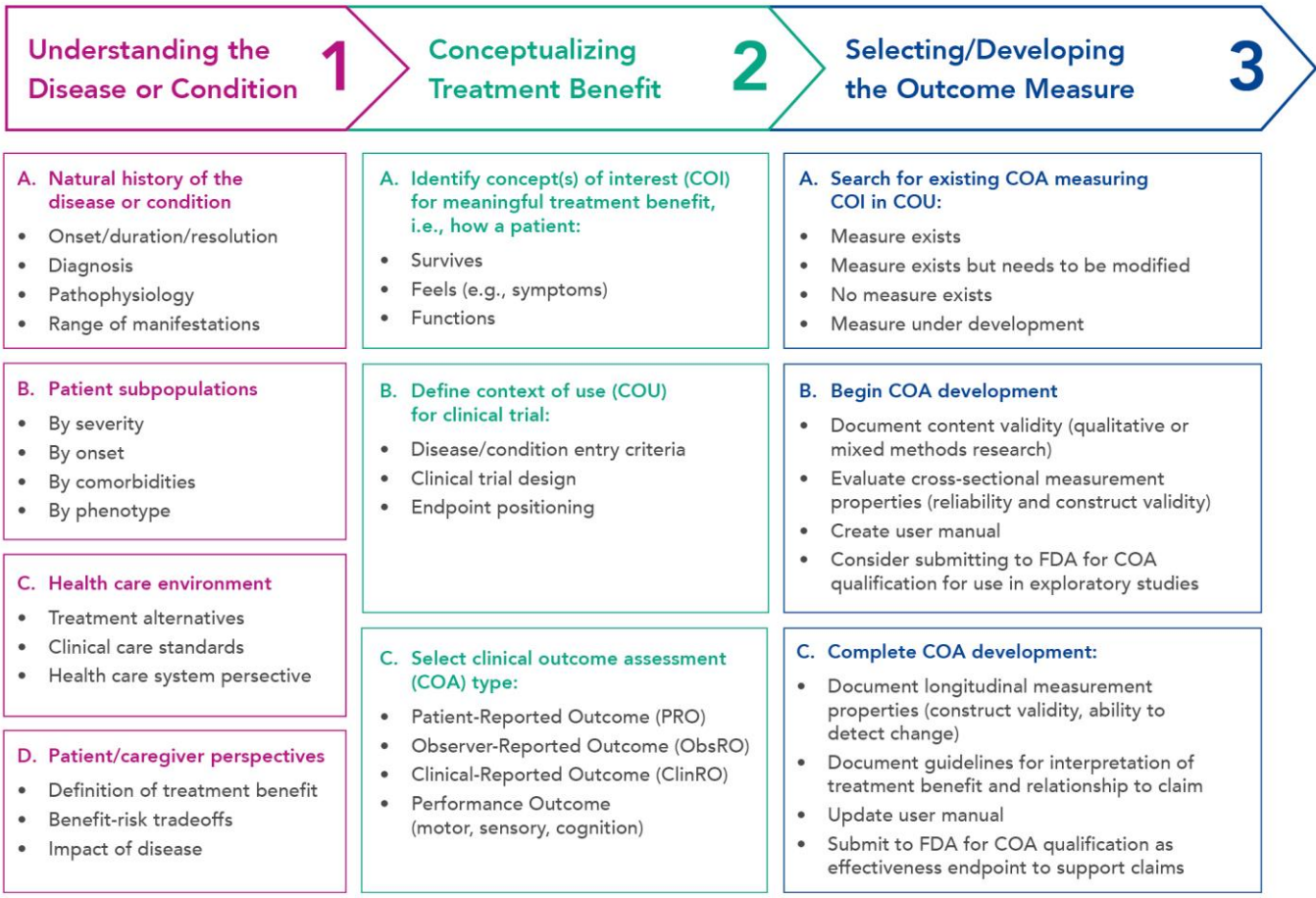

Figure 2. Confirmatory factor analysis of the TSQM.

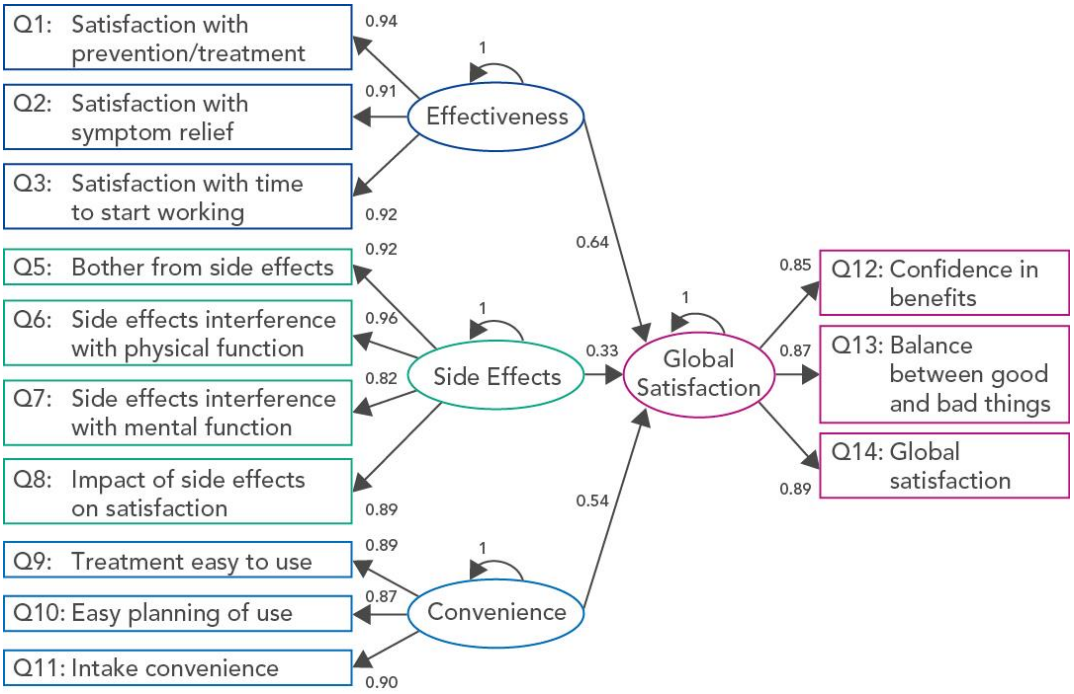

Figure Legends

Figure 1. FDA roadmap to patient-focused outcome measurement in clinical trials.[3]

Reproduced by permission of the US Food and Drug Administration.

Figure 2. Confirmatory factor analysis of the TSQM.

Ovoids represent unobserved variables (domains); rectangles represent observed variables (items); arrows represent the hypothesized links between the variables; parameters relative to each arrow are standardized estimates of the strength of association between the linked variables.

Root Mean Square Error of Approximation, 0.067; Normed Fit Index, 0.958; Goodness of Fit Index, 0.925; Adjusted Goodness of Fit Index, 0.884; Standardized Root Mean Square Residual, 0.044.

TSQM: Treatment Satisfaction Questionnaire for Medication (version 1.4).[6]

Supplemental Materials

Appendix 1

TSQM (Version 1.4)

Treatment Satisfaction Questionnaire for Medication

Instructions: Please take some time to think about your level of satisfaction or dissatisfaction with the medication you are taking in the clinical trial. We are interested in your evaluation of the effectiveness, side effects, and convenience of the medication over the last two to three weeks, or since you last used it. For each question, please place a single check mark next to the response that most closely corresponds to your own experience.

1. How satisfied or dissatisfied are you with the ability of the medication to prevent or treat your condition?

   ☐ 1 Extremely Dissatisfied
   ☐ 2 Very Dissatisfied
   ☐ 3 Dissatisfied
   ☐ 4 Somewhat Satisfied
   ☐ 5 Satisfied
   ☐ 6 Very Satisfied
   ☐ 7 Extremely Satisfied

2. How satisfied or dissatisfied are you with the way the medication relieves your symptoms?

   ☐ 1 Extremely Dissatisfied
   ☐ 2 Very Dissatisfied
   ☐ 3 Dissatisfied
   ☐ 4 Somewhat Satisfied
   ☐ 5 Satisfied
   ☐ 6 Very Satisfied
   ☐ 7 Extremely Satisfied

3. How satisfied or dissatisfied are you with the amount of time the medication takes to start working?

   ☐ 1 Extremely Dissatisfied
   ☐ 2 Very Dissatisfied
   ☐ 3 Dissatisfied
   ☐ 4 Somewhat Satisfied
   ☐ 5 Satisfied
   ☐ 6 Very Satisfied

☐ 7 Extremely Satisfied

4. As a result of taking this medication, do you experience any side effects at all?

☐ Yes
☐ No (if No, then please skip to Question 9)

5. How bothersome are the side effects of the medication you take to treat your condition?

☐ 1 Extremely Bothersome
☐ 2 Very Bothersome
☐ 3 Somewhat Bothersome
☐ 4 A Little Bothersome
☐ 5 Not at All Bothersome

6. To what extent do the side effects interfere with your physical health and ability to function
(i.e., strength, energy levels etc.)?

☐ 1 A Great Deal
☐ 2 Quite a Bit
☐ 3 Somewhat
☐ 4 Minimally
☐ 5 Not at All

7. To what extent do the side effects interfere with your mental function
(i.e., ability to think clearly, stay awake etc.)?

☐ 1 A Great Deal
☐ 2 Quite a Bit
☐ 3 Somewhat
☐ 4 Minimally
☐ 5 Not at All

8. To what degree have medication side effects affected your overall satisfaction with the medication?

☐ 1 A Great Deal
☐ 2 Quite a Bit
☐ 3 Somewhat
☐ 4 Minimally
☐ 5 Not at All

9. How easy or difficult is it to use the medication in its current form?

☐ 1 Extremely Difficult
☐ 2 Very Difficult
☐ 3 Difficult
☐ 4 Somewhat Difficult

- [ ] 5 Easy
- [ ] 6 Very Easy
- [ ] 7 Extremely Easy

10. How easy or difficult is it to plan when you will use the medication each time?

- [ ] 1 Extremely Difficult
- [ ] 2 Very Difficult
- [ ] 3 Difficult
- [ ] 4 Somewhat Difficult
- [ ] 5 Easy
- [ ] 6 Very Easy
- [ ] 7 Extremely Easy

11. How convenient or inconvenient is it to take the medication as instructed?

- [ ] 1 Extremely Inconvenient
- [ ] 2 Very Inconvenient
- [ ] 3 Inconvenient
- [ ] 4 Somewhat Convenient
- [ ] 5 Convenient
- [ ] 6 Very Convenient
- [ ] 7 Extremely Convenient

12. Overall, how confident are you that taking this medication is a good thing for you?

- [ ] 1 Not at All Confident
- [ ] 2 A Little Confident
- [ ] 3 Somewhat Confident
- [ ] 4 Very Confident
- [ ] 5 Extremely Confident

13. How certain are you that the good things about your medication outweigh the bad things?

- [ ] 1 Not at All Certain
- [ ] 2 A Little Certain
- [ ] 3 Somewhat Certain
- [ ] 4 Very Certain
- [ ] 5 Extremely Certain

14. Taking all things into account, how satisfied or dissatisfied are you with this medication?

- [ ] 1 Extremely Dissatisfied
- [ ] 2 Very Dissatisfied
- [ ] 3 Dissatisfied
- [ ] 4 Somewhat Satisfied
- [ ] 5 Satisfied
- [ ] 6 Very Satisfied
- [ ] 7 Extremely Satisfied

Appendix 2

Certified translations of the TSQM were provided as follows:

MAPI Institute

Arabic
English for the United States
French
Greek
Portugese

Oxford Outcomes Ltd

Afrikaans
Armenian
Austrian German
Belgian Dutch
Belgian French
Bulgarian
Canadian English
Canadian French
Chilean Spanish
Chinese for Malaysia
Chinese for Singapore
Chinese for Taiwan
Colombian Spanish
Croatian
Czech
Danish
Dutch
English for Australia and the UK
English for India
English for Malaysia
English for New Zealand
English for Singapore
English for South Africa
English for the Philippines
Finnish
French
German
Georgian
Hebrew
Hindi
Hungarian
Italian
Kannada
Korean
Malay
Malayalam
Marathi
Polish
Romanian
Russian
Slovakian
Slovenian
Spanish
Swedish
Swiss French
Swiss German
Tamil
Telugu
Thai
Turkish
Ukrainian