The Annals of Applied Statistics
2019, Vol. 13, No. 1, 466–491
https://doi.org/10.1214/18-AOAS1202

DEVELOPMENT OF A COMMON PATIENT ASSESSMENT SCALE ACROSS THE CONTINUUM OF CARE: A NESTED MULTIPLE IMPUTATION APPROACH[1]

BY CHENYANG GU AND ROEE GUTMAN

Harvard University and Brown University

Evaluating and tracking patients' functional status through the post-acute care continuum requires a common instrument. However, different post-acute service providers such as nursing homes, inpatient rehabilitation facilities and home health agencies rely on different instruments to evaluate patients' functional status. These instruments assess similar functional status domains, but they comprise different activities, rating scales and scoring instructions. These differences hinder the comparison of patients' assessments across health care settings. We propose a two-step procedure that combines nested multiple imputation with the multivariate ordinal probit (MVOP) model to obtain a common patient assessment scale across the post-acute care continuum. Our procedure imputes the unmeasured assessments at multiple assessment dates and enables evaluation and comparison of the rates of functional improvement experienced by patients treated in different health care settings using a common measure. To generate multiple imputations of the unmeasured assessments using the MVOP model, a likelihood-based approach that combines the EM algorithm and the bootstrap method as well as a fully Bayesian approach using the data augmentation algorithm are developed. Using a dataset on patients who suffered a stroke, we simulate missing assessments and compare the MVOP model to existing methods for imputing incomplete multivariate ordinal variables. We show that, for all of the estimands considered, and in most of the experimental conditions that were examined, the MVOP model appears to be superior. The proposed procedure is then applied to patients who suffered a stroke and were released from rehabilitation facilities either to skilled nursing facilities or to their homes.

1. Introduction.

1.1. *Overview.* To track and evaluate patients through the post-acute care continuum a common standardized evaluation tool is needed. Current evaluation tools have been largely developed within each type of health care provider, and cannot be easily compared. In inpatient rehabilitation facilities (IRFs), patients' functional

Received September 2017; revised July 2018.

[1] Supported in part by the Changing Long-Term Care in America project funded by the National Institute on Aging (P01AG027296).

Key words and phrases. Data augmentation, EM algorithm, missing data, nested multiple imputation, multivariate ordinal probit model, slice sampler.

status is evaluated by the Functional Independence Measure (FIM). After being discharged from IRFs, the functional status of patients who stay in skilled nursing facilities (SNFs) is collected using the Minimum Data Set (MDS), while the Outcome and Assessment Information Set (OASIS) is collected for patients who receive home health care provided by home health agencies. All of these assessments examine similar functional capabilities (e.g., eating, grooming, dressing, etc.), but the specific instruments, rating scales and instructions for scoring the different activities vary between these post-acute settings. Thus, it is difficult to evaluate and compare the rates of functional improvement experienced by patients treated in the different health care settings.

The Continuity Assessment Record and Evaluation item set is a standardized evaluation tool that was developed for use at acute hospital discharge and at post-acute care admission and discharge [Gage et al. (2012)]. This tool is intended to be a common evaluation tool for evaluating patients across the continuum of post-acute care, and considerable resources were invested in its development. However, implementing new instruments in all post-acute care settings may result in additional investments in administration and training as well as in changes to the reimbursement system [Li et al. (2017)]. Moreover, adopting new instruments would require translating past functional status scores so that comparison to the new scores is possible.

Equating setting-specific instruments so that functional status scores from one instrument could be used interchangeably with ones from another instrument is a possible approach to obtain a common evaluation tool [Kolen and Brennan (2014), Dorans, Pommerich and Holland (2007), von Davier (2011)]. Linking and equating scores across different standardized assessments has been a major focus in the field of educational testing for the past 90 years [see Dorans, Pommerich and Holland (2007), Chapter 2, for details]. Score equating methods have been recently used in health outcomes research. The conversion table method [Velozo, Byers and Joseph (2007)] was used to equate FIM assessments with MDS assessments. Conversion table equates the sum of individual item scores, also referred to as the total score, by matching on latent functional scores that are estimated from two different instruments using Item Response Theory models. This method was also used to equate scores from two physical functioning scales [ten Klooster et al. (2013)], as well as two depression scales [Fischer et al. (2011)]. Because conversion table ignores the variability from the estimation and the imputation processes, it may result in statistically invalid estimates when further analysis is performed using the imputed scores [Gu and Gutman (2017)]. Furthermore, a data set that comprises contemporaneous MDS and OASIS assessments is required in order to equate MDS and OASIS instruments. However, MDS and OASIS assessments are never jointly observed.

We propose a nested multiple imputation procedure [Shen (2000), Harel (2003), Rubin (2003)] to impute unmeasured assessments across the continuum of care.

This procedure enables evaluation and comparison of the rates of functional improvement across different health care settings, and it consists of an Equating step and a Translating step. In the Equating step, we impute the unmeasured assessments in MDS or OASIS that are close to the FIM assessment date to obtain a synthetic data set with simultaneous MDS and OASIS assessments. In the Translating step, we rely on the synthetic data set from the first step to estimate the relationship between MDS and OASIS that will be used to impute multiple unmeasured assessments in MDS or OASIS at later assessment dates. This two-step procedure accounts for the uncertainty in both steps, and provides flexibility for researchers to choose different models in the second step without the need to re-equate the instruments.

The Equating step imputes the missing instruments that consist of multiple ordinal items. The logistic and probit link functions are commonly used to model single ordinal variable. These link functions give similar model fit and predictive performance. Bayesian inference for these models relies on sampling from complex posterior distributions. However, by introducing auxiliary variables, sampling from these posterior distributions can become more efficient. Albert and Chib (1993) described this technique for the probit link function and more recently Holmes and Held (2006) and Polson, Scott and Windle (2013) proposed two possible approaches for the logistic link function. The multivariate ordinal probit (MVOP) model was proposed as an extension of the probit model [Albert and Chib (1993)] and the multivariate probit model [Ashford and Sowden (1970)] to multivariate ordinal responses. Similar extensions for logistic link function with multiple ordinal outcomes is an area of future research. Using the MVOP model, we can capture the complex dependence structure and the ordinal nature of the different functional assessment instruments as well as adjust for observed patients' covariates.

To generate multiple imputations of the unmeasured functional assessments using the MVOP model, we develop two computational approaches. The first approach combines the EM algorithm [Dempster, Laird and Rubin (1977)] for obtaining the maximum likelihood estimates of the parameters in the MVOP model and the bootstrap method to multiply and impute the missing values [Little and Rubin (2002)]. The second approach relies on the data augmentation (DA) algorithm [Tanner and Wong (1987)] to draw the unknown parameters of the MVOP model and the missing values from their joint posterior distribution. We compared the MVOP model to existing methods for imputing incomplete multivariate ordinal variables with respect to the biases, the sampling variances, and the RMSEs of their point estimates, as well as the widths and coverage rates of their interval estimates. For all of the estimands considered and in most of the experimental conditions that were examined, the MVOP model appears to be superior. In the Translating step, different models can be used to estimate the relationship between MDS and OASIS assessments. We illustrate this flexibility either by imputing the missing individual items using the MVOP model or by imputing the missing total scores using a linear regression model.

The remainder of this section describes the analytical data set, introduces the basic framework and reviews related work. Section 2 describes the MVOP models and their estimation methods. Section 3 presents the nested multiple imputation procedure. Section 4 compares the MVOP models to existing methods using a simulation study. Section 5 describes the empirical data analysis. Conclusions and discussions are provided in Section 6.

1.2. *Motivating example.* The analytical data set includes 72,575 patients who suffered a stroke and were discharged from IRFs between 2011 and 2014. Of these patients, 38,629 were released to SNFs, where the MDS assessments were collected for them. The other 33,946 patients were discharged home, where the OASIS assessments were used to measure their functional status. Patient assessments were collected on admission and at various time points during their post-acute stays. The median number of assessments for each patient was 5 for patients in SNFs and the range was 0 to 91. The median number of assessments for patients receiving home health care was 3 and the range was 0 to 46. Two assessments for each patient were included in our analyses. One assessment was collected at admission within 30 days from the IRF's discharge date. The other was recorded approximately 30 days after the first assessment. The primary research objective is to examine and compare the rates of functional improvement experienced by patients treated in the different health care settings after being discharged from IRFs. To describe the functional change for patients who were released to either SNFs or home, we will use the MDS scale.

1.3. *Basic framework.* We consider equating setting-specific patient assessments as a missing data problem. We assume that all patients have complete FIM assessments and complete demographic characteristics. Let $\mathbf{M} = \{M_i\}$, $i = 1, \dots, N$, where M_i is an indicator that is equal to 1 if patient i was discharged home and 0 otherwise. Let $\mathbf{Y}^{\text{fim}}$, $\mathbf{Y}^{\text{mds}} = (\mathbf{Y}_A^{\text{mds}}, \mathbf{Y}_B^{\text{mds}})$, and $\mathbf{Y}^{\text{oas}} = (\mathbf{Y}_A^{\text{oas}}, \mathbf{Y}_B^{\text{oas}})$ denote matrices of item responses in FIM, MDS, and OASIS, respectively, with rows referring to subjects and columns referring to variables, and where $\mathbf{Y}_A^{\text{mds}} = (\mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{Y}_{A,\text{mis}}^{\text{mds}})$, $\mathbf{Y}_B^{\text{mds}} = (\mathbf{Y}_{B,\text{obs}}^{\text{mds}}, \mathbf{Y}_{B,\text{mis}}^{\text{mds}})$, $\mathbf{Y}_A^{\text{oas}} = (\mathbf{Y}_{A,\text{mis}}^{\text{oas}}, \mathbf{Y}_{A,\text{obs}}^{\text{oas}})$, and $\mathbf{Y}_B^{\text{oas}} = (\mathbf{Y}_{B,\text{mis}}^{\text{oas}}, \mathbf{Y}_{B,\text{obs}}^{\text{oas}})$. The subscripts A and B denote the assessments on admission and at the later date, respectively. The subscripts *obs* and *mis* denote the observed and missing assessments, respectively. In addition, let $\mathbf{X}$ denote a set of fully observed covariates.

The joint posterior distribution of the missing data and the parameters can be written as

$$
\begin{aligned}
& f\bigl(\mathbf{Y}_{A,\text{mis}}^{\text{mds}}, \mathbf{Y}_{B,\text{mis}}^{\text{mds}}, \mathbf{Y}_{A,\text{mis}}^{\text{oas}}, \mathbf{Y}_{B,\text{mis}}^{\text{oas}}, \boldsymbol{\psi}_A, \boldsymbol{\psi}_B | \\
& \quad \mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{Y}_{B,\text{obs}}^{\text{mds}}, \mathbf{Y}_{A,\text{obs}}^{\text{oas}}, \mathbf{Y}_{B,\text{obs}}^{\text{oas}}, \mathbf{Y}^{\text{fim}}, \mathbf{X}, \mathbf{M}\bigr) \\
& \quad = f\bigl(\mathbf{Y}_{A,\text{mis}}^{\text{mds}}, \mathbf{Y}_{A,\text{mis}}^{\text{oas}}, \boldsymbol{\psi}_A | \mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{Y}_{A,\text{obs}}^{\text{oas}}, \mathbf{Y}^{\text{fim}}, \mathbf{X}, \mathbf{M}\bigr) \qquad (1.1)
\end{aligned}
$$

$$\times f(\mathbf{Y}^{\mathrm{mds}}_{B,\mathrm{mis}}, \mathbf{Y}^{\mathrm{oas}}_{B,\mathrm{mis}}, \boldsymbol{\psi}_B | \mathbf{Y}^{\mathrm{mds}}_{A,\mathrm{mis}}, \mathbf{Y}^{\mathrm{oas}}_{A,\mathrm{mis}}, \mathbf{Y}^{\mathrm{mds}}_{A,\mathrm{obs}}, \mathbf{Y}^{\mathrm{mds}}_{B,\mathrm{obs}}, \mathbf{Y}^{\mathrm{oas}}_{A,\mathrm{obs}}, \mathbf{Y}^{\mathrm{oas}}_{B,\mathrm{obs}}, \mathbf{Y}^{\mathrm{fim}}, \mathbf{X}, \mathbf{M}),$$

where $\boldsymbol{\psi}_A$ and $\boldsymbol{\psi}_B$ index the imputation models in the Equating and Translating steps, respectively. The Equating step is performed once, and the Translating step can be performed multiple times. To reduce the computational complexity and to provide flexibility to researchers we assumed in equation (1.1) that $\boldsymbol{\psi}_A$ and $\boldsymbol{\psi}_B$ are conditionally independent. The data setting that consists of patients' covariates, FIM assessments, and first MDS or OASIS assessments resembles the statistical matching setup [D'Orazio, Di Zio and Scanu (2006)]. In this setup, the joint distribution of $\{\mathbf{Y}^{\mathrm{fim}}, \mathbf{Y}^{\mathrm{mds}}_A, \mathbf{Y}^{\mathrm{oas}}_A, \mathbf{X}\}$ is not identifiable based on observed data, because MDS and OASIS are never jointly observed.

This setup also arises in the test equating literature when using common-item nonequivalent groups design [Kolen and Brennan (2014), Dorans, Pommerich and Holland (2007)]. This design assumes that different groups of examinees are assessed using two different test forms that share a common item set. When used for equating, the common-item set should be representative of the total test forms in content and statistical characteristics [Kolen and Brennan (2014), Dorans, Pommerich and Holland (2007)]. This is commonly attained by ensuring that the items are exactly the same in both forms and are at the same location in the form. Here, $\mathbf{Y}^{\mathrm{fim}}$ is similar for all patients, and it is administered prior to and within a short time frame from the initial MDS and OASIS assessments. In addition, $\mathbf{Y}^{\mathrm{fim}}$ includes similar content to the MDS and OASIS assessments, because it attempts to approximate the same underlying functional status. Based on these observations, a natural starting point is to apply the conditional independence assumption, $f(\mathbf{Y}^{\mathrm{mds}}_A, \mathbf{Y}^{\mathrm{oas}}_A | \mathbf{Y}^{\mathrm{fim}}, \mathbf{X}, \mathbf{M}, \boldsymbol{\psi}_A) = f(\mathbf{Y}^{\mathrm{mds}}_A | \mathbf{Y}^{\mathrm{fim}}, \mathbf{X}, \mathbf{M}, \boldsymbol{\psi}_A) f(\mathbf{Y}^{\mathrm{oas}}_A | \mathbf{Y}^{\mathrm{fim}}, \mathbf{X}, \mathbf{M}, \boldsymbol{\psi}_A)$ [D'Orazio, Di Zio and Scanu (2006)]. This assumption is often implicitly made in test equating applications using only $\mathbf{Y}^{\mathrm{fim}}$. Here, we include other patient characteristics as well.

We further assumed that the unmeasured assessments are missing at random (MAR) [Little and Rubin (2002)], because a major determinant of patients' discharge destination from a rehabilitation facility is their functional status, which is measured using the validated FIM instrument. Under the conditional independence and the MAR assumptions, we can impute $\mathbf{Y}^{\mathrm{mds}}_{A,\mathrm{mis}}$ using the posterior distribution $f(\mathbf{Y}^{\mathrm{mds}}_{A,\mathrm{mis}} | \mathbf{Y}^{\mathrm{fim}}, \mathbf{Y}^{\mathrm{mds}}_{A,\mathrm{obs}}, \mathbf{X}, \boldsymbol{\psi}_A)$ in the Equating step. These two assumptions cannot be inferred from the data and may not always be plausible. To examine the plausibility of these assumptions, we conducted a sensitivity analysis to investigate whether our results changed in a substantial way when these assumptions are violated [Rubin (1986), Heitjan, Landis and Richard (1994)].

The Equating step generates complete synthetic data sets that comprise MDS and OASIS assessments simultaneously for patients who were discharged home.

Assuming MAR and that the relationship between contemporary imputed and observed instruments does not change across the continuum of care, we can simplify the third line of equation (1.1):

$$\begin{aligned} f&\big(\mathbf{Y}^{\text{mds}}_{B,\text{mis}}, \mathbf{Y}^{\text{oas}}_{B,\text{mis}}, \boldsymbol{\psi}_B | \mathbf{Y}^{\text{mds}}_{A,\text{mis}}, \mathbf{Y}^{\text{oas}}_{A,\text{mis}}, \mathbf{Y}^{\text{mds}}_{A,\text{obs}}, \mathbf{Y}^{\text{mds}}_{B,\text{obs}}, \mathbf{Y}^{\text{oas}}_{A,\text{obs}}, \mathbf{Y}^{\text{oas}}_{B,\text{obs}}, \mathbf{Y}^{\text{fim}}, \mathbf{X}, \mathbf{M}\big) \\ &= f\big(\mathbf{Y}^{\text{mds}}_{B,\text{mis}}, \boldsymbol{\psi}^{\text{mds}}_B | \mathbf{Y}^{\text{mds}}_{A,\text{mis}}, \mathbf{Y}^{\text{oas}}_{A,\text{obs}}, \mathbf{Y}^{\text{oas}}_{B,\text{obs}}\big) \\ &\quad \times f\big(\mathbf{Y}^{\text{oas}}_{B,\text{mis}}, \boldsymbol{\psi}^{\text{oas}}_B | \mathbf{Y}^{\text{oas}}_{A,\text{mis}}, \mathbf{Y}^{\text{mds}}_{A,\text{obs}}, \mathbf{Y}^{\text{mds}}_{B,\text{obs}}\big), \end{aligned}$$

where $\boldsymbol{\psi}_B = (\boldsymbol{\psi}^{\text{mds}}_B, \boldsymbol{\psi}^{\text{oas}}_B)$, and in the Translating step we impute $\mathbf{Y}^{\text{mds}}_{B,\text{mis}}$ using $f(\mathbf{Y}^{\text{mds}}_{B,\text{mis}} | \mathbf{Y}^{\text{mds}}_{A,\text{imp}}, \mathbf{Y}^{\text{oas}}_{A,\text{obs}}, \mathbf{Y}^{\text{oas}}_{B,\text{obs}}, \boldsymbol{\psi}^{\text{mds}}_B)$, where $\mathbf{Y}^{\text{mds}}_{A,\text{imp}}$ denotes the imputed MDS assessments at admission.

1.4. *Related work.* The Equating and the Translating steps require methods that impute multivariate ordinal variables. Multivariate imputation methods can be classified into two types of methods: fully conditional specification and joint modeling. Fully conditional specification [van Buuren (2007)] involves a series of univariate conditional models that impute missing values sequentially with current model estimates. In practice, users only include main effects in these models, because it is challenging to identify and include higher-order interactions and nonlinear terms at each of the conditional models [Vermunt et al. (2008)]. With multiple ordinal variables, the default implementation of fully conditional specification relies on the ordered logit model. Gu and Gutman (2017) noted that this implementation fails to capture the full correlation structure of the imputed items when the proportional odds assumption is violated. Recently, a multi-level model based on the probit link function was proposed as a possible imputation model for the missing ordinal variable [Enders, Keller and Levy (2018)].

The joint modeling approach [Schafer (1997)] specifies a joint probability model for all the data. Imputation of missing values is performed from the implied distribution of the missing variables conditional on the observed data. Yucel, He and Zaslavsky (2011) proposed a method that is based on the multivariate normal model to impute ordinal variables and supplemented it with a rounding technique that preserves the observed marginal distribution of the ordinal variables. When there is a large proportion of missing values, propagation of errors in the underlying modeling approximation can compound and result in invalid statistical inferences [Yucel, He and Zaslavsky (2011), Gu and Gutman (2017)].

Imputation by Propensity score matching (IPSM) can be embedded in a joint modeling approach to define cells within which hot-deck imputations can be drawn [Andridge and Little (2009)]. The propensity score is defined as the probability of a unit to have missing values. IPSM imputes missing values with observed values from units with similar estimated propensity scores. IPSM is a generally valid statistical method, but its performance is sensitive to the specification of the propensity score model [Gu and Gutman (2017)].

Latent variable matching (LVM) [Gu and Gutman (2017)] is a recently proposed procedure that combines IRT models with multiple imputation [Rubin (1987)] to impute unmeasured assessments. LVM is also a hot-deck imputation method, which matches units using the underlying functional status estimated from IRT models. In its original form, LVM ignores patient covariates, which may violate the MAR assumption [Rubin (1994)]. LVM can be extended to account for a set of discrete and continuous covariates by applying it within subgroups of the covariates; however, this approach may become computationally intensive when the number of possible covariate values is large.

Among these methods, IPSM and LVM are the strongest candidate methods in terms of validity and efficiency for imputing the missing assessments in our datasets [Gu and Gutman (2017)]. Thus, we only compared these two methods with the newly proposed procedure in Section 4.

2. The multivariate ordinal probit model. Let $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}) = \{y_{ij}, i = 1, \dots, N, j = 1, \dots, J\}$ denote a generic matrix of item responses, where $\mathbf{Y}_{\text{mis}}$ and $\mathbf{Y}_{\text{obs}}$ are the matrices of missing and observed item responses, respectively, $y_{ij} \in \{1, \dots, c_j\}$ is the response of patient i to item j, and c_j is the number of response levels of item j. For example, in the Equating step, $\mathbf{Y}_{\text{mis}}$ corresponds to the unmeasured assessments in MDS, $\mathbf{Y}^{\text{mds}}_{A,\text{mis}}$, and $\mathbf{Y}_{\text{obs}}$ corresponds to the observed assessments in FIM and MDS, $\mathbf{Y}^{\text{fim}}$ and $\mathbf{Y}^{\text{mds}}_{A,\text{obs}}$. The MVOP model introduces a matrix of latent response variables $\mathbf{Z} = \{z_{ij}, i = 1, \dots, N, j = 1, \dots, J\}$ such that $y_{ij} = g_j(z_{ij}) = l$, if $\gamma_{j,l-1} < z_{ij} \le \gamma_{j,l}$, where $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,c_j-1} < \gamma_{j,c_j} = +\infty$ are unknown threshold parameters. The MVOP model assumes that $\mathbf{z}_i = (z_{i1}, \dots, z_{iJ})^\top \sim \mathcal{N}(\boldsymbol{\beta}\mathbf{x}_i, \boldsymbol{\Sigma})$, $i = 1, \dots, N$, where $\mathbf{x}_i$ is a $P \times 1$ vector of covariates for patient i, $\boldsymbol{\beta}$ is a $J \times P$ matrix of unknown regression coefficients and $\boldsymbol{\Sigma}$ is an unknown covariance matrix. Statistical inferences for $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ are based on the likelihood

$$
\begin{aligned}
L(\boldsymbol{\psi} | \mathbf{Y}_{\text{obs}}, \mathbf{X}) &= c \prod_{i=1}^{N} \int f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\psi})\, d\mathbf{y}_{\text{mis},i} \\
&= c \prod_{i=1}^{N} \int \int_{\Gamma_{iJ}} \cdots \int_{\Gamma_{i1}} \mathcal{N}_J(\mathbf{z}_i; \boldsymbol{\beta}\mathbf{x}_i, \boldsymbol{\Sigma})\, d\mathbf{z}_i\, d\mathbf{y}_{\text{mis},i},
\end{aligned}
\tag{2.1}
$$

where Γ_{ij} is the interval $(\gamma_{i,l-1}, \gamma_{i,l}]$ if $y_{ij} = g_j(z_{ij}) = l$, and c is a normalizing constant.

The vector parameter $\boldsymbol{\psi}$ is not identifiable because the likelihood (2.1) is invariant to location and scale transformations on $\mathbf{Z}$. Threshold constraints and correlation constraints are two types of identification constraints that are commonly made with the MVOP model. The threshold constraints fix two threshold parameters for each outcome. For example, one could set $\gamma_{j,1} = 0$ and either $\gamma_{j,2} = 1$ or $\gamma_{j,c_j-1} = 1$ $\forall j$. Applying these constraints allows us to sample the covariance matrix $\mathbf{\Sigma}$ from a known probability distribution [Chen and Dey (2000), Jeliazkov, Graves and Kutzbach (2008)]. The correlation constraints either fix $\gamma_{j,1} = 0$ or $\beta_{j,1} = 0\ \forall j$, and restrict $\mathbf{\Sigma}$ to be a correlation matrix $\mathbf{R}$ [Lawrence et al. (2008), Zhang et al. (2017)]. We refer to the MVOP model under the threshold constraints and correlation constraints as the MVOPT model and MVOPC model, respectively.

MVOP models have been analyzed using likelihood-based methods, including a direct likelihood approach that involves the evaluation of integrals using Gaussian–Hermite quadrature [Li and Schafer (2008)], an approximate EM algorithm [Guo et al. (2015)], a pseudo-likelihood approach [Varin and Czado (2010)] and Bayesian approaches [Chen and Dey (2000), Lawrence et al. (2008), Zhang et al. (2017)]. Of these methods, only Zhang et al. (2017) extended the MVOP model to handle incomplete correlated ordinal responses. Here, when some of the outcomes are missing, we propose Monte Carlo EM (MCEM) algorithms [Wei and Tanner (1990)] for maximum likelihood estimation and DA algorithms for Bayesian inference under the MVOP models to produce imputed values.

2.1. *MCEM algorithm.* We first consider the MVOPT model, and fix $\gamma_{j,1} = 0$ and $\gamma_{j,c_j-1} = 1\ \forall j$. The complete-data likelihood is

$$L_{\text{com}}(\boldsymbol{\psi}|\mathbf{Y},\mathbf{X},\mathbf{Z}) \propto |\mathbf{\Sigma}|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{\Sigma}^{-1}\sum_{i=1}^{N}(\mathbf{z}_i-\boldsymbol{\beta}\mathbf{x}_i)(\mathbf{z}_i-\boldsymbol{\beta}\mathbf{x}_i)^\top\right)\right\}$$
$$\times \prod_{i=1}^{N}\prod_{j=1}^{J}\mathbf{1}\{z_{ij}\in\Gamma_{ij}\}.$$

The E-step of the EM algorithm, given the current value of the parameter, $\boldsymbol{\psi}$, involves evaluating the expectation

$$\begin{aligned} Q(\boldsymbol{\psi}^*|\boldsymbol{\psi}) &= E\{\log L(\boldsymbol{\psi}^*|\mathbf{Y},\mathbf{X},\mathbf{Z})|\mathbf{Y}_{\text{obs}},\mathbf{X},\boldsymbol{\psi}\} \\ &= \int \log L(\boldsymbol{\psi}^*|\mathbf{Y},\mathbf{X},\mathbf{Z}) f(\mathbf{Z}|\mathbf{Y}_{\text{obs}},\mathbf{X},\boldsymbol{\psi})\,d\mathbf{Z}, \end{aligned}$$

which consists of multiple integrations with respect to a truncated multivariate normal distribution of $\mathbf{Z}$. $Q(\boldsymbol{\psi}^*|\boldsymbol{\psi})$ cannot be calculated analytically, but Monte Carlo methods can be used to approximate it. We extend the slice sampler algorithm proposed by Damien and Walker (2001) for bivariate normal distribution to sample from the truncated multivariate normal distribution. The algorithm introduces a latent variable so that the slice sampler runs on a sequence of conditional distributions which can all be sampled directly using uniform distributions. This algorithm has a faster mixing rate than the Gibbs sampling algorithm [Geweke (1991)]. Details of the algorithm are described in Section 1 of the Supplementary Material [Gu and Gutman (2019)].

In the M-step, we rely on conditional maximization [Meng and Rubin (1993)] to update $Q(\boldsymbol\psi^*|\boldsymbol\psi)$ in successive steps with respect to $\boldsymbol\beta$ and $\boldsymbol\Sigma$:

$$\hat{\boldsymbol\beta} = \sum_{i=1}^{N}\left\{\frac{1}{G}\sum_{g=1}^{G}\tilde{\mathbf{z}}_i^{(g)}\mathbf{x}_i^\top\right\}\left\{\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right\}^{-1},$$

$$\hat{\boldsymbol\Sigma} = \frac{1}{N}\left\{\sum_{i=1}^{N}\left\{\frac{1}{G}\sum_{g=1}^{G}\tilde{\mathbf{z}}_i^{(g)}\tilde{\mathbf{z}}_i^{(g)\top}\right\} - \hat{\boldsymbol\beta}\left\{\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\right\}\hat{\boldsymbol\beta}^\top\right\},$$

where $\{\tilde{\mathbf{z}}^{(g)}, g = 1, \ldots, G\}$ are G draws from $f(\mathbf{Z}|\mathbf{Y}_{\text{obs}}, \mathbf{X}, \boldsymbol\psi)$. To decrease the Monte Carlo errors, Wei and Tanner (1990) suggested using a large G.

To complete the estimation process, we derived a consistent estimator for $\boldsymbol\gamma$. The estimator is based on the empirical marginal distribution of the observed and imputed responses in the absence of threshold constraints [Olsson (1979)],

$$\tilde\gamma_{j,l} = \frac{1}{G}\sum_{g=1}^{G}\Phi^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\tilde y_{ij}^{(g)} \in (-\infty, l)\}\right\},$$

$$l = 1, \ldots, c_j - 1,\ j = 1, \ldots, J,$$

where $\tilde y_{ij}^{(g)} = y_{\text{obs},ij}$ if $M_i = 0$, $\tilde y_{ij}^{(g)} = \tilde y_{\text{imp},ij}^{(g)}$ and $\tilde y_{\text{imp},ij}^{(g)}$ is imputed through the indicator function $\mathbf{1}\{\tilde z_{ij}^{(g)} \in \Gamma_{ij}\}$ given the current estimate of $\boldsymbol\gamma_j$ if $M_i = 1$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution. The estimate of $\gamma_{j,l}$ given the threshold constraints is

$$\hat\gamma_{j,l} = \frac{\tilde\gamma_{j,l} - \min\tilde{\boldsymbol\gamma}_j}{\max\tilde{\boldsymbol\gamma}_j - \min\tilde{\boldsymbol\gamma}_j},$$

where $\tilde{\boldsymbol\gamma}_j = (\tilde\gamma_{j,1}, \ldots, \tilde\gamma_{j,c_j-1})$.

2.2. *Data augmentation algorithm.* For Bayesian inference of the MVOPT model, we assign a $\mathcal{N}(\mathbf{0}, 10^4 \times \mathbf{I})$ prior distribution for $\boldsymbol\beta$ and a $\mathcal{IW}(J+2, (J+2)\times\mathbf{I}_{J\times J})$ prior distribution for $\boldsymbol\Sigma$, where $\mathcal{IW}(\nu, S_0)$ denotes the inverse-Wishart distribution with ν degrees of freedom and scale matrix S_0. Based on the work of Albert and Chib (1993), we use a uniform prior distribution over the polytope $\mathcal{T} \subset \mathbb{R}^{c_j}$ for $\boldsymbol\gamma_j$, $j = 1, \ldots, J$. The feasible region for the parameter space of $\boldsymbol\gamma_j$:

$$\mathcal{T} = \{\boldsymbol\gamma_j = (\gamma_{j,2}, \ldots, \gamma_{j,c_j-1}) \in \mathbb{R}^{c_j} : \gamma_{j,l} > \gamma_{j,l-1}, \forall\, l = 2, \ldots, c_j - 1\}.$$

The DA algorithm for drawing samples from the posterior distribution of $\boldsymbol\psi$ consists of an Imputation step that draws $\mathbf{Z}$ from $f(\mathbf{Z}|\mathbf{Y}_{\text{obs}}, \mathbf{X}, \boldsymbol\psi)$ using the slice sampler algorithm described in Section 2.1, and three Posterior simulation (P) steps:

P-step 1: Draw $\widetilde{\boldsymbol\beta}|\mathbf{Z}, \boldsymbol\Sigma, \mathbf{X} \sim \mathcal{N}(\boldsymbol\mu_\beta, \boldsymbol\Sigma_\beta)$, where $\widetilde{\boldsymbol\beta} = (\boldsymbol\beta_1, \ldots, \boldsymbol\beta_J)^\top$, $\boldsymbol\beta_j$ is the jth row of $\boldsymbol\beta$, $\widetilde{\mathbf{X}}_i = \mathbf{I}_{J\times J} \otimes \mathbf{x}_i$, $\boldsymbol\Sigma_\beta = (\sum_{i=1}^N \widetilde{\mathbf{X}}_i^\top \boldsymbol\Sigma^{-1}\widetilde{\mathbf{X}}_i + 10^{-4} \times \mathbf{I}_{JP\times JP})^{-1}$ and $\boldsymbol\mu_\beta = \boldsymbol\Sigma_\beta \sum_{i=1}^N \widetilde{\mathbf{X}}_i^\top \boldsymbol\Sigma^{-1}\mathbf{z}_i$.

P-step 2: Draw $\boldsymbol{\Sigma}|\mathbf{Z},\mathbf{X} \sim \mathcal{IW}(N+J+2, \sum_{i=1}^{N}(\mathbf{z}_i - \boldsymbol{\beta}\mathbf{x}_i)(\mathbf{z}_i - \boldsymbol{\beta}\mathbf{x}_i)^\top + (J+2) \times \mathbf{I}_{J\times J})$.

P-step 3: Draw $\gamma_{jl}|\mathbf{z}_j, \mathbf{y}_{\text{obs},j} \sim \mathcal{U}(\max\{\max\{z_{ij} : y_{ij} = l\}, \gamma_{j,l-1}\}, \min\{\min\{z_{ij} : y_{ij} = l+1\}, \gamma_{j,l+1}\})$, for $l = 2, \dots, c_j - 2$, and $j = 1, \dots, J$, where $\mathcal{U}(a, b)$ denotes the uniform distribution with support (a, b).

After each cycle of the algorithm, we impute the missing responses $\mathbf{Y}_{\text{mis}}$ through the indicator functions $\mathbf{1}\{z_{ij} \in \Gamma_{ij}\}$, for $i = 1, \dots, N$ and $j = 1, \dots, J$ given the corresponding latent responses $\mathbf{Z}$ and threshold parameters $\boldsymbol{\gamma}$.

2.3. *Parameter expansion approach.* For the MVOPC model, we fix $\gamma_{j,1} = 0$ $\forall j$ and constrain the covariance matrix $\boldsymbol{\Sigma}$ to be a correlation matrix $\mathbf{R}$. In the MCEM algorithm, the M-step with respect to $\mathbf{R}$ does not have a closed form solution [Chib and Greenberg (1998)], and direct maximization of the expectation of the complete-data likelihood is computationally intensive. For the Bayesian inference, the posterior distribution of $\mathbf{R}$ does not follow a known probability distribution. Thus, we use the parameter expansion (PX) technique [Liu, Rubin and Wu (1998), Liu and Wu (1999)], and propose a PX-MCEM algorithm to obtain the maximum likelihood estimates, and a PX-DA algorithm to sample from the posterior distribution of $\boldsymbol{\psi}$, respectively. These algorithms are similar to the work of Zhang, Boscardin and Belin (2006), Lawrence et al. (2008) and Zhang et al. (2017).

We consider the following transformations:

$$\mathbf{z}_i^* = \mathbf{D}\mathbf{z}_i, \qquad \boldsymbol{\beta}^* = \mathbf{D}\boldsymbol{\beta}, \qquad \mathbf{R}^* = \mathbf{DRD}, \qquad \boldsymbol{\gamma}_j^* = d_j\boldsymbol{\gamma}_j, \tag{2.2}$$

so that $\mathbf{R}$ is transformed into a general covariance matrix, where $\mathbf{D} = \text{diag}(d_1, \dots, d_J)$ is a diagonal matrix with diagonal elements $d_j > 0$ $\forall j$. The PX-MCEM and the PX-DA algorithms using the transformed parameters $(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*, \mathbf{R}^*)$ and the latent responses $\mathbf{Z}^*$ proceed as the MCEM and DA algorithms described in Section 2.1 and Section 2.2, respectively. After each iteration, $(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{R})$ are updated via the inverse transformations of identities (2.2).

3. Nested multiple imputation procedure. Let $Q = Q(\mathbf{Y}_A^{\text{mds}}, \mathbf{Y}_B^{\text{mds}})$ be a quantity of interest. We summarize the proposed procedure to multiply impute $(\mathbf{Y}_{A,\text{mis}}^{\text{mds}}, \mathbf{Y}_{B,\text{mis}}^{\text{mds}})$:

Equating: Impute $\mathbf{Y}_{A,\text{mis}}^{\text{mds}}$ from the predictive distribution $f(\mathbf{Y}_{A,\text{mis}}^{\text{mds}}|\mathbf{Y}^{\text{fim}}, \mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{X})$.

Step 1: Draw K independent parameters $\tilde{\boldsymbol{\psi}}_A$ from the posterior distribution $p(\boldsymbol{\psi}_A|\mathbf{Y}^{\text{fim}}, \mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{X})$, or from the asymptotic distribution obtained by applying the EM algorithm to a bootstrapped sample of the cases.

Step 2: Impute $\mathbf{Y}_{A,\text{mis}}^{\text{mds}}$ through the indicator functions $\mathbf{1}\{\tilde{z}_{ij} \in \Gamma_{ij}\}$, where Γ_{ij} is determined by $\tilde{\boldsymbol{\psi}}_A$ and $\tilde{z}_{ij} \sim f(z_{ij}|\mathbf{Y}^{\text{fim}}, \mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{X}, \tilde{\boldsymbol{\psi}}_A)$, $\forall i, j$.

Step 3: Repeat steps 1 and 2 K times to create K imputed datasets $\mathbf{Y}_{A,\text{imp}}^{\text{mds},(k)}$, $k = 1, \dots, K$.

Translating: For each of the K imputed datasets in Stage 1, impute $\mathbf{Y}_{B,\text{mis}}^{\text{mds}}$ from the predictive distribution $f(\mathbf{Y}_{B,\text{mis}}^{\text{mds}} | \mathbf{Y}_{A,\text{imp}}^{\text{mds}}, \mathbf{Y}_{A,\text{obs}}^{\text{oas}}, \mathbf{Y}_{B,\text{obs}}^{\text{oas}})$.

Step 4: Draw L independent parameters $\tilde{\boldsymbol{\psi}}_B^{(k)}$ from the posterior distribution $p(\boldsymbol{\psi}_B | \mathbf{Y}_{A,\text{obs}}^{\text{oas}}, \mathbf{Y}_{A,\text{imp}}^{\text{mds},(k)})$, or from the asymptotic distribution obtained by applying the EM algorithm to a bootstrapped sample of the cases.

Step 5: Impute $\mathbf{Y}_{B,\text{mis}}^{\text{mds}}$ through the indicator functions $\mathbf{1}\{\tilde{z}_{ij} \in \Gamma_{ij}\}$, where Γ_{ij} is determined by $\tilde{\boldsymbol{\psi}}_B^{(k)}$ and $\tilde{z}_{ij} \sim f(z_{ij} | \mathbf{Y}_{A,\text{imp}}^{\text{mds},(k)}, \mathbf{Y}_{A,\text{obs}}^{\text{oas}}, \mathbf{Y}_{B,\text{obs}}^{\text{oas}}, \tilde{\boldsymbol{\psi}}_B^{(k)})$, $\forall i, j$.

Step 6: Repeat steps 4 and 5 L times to create L imputed datasets $\mathbf{Y}_{B,\text{imp}}^{\text{mds},(k,l)}$, $l = 1, \dots, L$.

Combining rules: The estimate of Q and its sampling variance are $\hat{Q}^{(k,l)} = \hat{Q}^{(k,l)}(\mathbf{Y}_{\text{com}}^{(k,l)})$ and $U^{(k,l)} = U^{(k,l)}(\mathbf{Y}_{\text{com}}^{(k,l)})$ respectively, where each of the complete datasets $\mathbf{Y}_{\text{com}}^{(k,l)} = (\mathbf{Y}_{A,\text{obs}}^{\text{mds}}, \mathbf{Y}_{B,\text{obs}}^{\text{mds}}, \mathbf{Y}_{A,\text{imp}}^{\text{mds},(k)}, \mathbf{Y}_{B,\text{imp}}^{\text{mds},(k,l)})$, for $k = 1, \dots, K$, and $l = 1, \dots, L$. The overall estimate of Q and its sampling variance are obtained using the nested multiple imputation combining rule, confidence intervals and significance tests are based on a Student-t reference distribution [Shen (2000), Harel (2003), Rubin (2003)].

4. Simulation study.

We examined the performance of the MVOP model in comparison to existing methods for imputing incomplete multivariate ordinal variables using a simulation study.

4.1. *Partially simulated data.*

The simulation study was based on observed FIM assessments and MDS assessments on admission for patients in SNFs, and missing MDS assessments were artificially generated. To generate incomplete data sets, we fitted a logistic regression model to the entire dataset where the explanatory variables comprised $\mathbf{Y}^{\text{fim}}$, patients' age and patients' gender,

$$\text{logit}\{\Pr(M_i = 1 | \mathbf{y}_i^{\text{fim}}, \mathbf{x}_i)\} = \alpha_0 + \sum_{j=1}^{J_1} \alpha_j y_{ij}^{\text{fim}} + \alpha_{J_1+1} x_{i1} + \alpha_{J_1+2} x_{i2}, \tag{4.1}$$

where $\mathbf{y}_i^{\text{fim}} = (y_{i1}^{\text{fim}}, \dots, y_{iJ_1}^{\text{fim}})$, $\mathbf{x}_i = (x_{i1}, x_{i2})$, and x_{i1} and x_{i2} denote the age and gender of patient i, respectively. This resulted in estimated regression coefficients $\hat{\boldsymbol{\alpha}}' = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_1')$, where $\hat{\alpha}_0$ is the estimated intercept and $\hat{\boldsymbol{\alpha}}_1 = (\hat{\alpha}_1, \dots, \hat{\alpha}_{J_1+2})$ is a vector of estimated regression coefficients for the other predictors. A simple random sample of $n = 1000$ patients was then drawn from the set of patients in SNFs, and M_i $(i = 1, \dots, n)$ was sampled from a Bernoulli distribution with probability $\Pr(M_i = 1 | \mathbf{y}_i^{\text{fim}}, \mathbf{x}_i) = F(\tilde{\alpha}_0 + \sum_{j=1}^{J_1} \hat{\alpha}_j y_{ij}^{\text{fim}} + \hat{\alpha}_{J_1+1} x_{i1} + \hat{\alpha}_{J_1+2} x_{i2})$, where $F(\cdot)$

is the c.d.f. of a specified distribution and $F^{-1}(\cdot)$ is the link function [McCullagh and Nelder (1989)]. We considered three choices of $F(\cdot)$, the logistic distribution, the Cauchy distribution, which is symmetric but has heavier tails than the logistic distribution, and the Box–Cox family distributions [Guerrero and Johnson (1982)]. The Box–Cox distribution takes the form

$$F_\lambda(x) = \begin{cases} 0, & x < -\dfrac{1}{\lambda}, \lambda > 0, \\ \dfrac{(1+\lambda x)^{1/\lambda}}{(1+\lambda x)^{1/\lambda}+1}, & 1+\lambda x > 0, \lambda \neq 0, \\ \dfrac{\exp(x)}{1+\exp(x)}, & \lambda = 0, \\ 1, & x > -\dfrac{1}{\lambda}, \lambda < 0. \end{cases}$$

This distribution allows us to assess the effect of skewness in the missing data mechanism. It is positively skewed for $\lambda > 0$ and negatively skewed for $\lambda < 0$; here, λ was fixed at either -0.3 or 0.3. The value of $\tilde{\alpha}_0$ was fixed so that $p_{\text{mis}} = n_1/n$ is either 20%, 40% or 60%, where n_1 is the number of patients who have missing assessments. MDS assessments for patient i were deleted to create an incomplete data set when $M_i = 1$. For each configuration, 1000 replications were produced.

The methods examined in the simulations were IPSM, LVM and the MVOP models implemented using both EM and DA algorithms: MVOPT-DA, MVOPT-EM, MVOPC-DA and MVOPC-EM. For IPSM, we estimated the propensity score using the logistic regression model: $\text{logit}\{e(\mathbf{y}_i^{\text{fim}}, \mathbf{x}_i)\} = \xi_0 + \sum_{j=1}^{J_1} \xi_j y_{ij}^{\text{fim}} + \xi_{J_1+1} x_{i1} + \xi_{J_1+2} x_{i2}$. For both IPSM and LVM, we used the nearest-neighbor matching algorithm to find a potential donor. Ten multiple imputations were generated using each of the methods.

We examined the performance of the different methods on two estimands: (1) the population mean total score of items in MDS, $Q_1 \equiv E(S^{\text{mds}})$, where $S_i^{\text{mds}} = \sum_j y_{ij}^{\text{mds}}$, $i = 1, \dots, n$; (2) the pairwise Goodman and Kruskal's γ rank correlation coefficients [Goodman and Kruskal (1979)] between J_{mds} items in MDS, $Q_2 \equiv \{\gamma(\mathbf{y}_j^{\text{mds}}, \mathbf{y}_k^{\text{mds}}), 1 \leq j < k \leq J_{\text{mds}}\}$. For each method, at each configuration, and at each of the 1000 replications, we recorded $\hat{Q}_m$, $m = 1, 2$, their estimated sampling variances, the corresponding root mean square errors (RMSEs), the 95% interval estimate widths, and determined whether the intervals covered or did not cover the true value. Using these values, we calculated for each approach and each configuration, the average coverage rate, the bias, the mean estimated sampling variance, the mean RMSE, and the mean interval width. Because the simulations are based on 1000 replicates for each configuration, observed coverage of 93.7% or above is not statistically distinguishable from the nominal level. In addition, we view observed coverage of 90% as indicative of a modest deficit in coverage.

For each configuration, we also calculated a loss function based on the negatively oriented interval scores [Gneiting and Raftery (2007), equation (61)]. This loss function provides flexible assessment of coverage by accounting for the distance between the interval estimate and the estimand. For estimand Q_m, the loss function for interval estimate I, has the form

$$\lambda(I) + \frac{2}{\alpha} \inf_{\eta \in I} |Q_m - \eta|, \tag{4.2}$$

where $\alpha = 0.05$ and $\lambda(I)$ denotes the Lebesgue measure of the interval estimate I.

The simulations were implemented using R 3.1.0 [R Core Team (2014)]. The proposed EM and DA algorithms were implemented in C++ for efficiency. For the EM algorithms, we generated $G = 100$ samples from $f(\mathbf{Z}|\mathbf{Y}_{\text{obs}}, \mathbf{X}, \boldsymbol{\psi})$ in the E-step, and calculated the observed-data likelihood using a Monte Carlo method [Genz (1992)] to monitor the convergence of the MCEM algorithm. For the DA algorithms, multiple parallel chains of 50,000 iterations with dispersed initial values were generated. Standard MCMC convergence diagnostics such as Gelman–Rubin Statistic [Gelman and Rubin (1992)], trace plots and autocorrelation plots were examined for a small sample of the simulations, and did not indicate failure to converge.

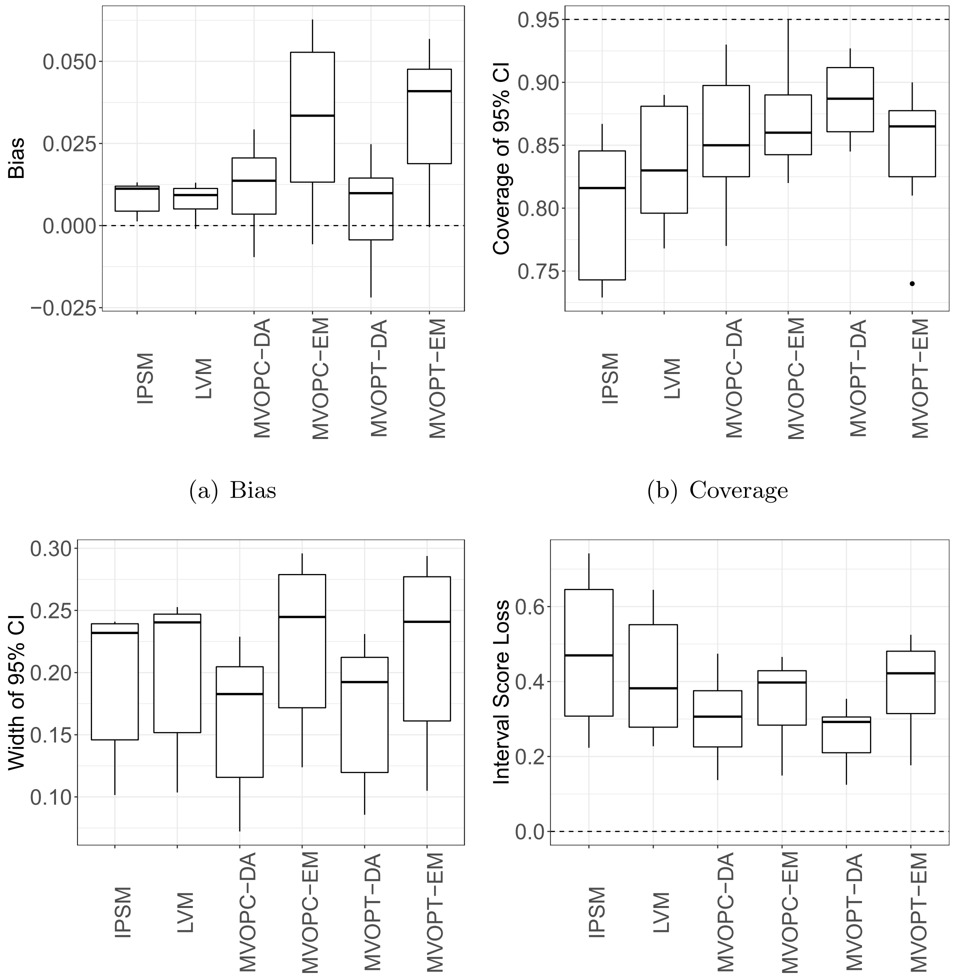
4.2. *Results*. Table 1 displays the mean biases, variances, RMSEs, coverages, interval widths and interval estimate loss function of the population mean total score of items in MDS, Q_1, for configurations, where $F(\cdot)$ is the logistic distribution and $p_{\text{mis}} = \{0.2, 0.4, 0.6\}$. Although some methods show modest deficits in coverage in some scenarios, all of the methods yield coverage that is generally either at or above the nominal level, statistically indistinguishable from the nominal level, or indicative of only a modest deficit in coverage. Compared to all of the methods that were examined, MVOP models implemented using the DA algorithms have coverages that are closest to nominal across all configurations. IPSM has coverages that are slightly smaller than LVM for $p_{\text{mis}} = 0.2, 0.4$, and worse than LVM for $p_{\text{mis}} = 0.6$. When $p_{\text{mis}} = 0.6$, the parametric models underlying LVM impute the missing values with less bias than the propensity score model used in IPSM. Similar results were observed when predictive mean matching was compared to IPSM [Andridge and Little (2009)]. The MVOP models implemented using the DA algorithms generally have the smallest biases and RMSEs, while the MVOP models implemented using the EM algorithms and the bootstrap method have the largest biases, variances, RMSEs and interval widths. Because some methods have lower coverage but with shorter intervals, and some have higher coverage with wider intervals, we used the loss function in equation (4.2) to compare the methods on coverage and interval width simultaneously. Generally, the MVOPT models implemented using the DA algorithms have the smallest interval score loss followed by LVM.

TABLE 1
Biases, variances, RMSEs, 95% interval coverages, 95% confidence interval widths and interval estimate loss function [Equation (4.2)] for the population mean total score of items in MDS, Q_1, given that $n = 1000$ and $F(\cdot)$ is the logistic distribution

p_{mis}	Method	Bias	Variance	RMSE	Coverage	Width	Equation (4.2)
0.2	IPSM[a]	−0.013	0.014	0.124	93.2	0.460	0.607
	LVM[b]	−0.047	0.014	0.124	93.9	0.459	0.588
	MVOPT[c]-DA[d]	−0.015	0.014	0.116	95.5	0.472	0.547
	MVOPC[e]-DA	−0.012	0.013	0.115	94.3	0.453	0.569
	MVOPT-EM[f]	−0.017	0.029	0.133	98.0	0.735	0.678
	MVOPC-EM	−0.032	0.030	0.130	97.8	0.752	0.677
0.4	IPSM	−0.015	0.016	0.147	90.4	0.500	0.752
	LVM	−0.070	0.020	0.159	91.8	0.565	0.767
	MVOPT-DA	−0.030	0.025	0.139	96.1	0.641	0.715
	MVOPC-DA	−0.025	0.020	0.141	94.8	0.564	0.793
	MVOPT-EM	−0.070	0.039	0.175	97.3	0.885	0.823
	MVOPC-EM	−0.073	0.039	0.166	97.2	0.886	0.836
0.6	IPSM	0.062	0.034	0.228	89.6	0.748	1.191
	LVM	−0.050	0.035	0.196	92.2	0.757	1.007
	MVOPT-DA	−0.027	0.050	0.183	97.9	0.931	0.984
	MVOPC-DA	−0.007	0.033	0.179	95.3	0.749	0.871
	MVOPT-EM	−0.180	0.061	0.269	91.7	1.137	1.084
	MVOPC-EM	−0.167	0.063	0.263	92.4	1.164	1.024

[a] IPSM: imputation by propensity score matching;
[b] LVM: latent variable matching;
[c] MVOPT: multivariate ordinal probit model with threshold constraints;
[d] DA: data augmentation algorithm;
[e] MVOPC: multivariate ordinal probit model with correlation constraints;
[f] EM: expectation-maximization algorithm.

Figure 1 displays the distribution of biases, 95% interval coverages, interval widths and interval score loss of the pairwise rank correlation coefficients between items in MDS, Q_2, for configurations where $F(\cdot)$ is the logistic distribution and $p_{\text{mis}} = 0.6$. The MVOP models except for MVOPC-DA have coverages that are close to nominal, while IPSM and LVM have median coverage that is lower than 85%. However, except for MVOPT-DA, the other MVOP models have biases that are larger than IPSM and LVM. As with Q_1, MVOPT-EM and MVOPC-EM have the largest biases and interval lengths, but their coverages are closer to nominal when compared to LVM and IPSM. Lastly, MVOPT-DA has better coverages and smaller interval score loss than MVOPC-DA. These trends are similar to the ones observed with $p_{\text{mis}} = \{0.2, 0.4\}$ [see Figures 5–6 in Section 2 of the Supplementary Material [Gu and Gutman (2019)]].



(a) Bias

(b) Coverage

(c) CI Width

(d) Interval Score Loss

FIG. 1. *Distribution across* 1000 *simulation replications of* (a) *biases*, (b) *coverages of* 95% *confidence interval*, (c) *widths of* 95% *confidence interval and* (d) *interval score loss for the pairwise rank correlation coefficients between items in MDS*, Q_2, *given that* $F(\cdot)$ *is the logistic distribution and* $p_{\text{mis}} = 0.6$.

Because MVOPT-DA generally has the best operating characteristics when $F(\cdot)$ is the logistic distribution, we only include this method when examining the effects of propensity score model misspecification. Table 1 in Section 2 of the Supplementary Material [Gu and Gutman (2019)] displays the results for the population mean total score of items in MDS when $F(\cdot)$ is the Cauchy distribution or the Box–Cox family distribution with $\lambda = \{-0.3, 0.3\}$. The performance of IPSM is sensitive to misspecification of the propensity score model. For example, when $p_{\text{mis}} = 0.6$ and $F(\cdot)$ is the Box–Cox family distribution with $\lambda = -0.3$, the coverage of IPSM is only 82% and its interval score loss is larger than LVM and MVOPT-DA. In contrast, LVM and MVOPT-DA are robust to different link functions. MVOPT-DA has better coverages and smaller biases than LVM across all of the configurations that were examined, and generally has smaller interval score loss than LVM. Figures 7–15 in Section 2 of the Supplementary Material [Gu and Gutman (2019)] display the results for the pairwise rank correlation coefficients between items in MDS when $F(\cdot)$ is the Cauchy distribution and the Box–Cox family distribution with $\lambda = \{-0.3, 0.3\}$, respectively. IPSM, LVM and MVOPT-DA have similar point estimates, but MVOPT-DA has better coverages and smaller interval score loss than LVM and IPSM in most of the examined configurations. IPSM has the lowest coverages, and the median of its coverages is about 72% when $F(\cdot)$ is the Box–Cox distribution with $\lambda = -0.3$ and $p_{\text{mis}} = 0.6$. When the percentage of missingness decreases, the coverages of MVOPT-DA are closer to nominal.

4.3. *Sensitivity of the methods to the conditional independence and MAR assumptions.* The proposed methods rely on the validity of the conditional independence and MAR assumptions (Section 1.3). We conducted an additional simulation study to examine the plausibility of these assumptions in this analysis.

One clinical variable that is recorded for patients in IRFs is their swallowing status at discharge. Swallowing status is a categorical variable with three categories: “Regular Food”, “Modified Food Consistency/Supervision” and “Tube/Parenteral”. Swallowing status is correlated with patients’ self-care functional status as well as patients’ discharge destination. We recoded the swallowing status using two dummy variables, which were added to equation (4.1). The setup of the coefficients in equation (4.1) was similar to the one in Section 4.1, and we also considered the three different link functions and three possible values for p_{mis}. Because MVOPT-DA has the best operating characteristics, we only examined the validity of the conditional independence and MAR assumptions with this model. Swallowing status was not included when fitting the MVOPT-DA model to address the possibility that physicians may have used unobserved clinical information when selecting between the two possible discharge destinations. The fitted MVOP model potentially violates both the conditional independence and the MAR assumptions.

Table 2 and Figure 16–19 in Section 2 of the Supplementary Material [Gu and Gutman (2019)] display the results for the population mean total score of items in MDS and for the pairwise rank correlation coefficients between items in MDS, respectively. MVOPT-DA generally provides statistically valid inferences. When the percentage of missingness decreases, the biases, variances, RMSEs and interval widths of MVOPT-DA decrease.

5. Motivating example revisited.

5.1. *Data.* FIM, MDS and OASIS include similar functional status items, but they have differences in the rating levels (i.e., "independence" is reflected by a higher score in FIM but a lower score in MDS). To increase the consistency of the items in these three instruments, we reversed the rating levels of FIM prior to the analysis such that in all three instruments lower rating levels represent better functional status. In addition, we recoded any MDS items with score of 7 or 8 (activity occurred only once or twice or activity did not occur) as a score of 4 (totally dependent) [Wysocki, Thomas and Mor (2015)]. We also combined the scores 3, 4 and 5 in the item "Feeding or Eating" in OASIS due to a small proportion ($<1\%$) of patients responding at these levels. After recoding, the items in FIM, MDS and OASIS have seven, five and four rating levels, respectively, except for the item "bathing" in OASIS that has seven levels.

Patients' demographic characteristics are summarized in Table 2. Table 3 displays patients' functional assessments in the three instruments. Patients who were discharged home have an average FIM total score of 17.19 (SD = 6.21), while the average of the total score for patients who were discharged to SNFs is 27.41 (SD = 7.46). This suggests that patients that were released home have better functional status when they were discharged from IRFs. Table 3 also shows that patients who were either released home or to SNFs have smaller average total scores at the later assessment date, suggesting that the functional status for most of patients improves over the course of their post-acute stay. The magnitude of improvement among the subsample of patients who received home health care appears to be larger than those who stayed in SNFs. 84.5% of the patients who recived home health improve their functional status, while only 48.2% of the patients in SNFs.

5.2. *Imputation model.* We illustrate the proposed nested multiple imputation procedure using the complete data set of 72,575 patients. In the first imputation stage, we impute the unmeasured assessments in MDS using the MVOPT-DA method described in Section 2.2. Age, gender, race and marital status are included in the model. Ten parallel chains of 50,000 iterations with dispersed initial values are generated, resulting in ten imputed data sets.

TABLE 2
Summary of the observed covariates for patients

Variable	**SNF**	**Home health**	**Overall**
Age	77.17 (9.62)	76.40 (10.05)	76.81 (9.83)
Gender, female (%)	53.0	53.2	53.1
Race, white (%)	81.2	77.0	79.2
Marital status, married (%)	42.2	50.0	45.9

TABLE 3
Summary of patients' functional outcomes in three instruments

Instrument	Variable	SNF	Home health	Overall
FIM	Score	27.41 (7.46)	17.19 (6.21)	22.63 (8.59)
	Score[a] at time 1	18.38 (2.85)	–	–
MDS	Score at time 2	17.43 (3.66)	–	–
	Difference[b]	−0.95 (2.41)	–	–
	Improved[c] (%)	48.2	–	–
	Score at time 1	–	15.60 (4.12)	–
OASIS	Score at time 2	–	10.59 (4.75)	–
	Difference	–	−5.01 (4.07)	–
	Improved (%)	–	84.5	–

[a]Score: the total score of functional assessments in each instrument;
[b]Difference: the difference in total scores measured on admission and at a later assessment date;
[c]Improved: the proportion of patients who experience functional improvement.

In the Translating step, we consider two possible models to illustrate the flexibility of the proposed procedure for translating assessments without re-equating the instruments. The first model is a linear regression model, $E(s_1|s_2) = \xi_0 + \xi_1 s_2$, where s_1 and s_2 denote the total scores of the imputed and observed items in MDS and OASIS on admission, respectively. The unmeasured total scores in MDS at the later assessment date are imputed using the estimates of ξ_0 and ξ_1 and the observed total scores in OASIS at the later assessment date. The second model is the MVOPT model, which models the joint distribution of all individual items in the imputed MDS and the observed OASIS instruments, $f(\mathbf{Y}^{\text{mds}}_{A,\text{imp}}, \mathbf{Y}^{\text{oas}}_{A,\text{obs}}|\boldsymbol{\psi})$. The unmeasured individual items in MDS at the later assessment date are imputed using the estimates of $\boldsymbol{\psi}$ and the observed individual items in OASIS at the later assessment date, $\mathbf{Y}^{\text{oas}}_{B,\text{obs}}$. For the MVOPT model, the DA algorithm in Section 2.2 is used to generate multiple imputations. Ten imputed data sets are generated in the second stage, resulting in 100 complete data sets.

We also examine the conversion table method and LVM to equate the MDS and OASIS instruments in the Equating step, and the linear regression model to impute the missing total scores in MDS in the Translating step. For LVM, in order to accommodate patients' covariates, we first partition the sample into five subclasses by sub-classifying at the quintiles of the distributions of the estimated propensity scores, $\widehat{\Pr}(M_i = 1|\mathbf{x}_i)$, and then impute the unmeasured assessments within each subclass.

5.3. *Model diagnostics.* As suggested by Gelman et al. (2005) and Abayomi, Gelman and Levy (2008), we evaluated the imputation model by comparing the distributions of the observed and the imputed values. Patients who were released

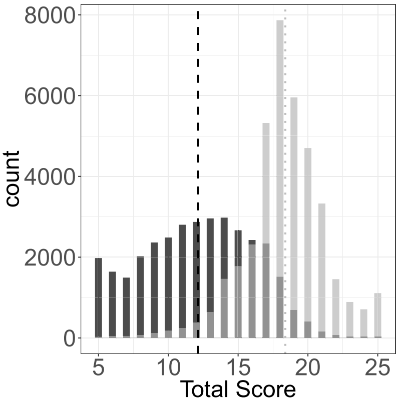


FIG. 2. *Histograms of the observed* (*gray*) *and imputed* (*black*) *total scores in MDS. The gray dotted line and black dashed line are the average observed and imputed total scores*, *respectively.*

home have smaller total MDS scores (see Figure 2), and are more likely to be at lower levels of each item (not shown). These patterns are consistent with the patterns that are observed in FIM.

We further examined the imputation model using posterior predictive checks [Gelman, Meng and Stern (1996), Burgette and Reiter (2010), He and Zaslavsky (2012), Si and Reiter (2013), Si, Reiter and Hillygus (2016)]. We first created $S = 1000$ complete data sets $D^{(s)} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(s)})$ $(s = 1, \dots, S)$ and replicated data sets $R^{(s)} = \mathbf{Y}_{\text{rep}}^{(s)}$ in which both $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ are simulated from the imputation model. We then compared each $D^{(s)}$ with its corresponding $R^{(s)}$ on three test statistics in the first stage imputation: (1) the mean total score of items in MDS, $T_1 \equiv \sum_{i,j} y_{ij}^{\text{mds}}/N$; (2) the proportion of response levels in each of the J_{mds} items in MDS, $T_2 \equiv \{n_{lj}/N, l = 1, \dots, c_j, j = 1, \dots, J_{\text{mds}}\}$, where n_{lj} is the number of responses at level l in item j; and (3) the pairwise Goodman and Kruskal's γ rank correlation coefficients between items in MDS, $T_3 \equiv \{\gamma(\mathbf{y}_j, \mathbf{y}_k), 1 \leq j < k \leq J_{\text{mds}}\}$. Let $T_{m,D^{(s)}}$ and $T_{m,R^{(s)}}$, $m = 1, 2, 3$, be the values of T_m computed with $D^{(s)}$ and $R^{(s)}$, respectively. For each T_m $(m = 1, 2, 3)$, we computed the two-sided posterior predictive probability (*ppp*),

$$ppp_m = (2/S) \times \min\left(\sum_{s=1}^{S} \mathbf{1}(T_{m,D^{(s)}} > T_{m,R^{(s)}}), \sum_{s=1}^{S} \mathbf{1}(T_{m,D^{(s)}} < T_{m,R^{(s)}})\right),$$

where $\mathbf{1}(\cdot)$ is the indicator function that is equal to 1 if the condition is satisfied and 0 otherwise. A small *ppp* indicates that $T_{D^{(s)}}$ and $T_{R^{(s)}}$ deviate from each other in one direction, which suggests that the imputation model distorts the data characteristics captured by T_m.

To obtain the pairs $(D^{(s)}, R^{(s)})$, we added a step to the DA algorithm that replaced all the values of $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ using the sampled parameter values at each

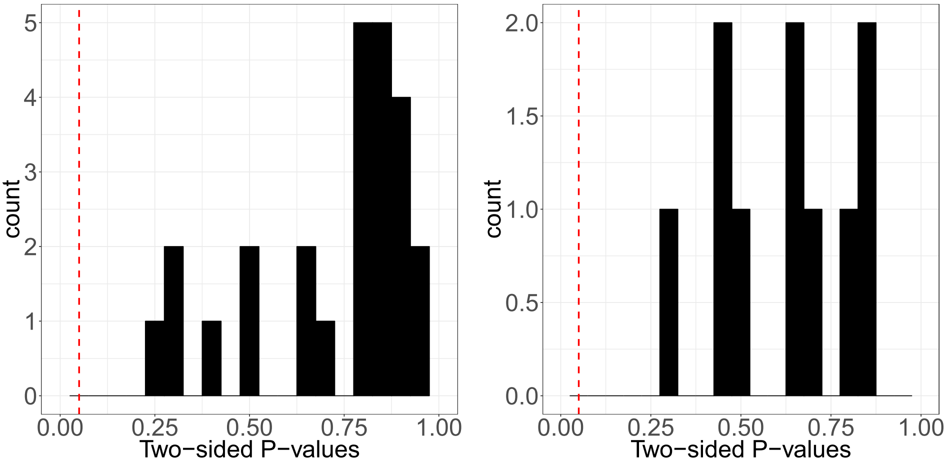


FIG. 3. *Histograms of the two-sided posterior predictive probabilities* (*ppp*) *for* T_2 (*left panel*) *and* T_3 (*right panel*). *The dashed line corresponds to a threshold value of* 0.05.

iteration. We calculated the test statistics T_1 based on 1000 complete and replicated data sets, and their differences $T_{1,D^{(s)}} - T_{1,R^{(s)}}$, $s = 1, \ldots, 1000$. The estimated two-sided $ppp_1 = 0.446$, which does not indicate a deficiency in the imputation model for T_1. The left and right panel of Figure 3 show the histogram of the two-sided ppp values for T_2 and T_3, respectively. None of the ppp_2 and ppp_3 values are below 0.05. Thus, we do not observe implausible imputations. Similar model diagnostics were performed for the second stage imputation, and no significant lack of model fit was detected (not shown).

Because posterior predictive checks may not be well calibrated [Hjort, Dahl and Steinbakk (2006)], we also examined the imputation performance using a sample partitioning method. Patients in SNFs were partitioned into a training sample that included 90% of the patients, and the remaining 10% served as a test sample. We fit the MVOPT model to the training sample and predicted the assessments of the test sample. We repeated this partitioning and prediction process 10 times, and in each replication we compared the distributions of the total mean score and the pairwise rank correlation coefficients of the predicted MDS assessments to the observed ones. Table 3 of the Supplementary Material [Gu and Gutman (2019)] shows the results of the 10 replications. No significant lack of model fit is detected.

5.4. *Analysis of multiply imputed data.* We compared the rates of functional change experienced by patients treated in SNFs and those treated by home health agencies using the observed and imputed assessments in MDS.

We define d_{snf} and d_{hh} to be the average change in total scores of the items in MDS over two assessments after discharge from IRFs for patients treated in SNFs and by home health agencies, respectively:

$$d_{\text{snf}} = \frac{1}{N_1}\sum_{i=1}^{N_1} S_{2,i}^{\text{snf}} - \frac{1}{N_1}\sum_{i=1}^{N_1} S_{1,i}^{\text{snf}}, \quad \text{and} \quad d_{\text{hh}} = \frac{1}{N_2}\sum_{i=1}^{N_2} S_{2,i}^{\text{hh}} - \frac{1}{N_2}\sum_{i=1}^{N_2} S_{1,i}^{\text{hh}},$$

where $S_{1,i}^{\text{snf}} = \sum_j y_{Aij}^{\text{mds}}$ and $S_{2,i}^{\text{snf}} = \sum_j y_{Bij}^{\text{mds}}$ are the total scores of the observed items in MDS for patients in SNFs on admission and at the later assessment, respectively, $S_{1,i}^{\text{hh}} = \sum_j \tilde{y}_{Aij}^{\text{mds}}$ and $S_{2,i}^{\text{hh}} = \sum_j \tilde{y}_{Bij}^{\text{mds}}$ are the total scores of the imputed items in MDS for patients receiving home health on admission and at the later assessment, respectively, N_1 and N_2 are the number of patients treated in SNFs and by home health agencies, respectively, and $N_1 + N_2 = N$. We also define p_{snf} and p_{hh} to be the proportion of patients whose functional status improved during the course of the post-acute stay in SNFs and home health care, respectively:

$$p_{\text{snf}} = \frac{1}{N_1}\sum_{i=1}^{N_1} \mathbf{1}\{S_{2,i}^{\text{snf}} < S_{1,i}^{\text{snf}}\}, \quad \text{and} \quad p_{\text{hh}} = \frac{1}{N_2}\sum_{i=1}^{N_2} \mathbf{1}\{S_{2,i}^{\text{hh}} < S_{1,i}^{\text{hh}}\},$$

where $\mathbf{1}\{A\}$ is an indicator function that is equal to 1 if A is true and 0 otherwise.

We apply the proposed NMI procedure to examine two quantities: (1) the difference in average change of total scores over the course of post-acute stay between patients in SNFs and those receiving home health; and (2) the difference in proportions of patients whose functional status improved during the post-acute stay between patients in SNFs and those receiving home health, $p_{\text{hh}} - p_{\text{snf}}$.

Table 4 displays the point and interval estimates of $d_{\text{hh}} - d_{\text{snf}}$ and $p_{\text{hh}} - p_{\text{snf}}$. The point and interval estimates with nested multiple imputation using either the regression translating model or the MVOPT translating model, as well as with

TABLE 4

Comparison of the estimated differences in average change of total score and estimated differences in proportion of patients whose functional status improve during the course of post-acute stay between patients treated in SNFs to those receiving home health care

	$d_{\text{hh}} - d_{\text{snf}}$			$p_{\text{hh}} - p_{\text{snf}}$ (%)		
	Estimate	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
CT[a]	−1.37	0.02	(−1.42, −1.34)	29.00	0.34	(28.32, 29.66)
LVM[b]	−0.01	0.05	(−0.11, 0.10)	12.08	0.53	(11.02, 13.13)
NMI (Reg)[c]	0.04	0.06	(−0.08, 0.16)	8.21	0.54	(7.15, 9.26)
NMI (MVOPT)[d]	−0.07	0.06	(−0.19, 0.05)	5.93	0.55	(4.84, 7.02)

[a] CT: conversion table method;

[b] LVM: latent variable matching;

[c] NMI (Reg): nested multiple imputation with linear regression translating model;

[d] NMI (MVOPT): nested multiple imputation with multivariate ordinal probit translating model.

LVM in the Equating step are similar. The results show that on average patients who received home health care do not have a significantly larger functional improvement than those who stayed in SNFs, but more patients who receive home health care improve their functional status during the post-acute stay than those in SNFs. In contrast, the results using the conversion table method suggest that on average patients who received home health care had a significantly higher rate of functional improvement than those who stayed in SNFs. In addition, a larger proportion of patients who received home health care experienced improved functional status in comparison to those who stayed in SNFs. Gu and Gutman (2017) noted that the conversion table has a poor performance when it is used to equate MDS and OASIS instruments. Thus, the estimates of the conversion table method in the Equating step may lead to implausible imputations in the Translating step, and overestimation of the rate of functional improvement for patients receiving home health.

The directions of the point estimates of $d_{\mathrm{hh}} - d_{\mathrm{snf}}$ are different for the different translating step methods, but their interval estimates are partially overlapping. The estimate of $p_{\mathrm{hh}} - p_{\mathrm{snf}}$ for the regression model is larger then that of the MVOPT model, suggesting that different translating models may result in different functional relationship between MDS and OASIS total scores. The MVOPT translating model incorporates more information by relying on all of the items in MDS and OASIS, which should result in a more accurate estimate.

6. Concluding remarks. We proposed a nested multiple imputation procedure to obtain a common patient assessment scale across the continuum of care by imputing unmeasured assessments at multiple dates in two steps. This procedure enables researchers to compare the rates of functional improvement experienced by patients treated in different health care settings using a common measure. This procedure accounts for the uncertainty in both the Equating and Translating steps, and it also provides flexibility for researchers to choose different translating models to impute multiple future assessments without the need to re-equate the instruments. The Equating step utilizes the MVOP model to impute the incomplete instruments that consist of multiple ordinal items. Simulations demonstrated that models based on MVOP are superior to existing methods for imputing incomplete multivariate ordinal variables in most of the experimental conditions that were examined. In addition, including observed covariates improves the point and interval estimates in the Equating step.

We applied the proposed procedure to analyze patients who had a stroke and were either released home or to SNFs after rehabilitation. Our analyses suggest that more patients who were discharged home and received home health care experience functional improvement in comparison to those who were released to SNFs, but on average the overall functional status improvement across all patients is similar across these settings. This analysis does not imply that one setting is more beneficial to patients than another, because the populations differ in patients' characteristics and initial functional status. However, using the proposed procedure, researchers can identify a subgroup of patients with similar characteristics and initial functional status who were discharged to either of the health care settings, and compare the rates of functional change in this subgroup of patients with the aim of identifying a setting that is more beneficial to certain patients. The proposed procedure can be further extended to impute unmeasured assessments at all assessment dates during patients' post-acute stays.

The newly proposed methods rely on the conditional independence and the missing at random assumptions. These assumptions are implicitly made in many educational testing applications with the common-items design. Here, these assumptions are somewhat defensible because all three instruments intend to determine the same underlying functional status, and they are all recorded within a close time period. In addition, the proposed methods performed well in a limited simulation analysis in which the two assumptions were violated. Nonetheless, developing procedures that accommodate departures from these assumptions is an area for future research.

One computational limitation of the MVOP model is the complexity of sampling from a truncated multivariate normal distribution, and this complexity is exacerbated when the dimension of the ordinal outcome variables is large. Another computational limitation is sampling the correlation matrix **R**. Here, we applied a parameter expansion technique to sample **R** efficiently. Recent work on using the prior distribution proposed in Lewandowski, Kurowicka and Joe (2009) and implemented in Stan [Carpenter et al. (2016)] is another possible solution.

In conclusion, we have proposed a procedure to obtain a common patient assessment scale across the continuum of care. This procedure is flexible and allows researchers to examine the rate of functional improvement using a single instrument.

SUPPLEMENTARY MATERIAL

Supplement to "Development of a common patient assessment scale across the continuum of care: A nested multiple imputation approach" (DOI: 10.1214/18-AOAS1202SUPP; .zip). The supplement includes the Slice Sampler Algorithm for the MVOP model, additional results in the simulation study of Section 4, results of posterior predictive checks in Section 5.3 and computer code for an example to illustrate the proposed procedure.

REFERENCES

ABAYOMI, K., GELMAN, A. and LEVY, M. (2008). Diagnostics for multivariate imputations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **57** 273–291. MR2440009

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ANDRIDGE, R. R. LITTLE, R. J. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **78** 40–64.

ASHFORD, J. R. and SOWDEN, R. R. (1970). Multi-variate probit analysis. *Biometrics* **26** 535–546.

BURGETTE, L. F. and REITER, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *Am. J. Epidemiol.* **172** 1070–1076.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P., RIDDELL, A. et al. (2016). Stan: A probabilistic programming language. *J. Stat. Softw.* **20** 1–37.

CHEN, M.-H. and DEY, D. K. (2000). A unified Bayesian approach for analyzing correlated ordinal response data. *Braz. J. Probab. Stat.* **14** 87–111. MR1838453

CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006). *Statistical Matching*: *Theory and Practice*. Wiley, Chichester. MR2268833

DAMIEN, P. and WALKER, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *J. Comput. Graph. Statist.* **10** 206–215. MR1939697

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DORANS, N. J., POMMERICH, M. and HOLLAND, P. W. (2007). *Linking and Aligning Scores and Scales*. Springer, New York.

ENDERS, C. K., KELLER, B. T. and LEVY, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychol. Methods* **23** 298–317.

FISCHER, H. F., TRITT, K., KLAPP, B. F. and FLIEGE, H. (2011). How to compare scores from different depression scales: Equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response Theory. *Int. J. Methods Psychiatr. Res.* **20** 203–214.

GAGE, B., CONSTANTINE, R., AGGARWAL, J., MORLEY, M., KURLANTZICK, V., BERNARD, S. et al. (2012). The development and testing of the Continuity Assessment Record and Evaluation. (CARE) item set: Final report on the development of the CARE item set.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., VAN MECHELEN, I., VERBEKE, G., HEITJAN, D. F. and MEULDERS, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61** 74–85. MR2135847

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1** 141–149.

GEWEKE, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing Science and Statistics*: *Proceedings of the* 23rd *Symposium on the Interface* 571–578. Interface Foundation of North America, Fairfax, VA.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GOODMAN, L. A. and KRUSKAL, W. H. (1979). *Measures of Association for Cross Classifications*. *Springer Series in Statistics* **1**. Springer, New York. MR0553108

GU, C. and GUTMAN, R. (2017). Combining item response theory with multiple imputation to equate health assessment questionnaires. *Biometrics* **73** 990–998. MR3713132

GU, C. and GUTMAN, R. (2019). Supplement to "Development of a common patient assessment scale across the continuum of care: A nested multiple imputation approach." DOI:10.1214/18-AOAS1202SUPP.

GUERRERO, V. M. and JOHNSON, R. A. (1982). Use of the Box–Cox transformation with binary response models. *Biometrika* **69** 309–314. MR0671968

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2015). Graphical models for ordinal data. *J. Comput. Graph. Statist.* **24** 183–204. MR3328253

HAREL, O. (2003). Strategies for data analysis with two types of missing values. PhD thesis, Pennsylvania State Univ., State College, PA.

HE, Y. and ZASLAVSKY, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Stat. Med.* **31** 1–18. MR2868986

HEITJAN, D. F., LANDIS and RICHARD, J. (1994). Assessing secular trends in blood pressure: A multiple-imputation approach. *J. Amer. Statist. Assoc.* **89** 750–759.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive p-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154

HOLMES, C. C. and HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1** 145–168. MR2227368

JELIAZKOV, I., GRAVES, J. and KUTZBACH, M. (2008). Fitting and comparison of models for multivariate ordinal outcomes. *Adv. Econom.* **23** 115–156.

KOLEN, M. J. and BRENNAN, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd ed. Springer, New York. MR3156933

LAWRENCE, E., LIU, C., BINGHAM, D. and NAIR, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics* **50** 182–191. MR2439877

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081

LI, Y. and SCHAFER, D. W. (2008). Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. *Comput. Statist. Data Anal.* **52** 3474–3492. MR2427359

LI, C.-Y., ROMERO, S., SIMPSON, K. N., BONILHA, H. S., SIMPSON, A. N., HONG, I. and VELOZO, C. A. (2017). Linking existing instruments to develop a continuum of care measure: Accuracy comparison using function-related group classification. *Qual. Life Res.* 1–10.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85** 755–770. MR1666758

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR3223057

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

OLSSON, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460. MR0554892

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1994). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. MR2055518

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability* **72**. Chapman & Hall, London. MR1692799

SHEN, Z. (2000). Nested multiple imputations. PhD thesis, Harvard Univ., Cambridge, MA.

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38** 499–521.

SI, Y., REITER, J. P. and HILLYGUS, D. S. (2016). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. *Ann. Appl. Stat.* **10** 118–143. MR3480490

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

TEN KLOOSTER, P. M., VOSHAAR, M. A. H. O., GANDEK, B., ROSE, M., BJORNER, J. B., TAAL, E., GLAS, C. A. W., VAN RIEL, P. L. C. M. and VAN DE LAAR, M. A. F. J. (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. *Health Qual Life Outcomes* **11** 199.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16** 219–242. MR2371007

VARIN, C. and CZADO, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11** 127–138.

VELOZO, C. A., BYERS, K. L. and JOSEPH, B. R. (2007). Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the functional independence measure and the minimum data set. *J. Rehabil. Res. Dev.* **44** 467.

VERMUNT, J. K., VAN GINKEL, J. R., DER ARK, V., ANDRIES, L. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Method.* **38** 369–397.

VON DAVIER, A. A., ed. (2011). *Statistical Models for Test Equating, Scaling, and Linking. Statistics for Social and Behavioral Sciences*. Springer, New York. MR2798186

WEI, C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WYSOCKI, A., THOMAS, K. S. and MOR, V. (2015). Functional improvement among short-stay nursing home residents in the MDS 3.0. *J. Am. Med. Dir. Assoc.* **16** 470–474.

YUCEL, R. M., HE, Y. and ZASLAVSKY, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Stat. Med.* **30** 3447–3460. MR2861625

ZHANG, X., BOSCARDIN, W. J. and BELIN, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *J. Comput. Graph. Statist.* **15** 880–896. MR2297633

ZHANG, X., LI, Q., CROPSEY, K., YANG, X., ZHANG, K. and BELIN, T. (2017). A multiple imputation method for incomplete correlated ordinal data using multivariate probit models. *Comm. Statist. Simulation Comput.* **46** 2360–2375. MR3625287

DEPARTMENT OF HEALTH CARE POLICY
HARVARD MEDICAL SCHOOL
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: gu@hcp.med.harvard.edu

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: roee_gutman@brown.edu