Identifying reprioritization response shift in a stroke caregiver population: a comparison of missing data methods

Tolulope T. Sajobi · Lisa M. Lix · Gurbakhshash Singh ·
Mark Lowerison · Jordan Engbers ·
Nancy E. Mayo

Abstract

Purpose Response shift (RS) is an important phenomenon that influences the assessment of longitudinal changes in health-related quality of life (HRQOL) studies. Given that RS effects are often small, missing data due to attrition or item non-response can contribute to failure to detect RS effects. Since missing data are often encountered in longitudinal HRQOL data, effective strategies to deal with missing data are important to consider. This study aims to compare different imputation methods on the detection of reprioritization RS in the HRQOL of caregivers of stroke survivors.

Methods Data were from a Canadian multi-center longitudinal study of caregivers of stroke survivors over a one-year period. The Stroke Impact Scale physical function score at baseline, with a cutoff of 75, was used to measure patient stroke severity for the reprioritization RS analysis. Mean imputation, likelihood-based expectation–maximization imputation, and multiple imputation methods were compared in test procedures based on changes in relative importance weights to detect RS in SF-36 domains over a 6-month period. Monte Carlo simulation methods were used to compare the statistical powers of relative importance test procedures for detecting RS in incomplete longitudinal data under different missing data mechanisms and imputation methods.

Results Of the 409 caregivers, 15.9 and 31.3 % of them had missing data at baseline and 6 months, respectively. There were no statistically significant changes in relative importance weights on any of the domains when complete-case analysis was adopted. But statistical significant changes were detected on physical functioning and/or vitality domains when mean imputation or EM imputation was adopted. There were also statistically significant changes in relative importance weights for physical functioning, mental health, and vitality domains when multiple imputation method was adopted. Our simulations revealed that relative importance test procedures were least powerful under complete-case analysis method and most powerful when a mean imputation or multiple imputation method was adopted for missing data, regardless of the missing data mechanism and proportion of missing data.

Conclusions Test procedures based on relative importance measures are sensitive to the type and amount of missing data and imputation method. Relative importance test procedures based on mean imputation and multiple imputation are recommended for detecting RS in incomplete data.

Keywords Missing data · Response shift · Relative importance analysis · Health-related quality of life

T. T. Sajobi (✉) · G. Singh
Department of Community Health Sciences, University of Calgary, 3280 Hospital Drive NW, Calgary T2N 4Z6, Canada
e-mail: tolu.sajobi@ucalgary.ca

T. T. Sajobi
Department of Clinical Neurosciences, University of Calgary, Calgary, Canada

T. T. Sajobi · M. Lowerison · J. Engbers
Clinical Research Unit, University of Calgary, Calgary, Canada

L. M. Lix
Department of Community Health Sciences, University of Manitoba, Winnipeg, Canada

N. E. Mayo
Division of Clinical Epidemiology, McGill University Health Center Research Institute, Montreal, Canada

Introduction

Response shift (RS), a change in an individual's internal standards, values, or conceptualizations of health and well-being, has been acknowledged as an important phenomenon that influences the assessment of longitudinal changes in health-related quality of life (HRQOL) studies [40, 41]. While RS is increasingly being investigated in longitudinal studies that adopted HRQOL instruments, this field of research has witnessed an increased number of advanced statistical methods such as structural equation modeling [30, 31], latent trajectory model [28]; classification and regression trees [3, 19, 20], and discriminant analysis and logistic regression analysis [23, 42] for RS detection in patient-reported HRQOL data.

Although these statistical procedures rest on different underlying model assumptions, they are all sensitive to the data analytic conditions that are inherent in longitudinal HRQOL data such as non-normality, missing data, and measurement error. Missing values arise because study participants drop out or are absent at intermittent occasions throughout the study [12, 22, 29] and because of questionnaire item non-response. Ignoring missing observations can lead to incorrect inferences about the size of a treatment effect or the magnitude of longitudinal change. Missing data may also result in incorrect conclusions about the presence of RS in longitudinal data.

Although several statistical methods have been developed to handle missing data in HRQOL studies [2, 29], the choice of a method is generally dependent on the pattern of missingness. Little and Rubin [22] summarized these patterns into three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR arises when the reason for the missingness is not related to any observed covariate or the true unknown values of the missing data. For the MAR mechanism, missing observations are unrelated to the true values of missing data but are associated with one or more *observed* covariates. Both MCAR and MAR are considered ignorable missing data mechanisms because the distribution of the missing observations is independent of the unobserved data, whereas MNAR is a non-ignorable pattern because missingness is related to *unobserved* underlying values of the missing data.

Missing data methods include conventional approaches such as complete-case analysis, last-observation-carried-forward, hot-deck imputation, pair-wise deletion, average-available-observations-carried-forward, and mean imputation [9, 18]. Although complete-case analysis has been shown to be valid when the missing observations are MCAR, such analyses may not be valid when the mechanism is MAR or MNAR, resulting in reduced statistical power and biased parameter estimates. However, all of these conventional methods for missing data have been shown to induce bias in the parameter estimates from longitudinal models under MAR or MNAR assumptions. A limitation of complete-case analysis is that it reduces sample size, resulting in a loss of statistical power. The true trajectory of change may be underestimated when the last observation is carried forward to account for missingness in longitudinal data. Moreover, mean imputation may underestimate the variability in the data, resulting in underestimated correlations among questionnaire items or summary scores [9]. Other statistical approaches for dealing with missing observations may be less frequently adopted by researchers because they are computationally intensive. These alternative techniques for parameter estimation include expectation–maximization (EM) imputation, full information maximum likelihood (FIML) imputation, multiple imputation, inverse-probability weighting, pattern mixture, and selection models [1, 9, 17, 18]. While imputation-based methods rest on the assumption that the missing observations are MAR, pattern mixture and selection models assume that the data are MNAR.

Despite the existing range of statistical methods for handling missing data in HRQOL studies, the influence of missing data on the detection of RS in longitudinal HRQOL data is unknown. We hypothesize that the choice of methods for handling missing data can influence the conclusion about the presence of RS in longitudinal HRQOL. More importantly, we argue that, although, imputation methods tend to result in increased sample size, these methods can result in either upward or downward bias in domain correlations and RS effect sizes, which in turn, may affect the statistical power to detect the presence of RS. Therefore, the purpose of this study is to investigate the effect of the choice of methods on conclusions about the presence of reprioritization RS. Data from an existing longitudinal data set of caregivers of stroke patients are used to demonstrate the implementations of these methods. Recommendations about strategies for handling missing data will be developed based on these study results.

Methods

Data Source

Data are from a multicentre observational longitudinal study "*Understanding Quality of Life Post-Stroke: A Study of Individuals and Their Caregivers*" [26, 27]. This study aimed to understand the effects of stroke severity on HRQOL of stroke survivors and their caregivers over a one-year period post-stroke. Data were collected on 678 stroke patients and their 409 caregivers at 1, 3, 6, and 12 months post-stroke. Study non-participation was low;

only 67 individuals who were approached refused to participate (16.3 %).

Data on stroke patients' and their caregivers' HRQOL were collected using both generic and stroke-specific measures, including the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36), Health Utility Index [11, 25], and Stroke Impact Scale [8] and other stroke-specific HRQOL measures such as the Stroke Impact Scale [7, 46]. The SF-36, which is the focus of this study, encompasses eight domains of HRQOL, including role physical, bodily pain, physical functioning, general health, role emotional, mental health, vitality, and social functioning. The domain scores are scaled to range in value from zero (poor health) to 100 (good health).

Severity of stroke was measured in survivors using the 16-item Stroke Impact Scale (SIS-16) [7]. The items encompass four physical functioning domains Scores range in value from zero (poor functioning) to 100 (good function). A score below 75 reflects severe disability in physical functioning, while scores above 75 indicate less-severe disability; these two stroke severity groups were used in the RS analysis. This instrument has been shown to possess good psychometric properties [7]. Data on caregivers' characteristics such as age, sex, previous experience as a caregiver, number of children, and relationship with patients were also collected at baseline.

Missing data methods

Three commonly adopted missing data methods for longitudinal data were selected for comparison. These were mean imputation, EM imputation, and multiple imputation [15, 32, 38].

Mean imputation

The mean imputation method for longitudinal data replaces missing values on each variable and measurement occasion by the mean of the total sample for each domain at each measurement occasion [15, 32, 38, 43]. While this imputation method is easy to implement, it can underestimate standard errors of parameter estimates, which may lead to erroneous conclusions on tests of significance. Mean imputation can be applied to missing item scores or domain scores. One advantage of item-level mean imputation over domain-level mean imputation is that it maximizes the number of observations. However, there have been few comparisons of the effect that item- and domain-level mean imputation has on study conclusions [13, 20].

EM likelihood-based imputation

Likelihood-based methods, such as EM and FIML, have been developed for handling missing data in cross-sectional data but also extended to longitudinal data [6, 10, 37]. The missing values are predicted based on maximum likelihood estimates from the available data. The EM algorithm is an iterative process that consists of two steps: the expectation (E) step and the maximization (M) step. In the E step, the expectation of the complete data log-likelihood is derived, given the observed data and the estimated parameters from a previous iteration. In the M step, the conditional expectation of the complete data log-likelihood is maximized. The observed data log-likelihood is increased until the algorithm converges. The EM imputation method is advantageous in that it allows the inclusion of the maximum number of observations. The EM method provides unbiased and efficient parameter estimates when data are MCAR or MAR but not when the data are MNAR.

Multiple imputation

For the multiple imputation method, several datasets are created based on random draws from a distribution of plausible missing values, the analytic technique is applied to each dataset, and the relevant parameter or estimate is summarized [14, 39]. This method overcomes the underestimation of standard errors of parameter estimates that arises with single imputation. Three widely used multiple imputation methods are the predictive model method, propensity score method, and Monte Carlo Markov Chain (MCMC) method. While the first two methods are appropriate for monotone missingness, the MCMC method can be applied to non-monotone missingness. A dataset is said to have a monotonic missing pattern when Y_{ijk}, the ith ($i = 1, \dots n$) subject's measurement on the kth ($k = 1, \dots, p$) domain variable in the jth ($j = 1, \dots J$) group, is missing for an individual implies that all subsequent measurements are missing for the individual. Standard statistical procedures are applied to each of the imputed datasets to estimate the parameters, and the results are combined, as recommended by Rubin [34]. Let Q_m be an estimate of the parameter of interest obtained from the mth ($m = 1, \dots, M$) imputed dataset and U_m is the standard error associated with Q_m. The overall parameter estimate is calculated as the average of the parameter estimates from the imputed datasets,

$$\overline{Q} = \frac{1}{M} \sum_{m=1}^{M} Q_m, \tag{1}$$

while its standard error is calculated as

$$T = \sqrt{\overline{U} + (1 + \frac{1}{M})B}, \qquad (2)$$

where $\overline{U} = \frac{1}{M}\sum_{m=1}^{M} U_m$ is the average of the within-imputation variance, and $B = \frac{1}{M}\sum_{m=1}^{M}(Q_m - \overline{Q})^2$ is the between-imputation variance. Previous research has shown that the efficiency of parameter estimates based on multiple imputations increased as the number of imputation increased [34] Although three to five imputations have been found to be sufficient in many applications especially when the proportion of missing data is <30 %, as many as 100 imputations have been recently recommended when the proportion of missing data is more than 30 % [14]. Standard multiple imputation strategies are readily implemented in standard statistical software packages including SAS [16] (SAS Inc 2011). Moreover, even though the multiple imputation model for continuous outcomes rests on the assumption of a multivariate normal distribution, previous research has shown that multiple imputation based on the MCMC method can be robust to non-normality when compared to other methods [24, 37]. As a result, the multiple imputation method for continuous outcomes can also be used to impute missing data on HRQOL items that are binary or ordinal. A rounding strategy is recommended so that the imputed responses can fit the range of values for the binary or ordinal variables [16, 37].

Relative importance measures for reprioritization RS

Statistical tests based on measures of relative importance derived from descriptive discriminant analysis and logistic

regression models have been recently developed for detecting RS in longitudinal HRQOL data [23]. These test procedures are based on the premise that there is a statistically significant longitudinal change in the relative importance of the HRQOL domains when reprioritization RS is present in the data (Figs. 1, 2).

Measures of relative importance derived from discriminant analysis or logistic regression models use weights (i.e., variable coefficients) or ranks (i.e., rank order of the coefficients based on magnitude). This approach has been proposed to quantify the relative importance of variables that discriminate between two or more groups defined a priori. These include standardized discriminant function coefficients (SDFC), discriminant ratio coefficients, standardized logistic regression coefficients (SLRC), and Pratt's index for logistic regression [35, 44, 45]. The first two measures are developed based on discriminant analysis, and the latter two are derived from logistic regression models. These measures have been described in detail elsewhere [35]. In the logistic regression model, the group-membership variable (e.g., stroke severity) is the dependent variable, while independent variables are the domain scores. In contrast, in discriminant analysis, the domains are the dependent variables model, while group-membership variable is the independent variable. To test for reprioritization RS, discriminant analysis or logistic regression models are conducted independently at each measurement occasion. To generate the sampling distribution of the differences in relative importance weights, bootstrap samples in each group were centered by subtracting the group means from the original data. The relative importance

Fig. 1 Changes in Standardized Discriminant Function Coefficients for SF-36 Domains between Baseline and 6 Months by Missing Data Methods *BP* SF-36 Bodily pain domain, *PF* SF-36 Physical functioning domain, *RP* SF-36 Role physical domain, *RE* SF-36 Role emotional domain, *SF* SF-36 Social functioning domain, *MH* SF-36 Mental health domain, *VT* SF-36 Vitality domain, and *GH* SF-36 General health domain

Fig. 2 Changes in Standardized Logistic Regression Coefficients for SF-36 Domains between Baseline and 6 Months by Missing Data Methods. *BP* SF-36 Bodily pain domain, *PF* SF-36 Physical functioning domain, *RP* SF-36 Role physical domain, *RE* SF-36 Role emotional domain, *SF* SF-36 Social functioning domain, *MH* SF-36 Mental health domain, *VT* SF-36 Vitality domain, and *GH* SF-36 General health domain
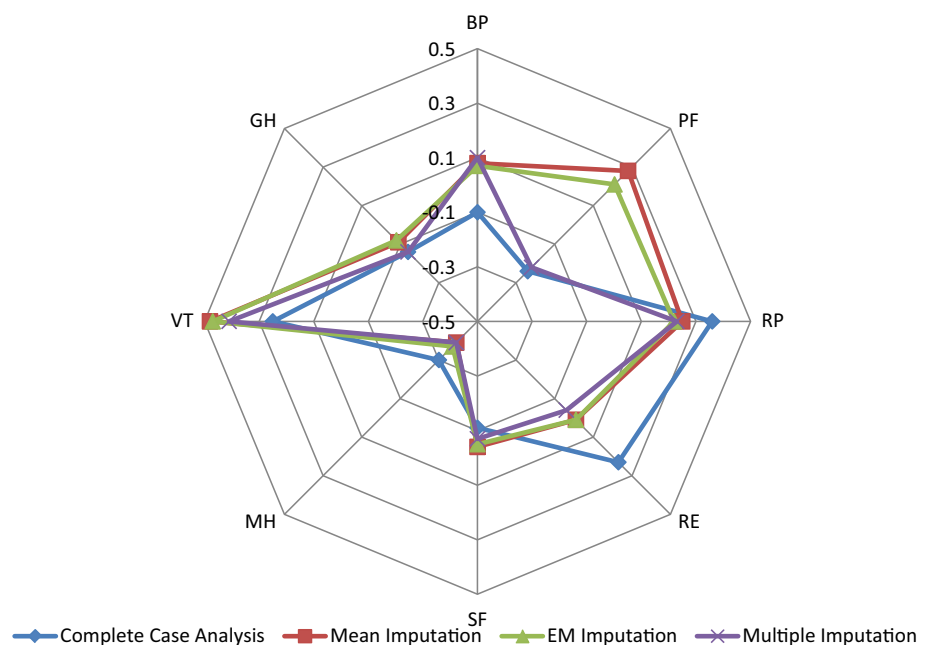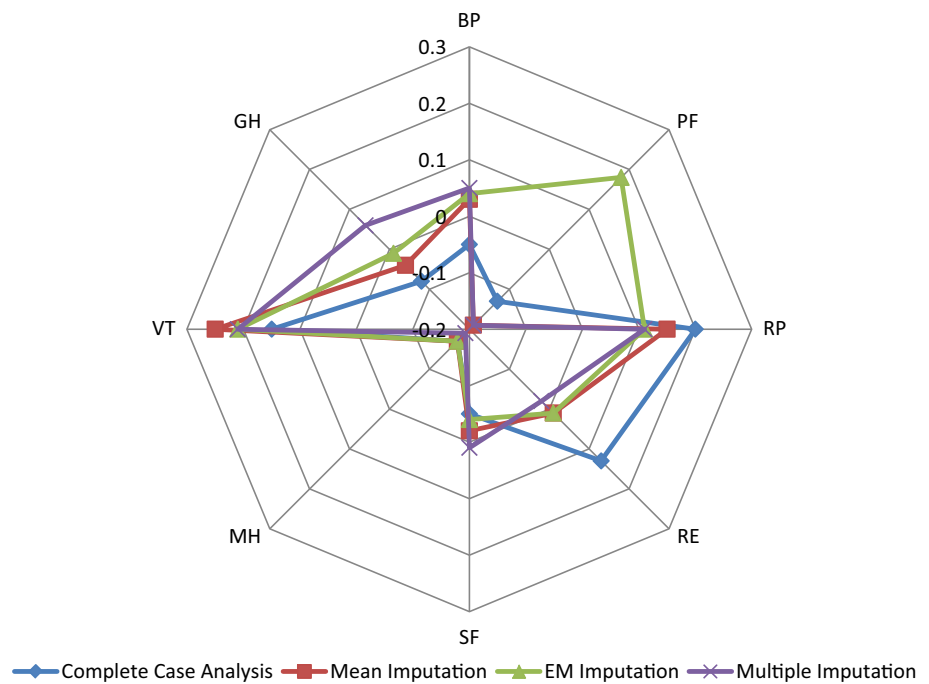

weights were then calculated from the centered data at each measurement occasion, and the bootstrap difference in relative importance weights was estimated. For each of the relative importance measures, reprioritization RS was considered present in a domain if the change in signed values of the relative importance weight or for a domain is statistically significant at a pre-specified level of significance (i.e., $\alpha = 5$ %) [23].

Statistical analysis

We investigated the impact of missing data methods on test procedures based on changes in the relative importance of HRQOL domains in detecting reprioritization RS in longitudinal data. Data from the self-reported HRQOL of stroke caregivers at baseline and month 6 post-stroke are used to demonstrate the implementation of these methods. This analysis focuses on detecting reprioritization RS among the SF-36 domains over the 6-month period post-stroke for caregivers of patients with severe and less-severe stroke. Preliminary descriptive analyses of the demographic characteristics and self-reported SF-36 domain scores for the stroke caregivers were conducted at baseline. To determine the pattern of missing data, patients with complete data at both occasions and those with incomplete data were compared on demographic characteristics and SF-36 domain scores at study baseline using t tests or χ^2 tests. The missing data methods were compared using test procedures based on changes in relative importance weights to detect reprioritization RS among the SF-36 domains.

For the multiple imputation model, each SF-36 domain at baseline and month 6 were included in the imputation model along with age of caregivers, number of children, gender, and relationship with spouse as ancillary variables. Given the apparent arbitrary missing data pattern observed for this dataset, a MCMC multiple imputation model was adopted and developed independently for each stroke severity group [24]. Given that about 40 % of the study participants had missing data on at least one measurement occasion, 100 imputed datasets were created from the multiple imputation model, in keeping with recommendations from previous research [14].

Given that no single relative importance measure has been uniformly recommended, the relative importance weights are derived from more than one relative importance measure, namely: SDFC derived from discriminant analysis and SLRC derived from the logistic regression model. For each measure, discriminant analysis and logistic regression analysis were conducted on data at baseline and 6-month post-baseline. Statistical significance of the changes in the relative importance weights was determined by the empirical bootstrap method, in which the distribution of the test statistic is approximated using $B = 500$ bootstrap samples, in keeping with previous research recommendations [4, 5]. Reprioritization RS was considered present in a domain if the change in the coefficients derived from each relative importance measure was statistically significant ($\alpha = 0.05$).

Given that our example data contain missing observation, the assessment of statistical significance may vary

depending on the missing data method adopted. For complete-case analysis, relative importance test procedures for RS detection were conducted on the remaining data, after case-wise deletion of observations with missing data. For mean imputation and EM imputation methods, test procedures of relative importance were conducted on the imputed datasets. For the multiple imputation method, the differences in relative importance weights over the two occasions for the original data were estimated by using Rubin's rule [34] to combine the imputation-specific estimates of differences in relative importance weights. The distribution of the differences in relative importance weights was obtained by repeatedly sampling from the original incomplete data with replacement (i.e., bootstrap) to create B bootstrap datasets. For each bootstrap sample, the differences in relative importance weights were estimated based on multiple imputation of the bootstrap sample. A detailed description of the implementation of relative importance test procedures for multiple imputed dataset is described in "Appendix 1". All analyses were conducted using SAS version 9.3 [36].

Results

Table 1 describes the demographic characteristics and self-reported SF-36 domain scores for caregivers of survivors with severe and less-severe stroke conditions at study baseline. Of the 409 caregivers, about 344 patients had complete data on the SF-36 domains at baseline, while 281 patients had complete data on all domains at month 6. Spousal caregivers comprised 64.7 % of the sample; 39.9 % of the caregivers were caring for patients with severe stroke conditions. The caregivers were predominantly female.

Of the 409 caregivers recruited into this study, 209 (50.0 %) has complete data on all the SF-36 domains and at all four measurement occasions. At baseline, 344 (84.1 %) of the participants had complete data, while 68.7 % of the participants had complete data at month 6. Further analyses shows that the pattern of missingness was significantly associated with severity of stroke for the patients receiving care; caregivers of patients with less-severe stroke conditions were more likely to have complete data on all four more occasions than caregivers of patients with severe stroke conditions. However, there was no significant association between the likelihood of missing data and the stroke severity after adjusting for caregiver's relationship with the stroke patients. We therefore assume that the missing data mechanism was MAR.

The results of the comparison of relative importance weight analyses as a function of the missing data method used are presented in Table 2. These analysis results are

Table 1 Characteristics of stroke caregivers at study baseline

Patients' characteristics	Severe stroke ($N_1 = 246$)	Less-severe stroke ($N_2 = 163$)
Age, mean (SD)	60.43 (14.82)	55.60 (14.62)
Gender (% male)	33.04	26.92
Relationship to patient (% spousal)	30.49	26.99
Children (% yes)	18.14	20.75
Experience as a caregiver (% yes)	56.44	56.33
Patient-reported outcomes, mean (SD)		
Stroke Impact Scale Physical	44.44 (20.51)	89.24 (8.05)
SF-36 bodily pain	76.69 (27.60)	79.94 (25.01)
SF-36 physical functioning	84.71 (21.26)	87.03 (19.27)
SF-36 role physical	73.06 (39.77)	81.13 (24.04)*
SF-36 role emotional	70.66 (40.52)	79.25 (35.64)*
SF-36 social functioning	71.63 (30.91)	84.11 (21.93)*
SF-36 mental health	71.83 (21.02)	75.14 (19.08)
SF-36 vitality	50.56 (25.32)	58.42 (23.16)*
SF-36 general health	73.92 (19.01)	75.23 (20.88)

SD standard deviation

* $p < 0.05$

based on changes in relative importance of the domains between baseline and 6-month follow-up. The magnitudes of changes in SDFCs were largest for role physical domain for complete-case analysis. The vitality domain had the largest changes in SDFC for mean imputed, EM imputed, and multiple imputed datasets.

When the complete-case analysis method was adopted, there was no evidence of statistical significant change in the SDFCs over the two measurement occasions for any of the SF-36 domains. But there were statistically significant changes in SDFC on physical functioning and vitality domains when mean imputation method was adopted. For EM imputation methods, the magnitude of change in SDFCs over time was statistically significant for the vitality domain only. There were statistically significant changes in SDFCs for mental health and vitality domains, when multiple imputation method was adopted.

Furthermore, unlike test procedures based on SDFC, the range of changes in the SLRCs among the domains and the corresponding standard errors were smaller. There was no statistically significant change in the SLRCs for any of the SF-36 domains when complete-case analysis method was adopted. But there were statistically significant changes in the coefficients for the physical functioning and vitality domains when mean imputation or EM imputation method was adopted. Statistically significant changes in SLRCs were also noted on physical functioning, mental health, and vitality domains when multiple imputation method was adopted.

Table 2 Changes in relative importance weights for SF-36 domains by missing data method

Test	Missing data method	BP	PF	RP	RE	SF	MH	VT	GH
Test procedures based on standardized discriminant function coefficients	Complete case	−0.10 (0.22)	−0.24 (0.18)	0.36 (0.24)	0.23 (0.25)	−0.11 (0.21)	−0.30 (0.24)	0.25 (0.22)	−0.14 (0.16)
	Domain mean imputation	0.08 (0.19)	**−0.28 (0.14)**	0.25 (0.21)	0.01 (0.20)	−0.04 (0.18)	−0.39 (0.21)	**0.48 (0.19)**	−0.09 (0.17)
	EM imputation	0.07 (0.18)	−0.21 (0.15)	0.22 (0.21)	0.1 (0.22)	−0.05 (0.16)	−0.37 (0.20)*	**0.47 (0.18)**	−0.08 (0.16)
	Multiple imputation	0.10 (0.23)	−0.22 (0.25)	0.23 (0.23)	−0.04 (0.18)	−0.07 (0.13)	**−0.39 (0.14)***	**0.41 (0.20)***	−0.14 (0.14)
Test procedures based on standardized logistic regression coefficients	Complete−case	−0.05 (0.10)	−0.13 (0.09)	0.20 (0.12)	0.13 (0.13)	−0.05 (0.10)	−0.17 (0.11)	0.15 (0.11)	−0.08 (0.08)
	Domain mean imputation	0.03 (0.09)	**−0.19 (0.07)**	0.15 (0.10)	0.01 (0.10)	−0.02 (0.08)	−0.17 (0.12)	**0.25 (0.10)**	−0.04 (0.07)
	EM Imputation	0.04 (0.09)	**−0.18 (0.08)**	0.11 (0.09)	0.01 (0.10)	−0.04 (0.09)	−0.17 (0.10)	**0.21 (0.09)**	−0.01 (0.08)
	Multiple Imputation	0.05 (0.08)	**−0.19 (0.09)**	0.11 (0.15)	−0.02 (0.12)	0.01 (0.09)	**−0.19 (0.08)**	**0.21 (0.10)**	−0.06 (0.09)

EM expectation–maximization, *BP* bodily pain *PF* physical functioning, *RP* role physical, *RE* role emotional, *SF* social functioning, *MH* mental health, *VT* vitality, *GH* general health

Numeric values in bold fonts corresponds to statistically significant changes in relative importance weights ($p < 0.05$)

Simulation study

A Monte Carlo simulation study was conducted to examine the statistical power of test procedures based on SDFC and SLRC under complete-case analysis, as well as with mean imputation, EM imputation, and multiple imputation methods. All procedures were investigated for two independent groups. The simulation conditions investigated included: (a) total sample size, (b) number of study variables, (c) missing data mechanisms, and (d) proportion of missing data. Total sample sizes of $N = 100$ and 250 were investigated, representing small and medium population sizes. For each group, unequal group sizes were considered using a 3:2 ratio. The number of outcome variables was fixed at $p = 4$. Given that the investigated procedures are likely to be influenced by the missing data mechanism, we compared the relative importance test procedures under MCAR, MAR, and MNAR missingness mechanisms. The percentages of missing observations were set at 15 and 25 %.

The multivariate longitudinal data over two measurement occasions for each group. First, the multivariate data at baseline were generated from a multivariate normal distribution with known means and group covariances using the *mvtnorm* package in R [33]. The data at the second measurement occasion was generated as a linear combination of baseline data, change from baseline, and multivariate latent factor representing reprioritization RS effects on the domains. The multivariate data at the second occasion were therefore obtained as.

$$\mathbf{Y}_{ij2} = \mathbf{Y}_{ij1} + \boldsymbol{\Delta}_{ij} + \mathbf{I}_{ij}$$

where $\mathbf{Y}_{ij1}$ is the $p \times 1$ vector of observations for the ith ($i = 1, \ldots, n_j$) subject in the jth ($j = 1, 2$) group at baseline, $\boldsymbol{\Delta}_{ij}$ is the $p \times 1$ change from baseline for the ith subject, and $\mathbf{I}_{ij}$ is the $p \times 1$ latent factor vector representing the RS effect for the ith subject. Assume $\mathbf{Y}_{ij1} \sim N_p(\boldsymbol{\mu}_{j1}, \boldsymbol{\Sigma}_1)$, $\boldsymbol{\Delta}_{ij} \sim N_p(\boldsymbol{\eta}_j, \boldsymbol{\Phi}_j)$, $\mathbf{I}_{ij} \sim N_p(\boldsymbol{\theta}_j, \boldsymbol{\Psi}_j)$, where $\boldsymbol{\eta}_j$, $\boldsymbol{\theta}_j$, $\boldsymbol{\Psi}_j$, and $\boldsymbol{\Phi}_j$ are the $p \times 1$ vector of mean change from baseline, $p \times 1$ vector of mean latent factors, $p \times p$ covariance matrix for the change from baseline, and $p \times p$ covariance matrix for the latent factors, respectively, for the jth group. The data for the second measurement occasion, $\mathbf{Y}_{ij2}$, were generated from a multivariate normal distribution with jth group mean as.

$$\boldsymbol{\mu}_{j2} = \boldsymbol{\mu}_{j1} + \boldsymbol{\eta}_j + \boldsymbol{\theta}_j$$

and the jth group covariance is

$$\boldsymbol{\Sigma}_{j2} = \boldsymbol{\Sigma}_{j1} + \boldsymbol{\Phi}_j + \boldsymbol{\Psi}_j$$

The mean configurations for the multivariate data at baseline, change from baseline, and the latent RS effects

are described in Table 3. For multivariate data at baseline, a quadratic mean configuration across the domains in Group 1 while a constant mean configuration, represented by a unit vector, was assumed as the mean vector for Group 2. On the other hand, a linearly decreasing mean configuration was assumed across the domains for θ_j in Group 1, while a null vector was assumed for θ_j in Group 2.

Moreover, relative importance measures are known to be influenced by both the magnitude and pattern of correlation among the observations [45]. A compound symmetric (CS) structure was assumed for the correlation among the domains at baseline. More specifically, the following CS correlation matrices were assumed for multivariate data at baseline and multivariate change from baseline: (a) Q_1: CS with $\rho = 0.3$ and (b) Q_2: CS with $\rho = 0.7$. On the other hand, a CS correlation structure was adopted for the latent factor variables such that the magnitude of the correlation among the latent factors was assumed $\rho = 0.2$.

Conditions for statistical power are investigated by simulating known RS effects, as measured by the multivariate latent factor, in the multivariate data at the second measurement occasion. RS effects were assumed to be present on all the domains in Group 1, but not in Group 2. The statistical power for each procedure was measured

using all-variable power (ALLV), the power of a test procedure to reject the p non-null hypothesis in the family of p hypotheses. A total of 500 simulations were conducted for each of the combination of simulation conditions. For each procedure, $B = 500$ bootstrap samples were used to generate the distribution of differences between relative importance weights. For $\alpha = 0.05$, a Bonferroni correction was used to adjust for multiple correlated tests of significance. Differences in statistical power <5 % were considered negligible, while differences greater than 10 % were considered substantial.

Simulation results

Table 4 describes the average ALLV power for relative importance test procedures based on SLRC and SDFC when complete-case analysis was conducted, as well as with mean imputation, EM imputation, and multiple imputation methods. These results were stratified by total sample size, percentage of missing data, and missingness mechanism. In general, the average ALLV power for test procedures based on the SDFC was at least 9.0 % higher than the average ALLV power for test procedures based on SLRCs, regardless of the percentage of missing data, missingness mechanism, and method for adjusting for missing data. On the other hand, the average ALLV power for each procedure decreased as the percentage of missing data increased, regardless of the sample size, mechanism of missing data, and imputation method. For example, when $N = 250$, the average ALLV power for tests based on SLRC and SDFC procedures were 19.3 and 30.8 %, respectively, for complete-case analysis method when 15 % of the data were missing via the MCAR mechanism.

Table 3 Description of mean patterns for the simulation study

Mean configurations	Group 1	Group 2
Mean at baseline (μ)	(0.5, 2, 1.5, 1)	(1, 1, 1, 1)
Mean change (η)	(1, 1, 1, 1)	(1, 1, 1, 1)
Latent variable mean (θ)	(3, 1.5, −1.5, −1.5)	(0, 0, 0, 0)

Table 4 Average all-variable power (%) for relative importance tests by missing data method when $p = 4$ (simulation results)

% Missing	Missing data mechanism	Standardized discriminant function coefficients				Standardized logistic regression coefficients			
		CC	Mean	EM	MI	CC	Mean	EM	MI
$N = 100$									
15	MCAR	30.8	41.3	40.4	41.5	19.3	26.1	27.5	27.9
	MAR	31.2	40.8	39.9	41.4	19.4	26.6	27.5	27.5
	MNAR	30.3	39.6	38.5	39.2	14.7	21.1	21.8	22.5
25	MCAR	25.5	43.7	41.4	44.3	15.5	29.0	29.1	29.9
	MAR	25.1	40.5	41.3	42.8	15.0	26.6	29.5	29.6
	MNAR	26.1	40.3	38.9	40.7	10.7	16.9	21.6	21.1
$N = 250$									
15	MCAR	88.6	92.5	90.9	91.6	79.3	84.7	83.8	85.1
	MAR	88.6	91.8	90.8	92.0	80.6	84.8	84.3	85.8
	MNAR	87.0	90.5	88.5	90.0	72.6	77.5	77.3	78.8
25	MCAR	82.9	90.2	87.9	90.7	70.9	82.7	81.4	83.0
	MAR	82.3	89.9	88.5	90.1	71.4	81.5	81.0	83.3
	MNAR	78.9	87.5	83.5	86.2	56.8	67.9	69.00	70.5

MCAR missing completely at random, *MAR* missing at random, *MNAR* missing not at random, *CC* complete-case analysis, *Mean* mean imputation, *EM* EM imputation, *MI* multiple imputation, *% Missing* percent of missing data

But when the percentage of missing data was 25 %, the average ALLV power for the former and latter procedures were 15.5 and 25.5 %, respectively.

There were negligible differences in average ALLV power of the test procedures under both MCAR and MAR missingness mechanisms. However, both test procedures were not equally robust to MNAR mechanism. More specifically, test procedure based on SDFC was not influenced by the pattern of missing data, but test procedure based on SLRC was substantially less powerful when the data were missing via the MNAR mechanism. For example, when 25 % of the data were missing, the average ALLV power for test procedure based on SDFC for MCAR, MAR, and MNAR were 82.9, 82.3, and 78.9 %, respectively, when complete-case analysis method was adopted. In contrast, the average ALLV for test procedure based on SLRC were 70.9, 71.4, and 56.8 %, when the data are MCAR, MAR, MNAR, respectively.

Moreover, both test procedures were sensitive to the missing data method. More specifically, the ALLV power for each procedure was smallest when the complete-case method was adopted and largest when an imputation method was adopted. On the other hand, there were no substantial differences in the average ALLV power for test procedures based on mean, EM, and multiple imputation methods. For example, for $N = 250$, when 25 % of the data were missing via the MCAR mechanism, the average ALLV power for test procedure based on SLRC under complete-case analysis, mean imputation, and multiple imputation methods were 70.2, 83.5, and 82.7 %, respectively.

Finally, the average ALLV power for each procedure when $N = 250$ was at least two times the average ALLV power when $N = 100$, regardless of the mean configuration, percentage of missing data, missing data mechanism, and missing data method. For example, when 15 % of the data were missing via the MCAR mechanism, the average ALLV power for test procedures based on SDFC and SLRC procedures under mean imputed data were 44.9 and 20.3 %,respectively, when $N = 100$. When $N = 250$, the average ALLV power for the test procedures under mean imputed data were 92.5 and 84.7 %, respectively.

Discussion

This study investigated the presence of reprioritization RS in a stroke caregiver population characterized by missing data. The choice of imputation methods used to address missing data will influence conclusions about the presence of RS. Neither of the two relative importance tests for reprioritization RS showed any evidence of RS when complete-case analysis was adopted. But the test procedures based on SLRC and SDFC revealed the presence of RS on physical function and vitality domains when mean

imputation or EM imputation was adopted. The multiple imputation method showed evidence of RS on physical functioning, mental health, and vitality domains.

The findings about RS detection in the numeric example are consistent with results of our simulation study, which revealed that relative importance tests procedure based on complete-case analysis are less powerful in detecting RS than tests based on imputation methods. The statistical power of these test procedures decreased as the proportion of missing data increased, but the power of the test procedures increased as the sample size increased. On the other hand, there were negligible differences in statistical power of the procedures when mean, EM, and multiple imputation methods were adopted. Furthermore, our simulation results revealed that a test procedure based on changes in SLRCs was substantially less powerful when the data were MNAR compared to the power results when the data are MCAR. There were negligible differences in the average ALLV power of the test procedure based on SDFCs when the data were missing via the MCAR or MNAR mechanisms, regardless of the choice of missing data method, percentage of missing data, and sample size.

Although results from the simulation study revealed that relative importance tests based on imputation methods are more powerful to detect RS than tests based on complete-case analysis, the analysis results of the example datasets showed variations in estimates of differences in relative importance weights over the 6-month period for various imputation methods for each relative importance test procedure. The observed variations can be attributed to several factors such as imputation methodology, domain correlations, with-in domain autocorrelations, and mechanism of missing data. For example, the choice of an imputation method can affect the estimates of the difference in relative importance weights on the domains. Given that, mean imputation tends to underestimate the correlations among the SF-36 domains, estimates of relative importance weights at each measurement occasions, and estimates of differences in relative importance weights over the two measurement occasions may be biased. In contrast, both EM and multiple imputation models preserve the correlation structure by estimating the missing data through maximum likelihood and MCMC methods, respectively. We recommend that the changes in relative importance weights as well as the corresponding standard errors for the imputation methods be considered when determining whether the choice of imputation methods might influence the interpretation of the original clinical study.

The study has some caveats. Our investigations of the mechanism of missing data in the example data suggest that the missing observations for participants over the two measurement occasions are likely to be MAR, because the pattern of missingness over the 6-month follow-up was associated with self-reported domain scores at baseline.

Nonetheless, it is not known whether the missing data mechanism might also be associated with unobserved variables (i.e., MNAR), and this assumption cannot be tested. On the other hand, the simulation study described here only focused on the statistical power of the test procedures when the correlations among the domains are assumed to be equal. This assumption may not be realistic in longitudinal HRQOL studies characterized by heterogeneous (i.e., unstructured) between-domain correlation matrix.

In summary, this study investigates test procedures based on relative importance measures for RS detection in incomplete longitudinal HRQOL data in the context of complete-case analysis, and three standard missing data imputation methods. Our findings revealed that the detection of reprioritization RS among the HRQOL domains is affected by the choice of both the proportion and type of missing data mechanism and the imputation methods. More specifically, relative importance test procedures for RS based on changes in SDFCs and SLRCs are most powerful when mean or multiple imputation methods are used to fill missing observations. We therefore recommend that one of these methods should be adopted when testing for RS in incomplete longitudinal HRQOL data.

Appendix 1

Relative importance tests for response shift detection in incomplete longitudinal data.

To implement relative importance tests under complete-case analysis method, observations with at least one missing observation on any of the domains are deleted. The differences in relative weights are conducted on the complete data. Similarly, for the mean imputation and EM imputation methods, tests of differences in relative importance weights are conducted on the mean imputed and EM imputed datasets, respectively.

In contrast, for multiple imputation methods, the following steps are taking to implement the relative importance tests.

1. Estimate the observed differences in relative importance weights for domains on the original data. This is accomplished as follows:

   a. Multiply impute the original incomplete longitudinal data by creating M copies of the imputed dataset.

   b. For each imputed dataset, estimate the relative importance weights at each occasions (w_{1mk} and w_{2mk}; $m = 1, \ldots, M$; $k = 1, \ldots, p$) and the differences in relative importance weights over the two occasions (i.e., $W_{mk} = w_{1mk} - w_{2mk}$).

   c. Use Rubin's rule as described in Sect. "Methods" to combine the estimates of the differences in relative importance weights for the M datasets. The average difference in relative importance weights is called the observed relative importance weights, $\hat{W}_k$.

2. Generate the distribution of the differences in relative importance weights by repeatedly sampling from the original incomplete data with replacement (say B times).

   a. For each bootstrapped incomplete dataset, implement steps 1a–c to estimate the bootstrap-specific differences in relative weights $\hat{W}_{bk}$.

   b. Sort the $\hat{W}_{bk}$ s ($b = 1, 2, \ldots, B$) from the smallest to the highest. This is the bootstrap distribution of the difference in relative importance weights from the B bootstrap samples.

   c. Estimate the W_{bk}s that corresponds to the 100 ($\alpha/2$)th percentile (W_{bkL}) and the 100 ($1-\alpha/2$)th percentile (W_{bkU}) of the bootstrap distribution.

3. Statistical significance of the differences in relative importance weights on a domain is considered present if W_k is outside the interval (W_{bkL}, W_{bkU}).

Appendix 2

Description of missing data patterns in the longitudinal study of stroke caregivers.
See Table 5.

Table 5 Description of missing data patterns in the SF-36 domains of the longitudinal study of stroke caregivers

Missing data pattern	Baseline	Month 3	Month 6	Month 12	N
1	O	O	O	O	209
2	O	O	X	X	43
3	O	O	O	O	32
4	O	X	X	X	60
5	X	O	O	O	37
6	X	O	X	X	19
7	X	X	O	O	3
8	X	X	X	X	6

O observed, *X* missing

Appendix 3

See Tables 6 and 7.

Table 6 Percentile confidence intervals and p values for relative importance tests based on standardized discriminant function coefficients and standardized logistic regression coefficients for complete-case analysis and mean imputed data

SF-36 domains	Complete-case analysis				Mean imputation			
	SDFC		SLRC		SDFC		SLRC	
	CI	p values	CI	p values	CI	p values	CI	p values
Bodily pain	(−0.63, 0.60)	0.62	(−0.34, 0.26)	0.65	(−0.49, 0.51)	0.64	(−0.30, 0.22)	0.67
Physical functioning	(−0.58, 0.48)	0.18	(−0.24, 0.20)	0.16	(−0.46, 0.42)	0.01	(−0.22, 0.21)	<0.01
Role physical	(−0.73, 0.68)	0.12	(−0.37, 0.30)	0.15	(−0.75, 0.52)	0.26	(−0.27, 0.27)	0.22
Role emotional	(−0.97, 0.77)	0.36	(−0.36, 0.31)	0.40	(−0.53, 0.65)	0.96	(−0.24, 0.28)	0.94
Social functioning	(−0.83, 0.47)	0.57	(−0.29, 0.29)	0.74	(−0.53, 0.58)	0.78	(−0.20, 0.24)	0.63
Mental health	(−0.72, 0.80)	0.20	(−0.34, 0.33)	0.22	(−0.64, 0.47)	0.06	(−0.24, 0.22)	0.07
Vitality	(−0.77, 0.57)	0.30	(−0.25, 0.33)	0.29	(−0.80, 0.58)	0.03	(−0.30, 0.22)	<0.01
General health	(−0.62, 0.42)	0.37	(−0.25, 0.20)	0.54	(−0.47, 0.45)	0.58	(−0.25, 0.22)	0.62

Table 7 Percentile confidence intervals and p values for relative importance tests based on standardized discriminant function coefficients and standardized logistic regression coefficients for EM imputation and multiple imputation methods

SF-36 domains	EM imputation				Multiple imputation			
	SDFC		SLRC		SDFC		SLRC	
	CI	p values	CI	p values	CI	p values	CI	p values
Bodily pain	(−0.56, 0.55)	0.66	(−0.25, 0.24)	0.68	(−0.49, 0.45)	0.68	(−0.29, 0.25)	0.70
Physical functioning	(−0.44, 0.40)	0.10	(−0.24, 0.20)	0.02	(−0.42, 0.36)	0.03	(−0.18, 0.18)	0.01
Role physical	(−0.65, 0.57)	0.23	(−0.34, 0.28)	0.25	(−0.65, 0.52)	0.01	(−0.31, 0.26)	0.06
Role emotional	(−0.48, 0.52)	0.97	(−0.28, 0.27)	0.95	(−0.62, 0.62)	0.92	(−0.27, 0.26)	0.96
Social functioning	(−0.44, 0.51)	0.80	(−0.23, 0.25)	0.57	(−0.50, 0.47)	0.68	(−0.24, 0.21)	0.61
Mental health	(−0.69, 0.48)	0.07	(−0.32, 0.26)	0.07	(−0.61, 0.52)	<0.01	(−0.25, 0.26)	0.01
Vitality	(−0.54, 0.41)	0.02	(−0.32, 0.25)	0.01	(−0.53, 0.44)	<0.01	(−0.28, 0.19)	0.03
General health	(−0.45, 0.45)	0.58	(−0.19, 0.29)	0.60	(−0.43, 0.43)	0.10	(−0.19, 0.19)	0.14

References

1. Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.
2. Allison, P. D. (2003). Missing data techniques for structural equation modelling. *Journal of Abnormal Psychology, 112*, 545–557.
3. Boucekine, M., Loundou, A., Baumstarck, K., Minaya-Flores, P., Pelletier, J., Ghattas, B., et al. (2013). Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: A cohort study. *BMC Medical Research Methodology, 13*, 20.
4. Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141–1164.
5. Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers*. New Jersey: Wiley.
6. Dempster, A. P., Laird, N. M., & Rubin, B. D. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*, 1–38.
7. Duncan, P. W., Lai, R. K., Bode, R. K., et al. (2003). Stroke Impact Scale-16: A brief assessment of physical function. *Neurology, 60*, 291–296.
8. Duncan, P., Wallace, D., Lai, S., Johnson, D., Embretson, S., & Laster, L. (1999). The stroke impact scale version 2.0. *Stroke, 30*, 2131–2140.
9. Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
10. Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modelling, 8*, 430–457.
11. Feeny, D. H., Torrance, G. W., & Furlong, W. J. (1996). Health utilities index. In Bert Spilker (Ed.), *Quality of life and*

pharmacoeconomics in clinical trials, 2 (Vol. 26, pp. 239–252). Philadelphia: Lippincott-Raven Press.

12. Fielding, S., Fayers, P. M., & Ramsay, C. R. (2009). Investigating the missing data mechanism in quality of life outcomes: A comparison of approaches. *Health and Quality of Life Outcomes, 7*, 57–66.

13. Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*, 1–25.

14. Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputations theory. *Prevention Science, 8*, 206–213.

15. Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (eds.) *Research methods in psychology* (Vol. 2 pp 87–114) of *Handbook of psychology* (I. B. Weiner, Editor-in-Chief). New York: Wiley.

16. Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete regression models. *The American Statistician, 61*, 79–90.

17. Ibrahim, J. G., & Molenbergh, G. (2009). Missing data methods in longitudinal studies: A review. *Test, 18*, 1–43.

18. Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*, 305–315.

19. Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality of life response shift in HIV/AIDS. *Journal of Clinical Epidemiology, 62*, 1138–1147.

20. Li, Y., & Schwartz, C. E. (2011). Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. *Quality of Life Research, 13*, 1543–1553.

21. Lin, T. H. (2006). Missing data imputation in quality of life assessment: Imputation for WHOQOL-BREF. *Pharmacoeconomics, 24*, 917–925.

22. Little, R. J. A., & Rubin, B. D. (2002). *Statistical analysis with missing data* (2nd ed.). New Jersey: Wiley.

23. Lix, L. M., Sajobi, T. T., Sawatzky, R., Liu, J., Mayo, N. E., Huang, Y., et al. (2013). Relative importance measures for reprioritization response shift. *Quality of Life Research, 22*, 695–703.

24. Ma, J., Akhtar-Danesh, N., Dolovich, L., & Thabane, L. (2011). Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology, 11*, 18.

25. Mathias, S., Bates, M., Pasta, D., Cisternas, M., Feeny, D., & Patrick, D. (1997). Use of the health utilities index with stroke patients and their caregivers. *Stroke, 28*, 1888–1894.

26. Mayo, N. E., Fellows, L. K., Scott, S. C., Cameron, J., & Wood-Dauphinee, S. (2009). A Longitudinal view of apathy and its impact after stroke. *Stroke, 40*, 3299–3307.

27. Mayo, N. E., Scott, S. C., Bayley, M., Cheung, A., Garland, J., Jutai, J., & Wood-Dauphinee, S. (2013). Modeling health-related quality of life in people recovering from stroke. *Quality of Life Research*. doi:10.1007/s11136-013-0605-4.

28. Mayo, N. E., Scott, S. C., Dendukuri, N., Ahmed, S., & Wood-Dauphinee, S. (2008). Identifying response shift statistically at the individual level. *Quality of Life Research, 17*, 627–639.

29. Myers, W. R. (2000). Handling missing data in clinical trials: An overview. *Drug Information Journal, 34*, 525–533.

30. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*, 587–598.

31. Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients 10 undergoing invasive surgery. *Quality of Life Research, 14*, 599–609.

32. Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556.

33. R Development Core Team (2012). *R: A language and environment for statistical computing, reference index version 3.1.0*. R Foundation for Statistical Computing, Vienna, Austria.

34. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

35. Sajobi, T. T., Lix, L. M., Clara, I., Walker, J., Graff, L. L., Rawsthorne, P., et al. (2012). Measures of relative importance for health-related quality of life. *Quality of Life Research, 21*, 1–11.

36. SAS Institute Inc. (2012). *SAS/STAT user's guide, version 9.2*. Cary, NC: SAS Institute Inc.

37. Schaffer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.

38. Schaeffer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

39. Schaeffer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.

40. Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine, 48*, 1531–1548.

41. Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology and Health, 19*, 51–69.

42. Schwartz, C. E., Sajobi, T. T., Lix, L. M., Quaranto, B. R., & Finkelstein, J. A. (2013). Changing values, changing outcomes: The influence of reprioritization response shift on outcome assessment after spine surgery. *Quality of Life Research, 22*, 2255–2264.

43. Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology, 6*, 57.

44. Thomas, D. R. (1992). Interpreting discriminant functions: A data analytic approach. *Multivariate Behavioural Research, 27*, 218–223.

45. Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics, 21*, 110–130.

46. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*, 473–483.