Published in final edited form as:

Stat Med. 2016 July 30; 35(17): 2991–3006. doi:10.1002/sim.6590.

Joint Multiple Imputation for Longitudinal Outcomes and Clinical Events Which Truncate Longitudinal Follow-up

Bo Hu^{1,*}, Liang Li², and Tom Greene³

¹Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH

²Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX

³Division of Clinical Epidemiology, University of Utah, Salt Lake City, UT

Abstract

Longitudinal cohort studies often collect both repeated measurements of longitudinal outcomes and times to clinical events whose occurrence precludes further longitudinal measurements. Although joint modeling of the clinical events and the longitudinal data can be used to provide valid statistical inference for target estimands in certain contexts, the application of joint models in medical literature is currently rather restricted due to the complexity of the joint models and the intensive computation involved. We propose a multiple imputation (MI) approach to jointly impute missing data of both the longitudinal and clinical event outcomes. With complete imputed datasets, analysts are then able to use simple and transparent statistical methods and standard statistical software to perform various analyses without dealing with the complications of missing data and joint modeling. We show that the proposed MI approach is flexible and easy to implement in practice. Numerical results are also provided to demonstrate its performance.

Keywords

multiple imputation; missing data; joint modeling; competing risk; longitudinal data

1. Introduction

Longitudinal clinical studies often include repeated measurements of biomarkers, patient reported outcomes or other longitudinal outcomes, as well as clinical events that preclude further measurement of the longitudinal outcomes. In particular, studies of chronic kidney disease (CKD) often center on longitudinal assessments of markers of disease severity or comorbidity, such as estimated glomerular filtration rate (eGFR), proteinuria, quality of life or physical function, etc., which are measured according to a designed schedule until death or kidney failure, at which point follow-up is terminated. When subjects are followed over time in a longitudinal study, it is almost inevitable that some longitudinal assessments will be missing [1, 2, 3]. Meanwhile the survival outcomes, measured as times to the clinical events, may also be missing due to loss to follow-up or administrative censoring.

*Correspondence to: hub@ccf.org.

Historically, most studies in the medical literature have considered the longitudinal and survival outcomes in separate analyses, in which planned longitudinal measurements occurring after the terminating events are treated as ignorable missing data. This practice has a number of drawbacks. When the longitudinal outcomes and the clinical events are correlated, failure to appropriately account for "non-ignorable" dropout can lead to biased statistical inference for the longitudinal outcomes. Furthermore, conducting separate analyses reduces efficiency by failing to incorporate the information provided jointly by both types of outcomes.

Joint modeling of the longitudinal and survival outcomes has gained popularity over the past decades (see [4] for a systematic review). Many joint modeling approaches fall under the framework of shared-parameter models [5, 6, 7]. A shared-parameter model typically consists of two sub models, one for the longitudinal outcome and the other for the survival outcome. The two sub models are linked through a shared parameter that represents subject-specific characteristics of the longitudinal profiles. Joint analysis may also be performed using multi-state approaches [8]. Hu et al. (2012) [9] proposed a multi-state representation that classifies subjects into discrete states defined jointly by the longitudinal and survival outcomes. Although a number of joint modeling approaches are available, their applications in medical literature are currently limited due to their statistical and computational complexity. For example, the shared-parameter models often involve a multi-dimensional integration, which may become intractable for a high dimensional shared parameter that is often needed in models with non-linear longitudinal profiles [10, 11]. The available software for joint modeling also has limited model options for the great diversity of problems encountered in practice [12].

In this paper we propose a computationally simple and transparent multiple imputation (MI) approach to impute missing data in both longitudinal and survival outcomes to support statistical inferences. We shall restrict our consideration to estimands that can be defined by the joint distribution of the survival outcomes and longitudinal measurements obtained prior to the occurrence of the clinical events. This avoids the conceptual problems that occur when attempting to conduct statistical inference on outcomes obtained after death or other terminating events [13]. With this restriction, we restrict imputations to the survival outcomes themselves and to truly missing longitudinal measurements that are missing due to dropout or intermittent missingness as opposed to the occurrence of the clinical events. For the survival part, we will also restrict attentions to estimands and imputed events within the follow-up period. With complete data after imputation, simple and direct statistical approaches can be used for analysis. In other words, imputation plus simple approaches can serve as a practical alternative to the complicated joint modeling approaches.

There is an extensive literature on the imputation of missing longitudinal data [2, 14, 15]. Ali and Siddiqui (2009) [16] specifically discussed the imputation of longitudinal data in the informative dropout setting, but their imputation approach does not depend on survival outcomes. There also exist many imputation methods for missing covariates in survival analysis (e.g., [17, 18, 19]), where the missing data are usually for the baseline covariates. It is generally recommended that the survival outcome should be used for the imputation of covariates. Several approaches have been proposed for the imputation of censored event

times (e.g., [20, 21]). [22] and [23] discussed the imputation of censored event times by using longitudinal data as auxiliary variables. [24] discussed the use of multiple imputation for the diagnostics of joint models. However, this approach requires fitting the joint model prior to multiple imputations. Our proposed approach differs from the current literature by imputing both missing survival and longitudinal data. The imputation proceeds in a sequential fashion so that the imputation of each outcome depends on the other.

Section 2 introduces a motivating example. Section 3 describes the general framework of the MI approach and a flexible parametric imputation approach under this framework that is easy to implement in practice. Section 4 applies the MI algorithms to a simulation study and the motivating example to examine the numerical performance. Section 5 concludes the paper with a brief discussion.

2. The AASK Study

The African American Study of Kidney Disease and Hypertension (AASK) is a randomized clinical trial with a follow-on cohort study in 1094 African Americans with chronic hypertensive kidney disease [25]. The follow-up time ranges from 8.5 to 11.5 years depending on the date of randomization of the participants.

In the AASK Study, as in most clinical trials of chronic kidney disease, follow-up of longitudinal outcomes was terminated at the occurrence End Stage Renal Disease (ESRD) or death, whichever occurred first. Therefore the survival data in AASK fall under the competing-risk framework since the occurrence of death censored the observation of ESRD and vice versa. ESRD is defined by the initiation of renal replacement therapy that dramatically alters the interpretation of many biomarkers measured in the serum or the urine. In general renal replacement therapy may be provided in form of dialysis or kidney transplantation, but in the AASK study all occurrences of ESRD were signified by the initiation of dialysis. 318 AASK patients reached dialysis and another 176 patients died prior to the end of follow-up period.

The AASK study database includes an extensive array of longitudinal data and provides a rare opportunity to study the long-term progression of chronic kidney disease. In this paper we consider the bivariate longitudinal outcome of estimated glomerular filtration rate (eGFR) and proteinuria. eGFR is calculated from serum creatinine, age and gender, and is the standard measure of the level of kidney function [26]. In the AASK study, eGFR was assessed at baseline and approximately every 6 months thereafter until the patient died, reached dialysis, or was otherwise lost to follow-up. Proteinuria, measured as the ratio of urinary protein and creatinine in a 24-hour urine collection, was assessed at baseline, every 6 months during the first five years and annually thereafter. The maximum number of eGFR measurements was 21 and the maximum number of proteinuria measurements was 15 for the AASK participants.

Figure 1 shows the status of data collection at each scheduled patient visit. For simplicity, we have defined a common right censoring time for both the longitudinal and survival outcomes, and have categorized any missing measurements of the longitudinal outcome

prior to right censoring and the clinical events as intermittently missing. At the time of each scheduled visit, a patient could be deceased, in ESRD, right censored, continuing in follow-up with intermittently missing longitudinal data (eGFR or proteinuria), or continuing in follow-up with observed longitudinal data. Under our framework, the status of the patient is regarded as known at a given visit if they previously reached ESRD or death, so that only the categories defined by right censoring or intermittent missingness indicate true missing data requiring imputation.

In AASK, 823 patients had intermittent missing patterns of the longitudinal data, with intermittent missing gap ranging from 9 to 64 months (with a mean of 16.5 months). 355 patients were lost to follow-up, causing missingness in both the longitudinal and survival data.

3. Multiple Imputation

3.1. Data Notations

We consider competing-risk type of survival outcome with M event types. T^* denotes the time to the first event and δ^* takes a value of $m = 1, \dots, M$ if event type m occurs first. In the AASK study, $m = 1$ for dialysis and $m = 2$ for death.

For the longitudinal data, we consider multivariate outcomes, $\mathbf{y}_1, \dots, \mathbf{y}_J$, where $\mathbf{y}_j$ is a vector of repeated measurements of the j th outcome that are scheduled to be measured at a set of time points $t_j = t_{j1}, \dots, t_{jm_j}$. $J = 2$ in the AASK example, where $\mathbf{y}_1 = (y_1^1, \dots, y_1^{21})^T$ is a vector of longitudinal eGFR measurements at $t_1 = (0, 6, \dots, 120)^T$ and $\mathbf{y}_2 = (y_2^1, \dots, y_2^{15})^T$ is a vector of longitudinal proteinuria measurements at $t_2 = (0, 6, \dots, 48, 60, \dots, 120)^T$. For simplicity of notation, we denote $\mathbf{Y}$ as the vector of all longitudinal outcomes that are pooled together and arranged chronologically. Outcomes with tied measurement times may be arranged randomly.

Although the longitudinal outcomes are scheduled to be measured at all time points, their measurements are not feasible or their meanings are significantly altered after the occurrence of the events [27]. Therefore we consider longitudinal outcomes with measurement times prior to the time to the first clinical event, denoted by $\mathbf{Y}_{T^*}$. Then $\mathbf{W} = (\mathbf{Z}^T, T^*, \delta^*, \mathbf{Y}_{T^*}^T)^T$ denotes the complete data, where $\mathbf{Z}$ is a vector of baseline covariates.

Let C be the censoring time for the clinical events, which we assume to be independent of the longitudinal and survival variables given baseline covariates $\mathbf{Z}$. The observed survival data are then $T = \min(T^*, C)$ and $\delta = \delta^* \mathbf{I}(T^* < C)$. The actual event time T^* is censored and can be considered as missing but greater than T if $\delta = 0$. Meanwhile we decompose the longitudinal data $\mathbf{Y}_{T^*}$ as $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$, where $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ are the observed and missing part of $\mathbf{Y}_{T^*}$, respectively.

Suppose that the study has n subjects, $\mathbf{W}_i = (\mathbf{Z}_i^T, T_i^*, \delta_i^*, \mathbf{Y}_i^T)^T$ is a vector of complete data for subject i , $i = 1, \dots, n$. Meanwhile $(\mathbf{Z}_i^T, T_i, \delta_i, \mathbf{Y}_{i,obs}^T)^T$ represents the observed data. For

simplicity, we assume no missing data for the baseline variables $\mathbf{Z}$ throughout this paper but our approach can be easily extended to account for missing baseline data.

3.2. General Framework

Multiple imputation replaces each missing data with a random sample of plausible values. More formally, imputed values should be drawn from the posterior predictive distribution of the missing data given the observe data. In our context, the target predictive distribution can be written as

$$p(T^*, \delta^*, \mathbf{Y}_{mis} | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}). \quad (1)$$

Since the missing data in (1) consist of both longitudinal and survival parts, we factor (1) as

$$p(T^*, \delta^*, \mathbf{Y}_{mis} | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) = p(T^*, \delta^* | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) p(\mathbf{Y}_{mis} | T^*, \delta^*, T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}). \quad (2)$$

For the missing survival data, we can further decompose $p(T^*, \delta^* | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z})$ in (2) as

$$p(T^* | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) p(\delta^* | T^*, T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}). \quad (3)$$

Therefore we can impute the composite event time T^* and event type δ^* sequentially.

Based on equations (2) and (3), it is intuitive to impute the missing data in two steps, where the first step imputes the missing survival data (T^*, δ^*) and the second step imputes the missing longitudinal data $\mathbf{Y}_{mis}$. For subjects with observed survival data, that is, $\delta > 0$, the target predictive distribution is $p(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}, T^*, \delta^*)$. Therefore imputation is only needed for missing longitudinal data. In principle, any appropriate approach can be applied to impute the missing data in each step. In the following section, we describe a simple parametric approach to illustrate the general framework.

3.3. A Parametric Approach

For the parametric imputation approach, we express (1) as

$$\begin{aligned}
 p(T^*, \delta^*, \mathbf{Y}_{mis} | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) &= \int p(T^*, \delta^*, \mathbf{Y}_{mis} | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}; \theta) p(\theta | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) d\theta \\
 &= \int p(T^* | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}; \boldsymbol{\alpha}) p(\delta^* | T^*, T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}; \boldsymbol{\beta}) \times p(\mathbf{Y}_{mis} | T^*, \delta^*, T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}; \boldsymbol{\gamma}) p(\theta | T, \delta=0, \mathbf{Y}_{obs}, \mathbf{Z}) d\theta,
 \end{aligned}
 \tag{4}$$

where $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})^T$ is a vector of unknown parameters and each component of θ indexes one corresponding conditional distribution in the second equality. Instead of evaluating the integral (4), the data augmentation (DA) algorithm [28] is generally more convenient, which proceeds by randomly sampling the parameters and missing data iteratively.

Based on (4), we propose an imputation approach based on sequential conditional parametric models. After l iterations, let $T^{*(l)} = (T_{mis}^{*(l)}, T_{obs}^*)$, $\delta^{*(l)} = (\delta_{mis}^{*(l)}, \delta_{obs}^*)$ and $\mathbf{Y}^{(l)} = (\mathbf{Y}_{mis}^{(l)}, \mathbf{Y}_{obs})$ denote the complete survival and longitudinal data, where $T_{mis}^{*(l)}$, $\delta_{mis}^{*(l)}$ and $\mathbf{Y}_{mis}^{(l)}$ represent imputed data. At iteration $l+1$, we impute the missing values of T^* , δ^* and $\mathbf{Y}$ in the following steps.

Step 1a. Impute event times—For each censored time t that imputation is needed, we assume a Weibull model for the composite event time T^* for the set of subjects at risk, that is, $R(t) = \{i/T_i \leq t\}$. More specifically, the hazard function is

$$h(s; \boldsymbol{\alpha}_t) = \alpha_{t0} s^{\alpha_{t0}-1} \exp(\alpha_{t1} + \alpha_{t2}^T f(\mathbf{Y}^{(l)}) + \alpha_{t3}^T \mathbf{Z}) \text{ for } s \geq t, \quad (5)$$

where $\boldsymbol{\alpha}_t = (\alpha_{t0}, \dots, \alpha_{t3})^T$ is part of the parameter vector $\boldsymbol{\alpha}$ in (4). $f(\mathbf{Y}^{(l)})$ represents some characteristics of the longitudinal data. For example, in the AASK study, $f(\mathbf{Y}^{(l)})$ could be the baseline eGFR and proteinuria, the latest eGFR and proteinuria prior to t or both baseline and the latest values [23].

Let $\hat{\boldsymbol{\alpha}}_t = (\hat{\alpha}_{t0}, \dots, \hat{\alpha}_{t3}^T)^T$ denote the vector of the maximum likelihood estimates of the parameters in model (5), which asymptotically follows a multivariate normal (MVN) distribution as $\sqrt{n_t}(\hat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t) \sim \text{MVN}(0, \sum \boldsymbol{\alpha}_t)$. n_t is the size of $R(t)$. For the imputation, we

1. draw $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_{t0}, \tilde{\alpha}_{t1}, \tilde{\alpha}_{t2}, \tilde{\alpha}_{t3})^T$ from the distribution $\text{MVN}(\hat{\boldsymbol{\alpha}}_t, \hat{\Sigma}_{\boldsymbol{\alpha}}/n_t)$;
2. compute the hazard function $h(s; \tilde{\boldsymbol{\alpha}})$ for subjects that need imputation, that is, $\{i/T_i = t, \delta_i = 0\}$;
3. draw $T^{*(l+1)}$ from the cumulative density function $1 - S(s; \tilde{\boldsymbol{\alpha}})/S(t; \tilde{\boldsymbol{\alpha}})$ for $s \leq t$, where $S(t; \tilde{\boldsymbol{\alpha}})$ is the survival probability function.

Step 1 above uses an empirical Bayes approach by drawing $\tilde{\boldsymbol{\alpha}}$ from the sampling distribution of $\hat{\boldsymbol{\alpha}}_t$ in fitting the Weibull model (5), which is a common approximation approach used for

regression-based imputation methods [29]. Based on model (5), the logarithm of the survival probability in the third step above can be readily calculated as

$$\begin{aligned}\log S(t; \tilde{\alpha}) &= -\int_{-\infty}^t h(x; \tilde{\alpha}) dx \\ &= -t^{\tilde{\alpha}_{t0}} \exp(\tilde{\alpha}_{t1} + \tilde{\alpha}_{t2}^T f(\mathbf{Y}^{(l)}) + \tilde{\alpha}_{t3}^T \mathbf{Z}).\end{aligned}$$

Step 1b. Impute event types—We assume a multinomial logistic model for the event indicator δ^* . Denote $p_m(\boldsymbol{\beta}_t) = \Pr(\delta^* = m)$ the probability that event m occurs first for the subjects with imputed event time $T_i^{*(l+1)}$ from Step 1a. Under the multinomial logistic model, we have

$$\log \frac{p_m(\boldsymbol{\beta}_t)}{p_1(\boldsymbol{\beta}_t)} = \beta_{t0}^m + \beta_{t1}^m \log T^{*(l+1)} + \beta_{t2}^m f(\mathbf{Y}^{(l)}) + \beta_{t3}^m \mathbf{Z} \quad (6)$$

for $m > 1$, where $\boldsymbol{\beta}_t = (\beta_{t0}^2, \dots, \beta_{t3}^M)^T$ is part of the parameter vector $\boldsymbol{\beta}$ in (4) and $\sum_{m=1}^M p_m(\boldsymbol{\beta}_t) = 1$. The log-transformation of the event time, $\log T^{*(l+1)}$, is included as a covariate in model (6) to model the dependence of the event type on the event time. Other transformations of $T^{*(l+1)}$ may be also considered, e.g., $H(T^{*(l+1)})$ with $H(\cdot)$ being the Nelson-Aalen estimator of the cumulative hazard function [19].

Fitting model (6) should be restricted to subjects in $R(t)$ with $\delta > 0$, which then yields a maximum likelihood estimate $\hat{\boldsymbol{\beta}}_t$ that asymptotically follows a multivariate normal distribution $\text{MVN}(\boldsymbol{\beta}_t, \Sigma_{\boldsymbol{\beta}_t})$. To impute the event types for subject i with $i \in \{i/T_i = t, \delta_i = 0\}$, we

- 1 draw $\tilde{\boldsymbol{\beta}}_t$ from $\text{MVN}(\hat{\boldsymbol{\beta}}_t, \hat{\Sigma}_{\boldsymbol{\beta}_t})$;
- 2 compute probabilities $p_m(\tilde{\boldsymbol{\beta}}_t)$, $m = 1, \dots, M$;
- 4 draw a M -vector r from a multinomial distribution with probabilities $p_m(\tilde{\boldsymbol{\beta}}_t)$, $m = 1, \dots, M$;
- 4 set $\delta^{*(l+1)} = m$ if the m th element of r is equal to 1.

Steps 1a and 1b are repeated for all distinct observed censoring times, after which all missing survival data are imputed. And the resulting data can be expressed as

$(T_{mis}^{*(l+1)}, T_{obs}^*, \delta_{mis}^{*(l+1)}, \delta_{obs}^*, \mathbf{Y}_{mis}^{(l)}, \mathbf{Y}_{obs})$. Step 1b can be skipped for standard survival data without competing events. In practice, the risk set $R(t)$ may have a small sample size for some t and fitting models (5) and (6) might encounter numerical problems. An approximation approach can be obtained by running these two imputation steps at a set of landmark times $(\tau_1 < \dots < \tau_K)$ instead of all observed censoring times. For each landmark time τ_k , models (5) and (6) will be fit to subjects with $T_i \geq \tau_k$ and the fitted models will be used to impute missing data for subjects with $T_i \in [\tau_k, \tau_{k+1})$ and $\delta_i = 0$. This basically assumes constant modeling relationship for each landmark interval $[\tau_k, \tau_{k+1})$. A similar

approach is used for the dynamic prediction by landmarking [30]. The landmark times could be selected to ensure proper fit of models (5) and (6).

Step 2. Impute longitudinal data—If the vector of complete data $\mathbf{W}$ or an appropriate transformation of $\mathbf{W}$ follows a multivariate normal distribution, the conditional distribution $p(\mathbf{Y}/T^*, \delta^*, \mathbf{Z}; \boldsymbol{\gamma})$ is also multivariate normal, where $\boldsymbol{\gamma}$ consists of the mean vector and the variance-covariance matrix of the MVN distribution. In this case, a direct approach is to use the Markov chain Monte Carlo (MCMC) method [31] to impute the missing longitudinal data. The MCMC method initially estimates $\boldsymbol{\gamma}$ using the Expectation-Maximization (EM) algorithm, and then iterates an I-step and a P-step. The I-step draws $\mathbf{Y}_{mis}^{(l+1)}$ from the conditional MVN distribution of $p(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, T^{*(l+1)}, \delta^{*(l+1)}, \mathbf{Z}; \boldsymbol{\gamma})$. The P-step simulates $\boldsymbol{\gamma}$ based on the observed and imputed values of $\mathbf{Y}$.

In cases where it is difficult to identify a suitable multivariate distribution for $\mathbf{W}$ or its transformations, a popular alternative is the fully conditional specification (FCS) approach [32]. The FCS approach is able to handle mixed variable types since each variable is imputed using its own imputation model. The FCS approach is also known as multiple imputation by chained equations (MICE) in the literature [29]. More specifically, let $\mathbf{Y}_j$ be the j th component of the longitudinal data $\mathbf{Y}$ and let $\mathbf{W}_{-j}$ denote the vector containing the remaining variables of $\mathbf{W}$ except Y_j . Under the assumption that the conditional distribution $p(\mathbf{Y}_j|\mathbf{W}_{-j}, \boldsymbol{\gamma}_j)$ follows a parametric regression model with parameter $\boldsymbol{\gamma}_j$, we

1. regress $\mathbf{Y}_j$ on $\mathbf{W}_{-j}$ by restricting to subjects with observed values of $\mathbf{Y}_j$;
2. draw $\tilde{\boldsymbol{\gamma}}_j$ from the sampling distribution of the estimates of $\boldsymbol{\gamma}_j$ in fitting the regression model;
3. draw $\mathbf{Y}_j^{(l+1)}$ from $p(\mathbf{Y}_j|\mathbf{W}_{-j}, \tilde{\boldsymbol{\gamma}}_j)$.

The regressors $\mathbf{W}_{-j}$ in the first step contain both observed and imputed values of other variables. The above process is repeated for all components of the longitudinal data $\mathbf{Y}$. Subsequently iteration $l + 1$ completes and data are updated as

$(T_{mis}^{*(l+1)}, T_{obs}^*, \delta_{mis}^{*(l+1)}, \delta_{obs}^*, \mathbf{Y}_{mis}^{(l+1)}, \mathbf{Y}_{obs})$. For subjects with observed survival data, the imputation consists only of Step 2.

The multi-step imputation procedure may iterate several times to achieve numerical convergence, which is similar to the chained equation approach for multiple imputation. While there is not a formal approach to determine the number of iterations, convergence can be monitored by examining the mean values of each variable under imputation over iterations [33]. Simulation works indicated that satisfactory performance can be achieved with just 5 or 10 iterations [34, 35].

Upon completion of the imputation, an administrative censoring time may be applied to each imputed dataset, which typically coincides with the planned follow-up period of the study. For example, AASK study ended on June 30th, 2007. Then any imputed event times later than the end date can be discarded.

4. Numerical Results

4.1. Simulation Studies

Simulation model—We considered a univariate longitudinal outcome in the simulation studies. The serial measurements of the longitudinal outcome $Y(t)$, $t = 0, 6, \dots, 120$, and the survival outcome (T^*, δ^*) were generated from the following models,

$$\begin{aligned} Y(t) &= \mathbf{b}_0 + \mathbf{b}_1 t + e(t), \\ \log T^* &= \alpha_0 + \alpha_1^T \mathbf{Z} + \alpha_2^T \mathbf{b} + \varepsilon, \\ \log \frac{P(\delta^*=2)}{P(\delta^*=1)} &= \beta_0 + \beta_1^T \mathbf{Z} + \beta_2^T \mathbf{b} + \beta_3 \log T^*, \end{aligned} \quad (7)$$

where $\delta^* = 1$ and 2 to denote dialysis and death, respectively; $\mathbf{Z} \sim \text{MVN}(\mu_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ represents a two-dimensional vector of baseline variables; $\mathbf{b} = (\mathbf{b}_0, \mathbf{b}_1)^T \sim \text{MVN}(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ represents the intercept and slope of the longitudinal trajectory; $e(t) \sim N(0, \sigma_e^2)$ is the residual error; and $\varepsilon \sim N(0, \sigma_{\varepsilon}^2)$. The values of all parameters, $\mu_{\mathbf{Z}}, \mu_{\mathbf{b}}, \Sigma_{\mathbf{Z}}, \Sigma_{\mathbf{b}}, \sigma_e, \sigma_{\varepsilon}, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, were estimated from the AASK data.

An independent censoring time C was simulated from a mixture of an exponential distribution $\exp(1/300)$ and a truncated normal distribution $\mathcal{N}(130, 30)\mathbf{I}_{(0,130)}$. The mixture distribution was chosen to yield an overall high censoring rate and an accelerated censoring rate towards the end of the follow-up period. C censors the observation of the time to event, leading to $T = \min(T^*, C)$ and $\delta = \delta^* \mathbf{I}(T^* < C)$ as the observed survival data.

Missing model of longitudinal data—We generated missing data of the longitudinal outcome according to the following strategy:

- (Dropout). We only simulated $Y(t)$ with $t \leq T$.
- (Intermittent missing). A simulated $Y(t)$ was randomly deleted according to the proportion of intermittent missing eGFR at time t in AASK.
- (Consecutively-intermittent missing). We simulated the number of consecutively-intermittent missing longitudinal measurements from a Poisson distribution, $n_c \sim \text{Poisson}(\lambda_c)$. We then randomly selected a time point and deleted all n_c subsequent longitudinal measurements.
- (Lost to follow-up). We simulated $n_l \sim \text{Poisson}(\lambda_l)$, and then deleted n_l consecutive measurements prior to the time T .

The baseline longitudinal outcome $Y(0)$ was retained for all subjects. We considered two missing mechanisms, missing at random (MAR) and missing not at random (MNAR), by selecting different models for λ_c and λ_l . For MAR, $\log \lambda_c$ and $\log \lambda_l$ were linear functions of the baseline longitudinal data $Y(0)$, T and δ , which were all kept observed in the simulation study. Then missingnesses of the longitudinal data only depended on observed data. For MNAR, λ_c and λ_l further depended on $\mathbf{b}_0$ and $\mathbf{b}_1$, the unobserved random intercept and slope

of the longitudinal trajectory. Details of the models for λ_c and λ_I can be found in the Supplementary Materials.

The simulation size was 100 and the sample size was 500 in each simulation. Figure 2 shows the proportion of missing measurements of the longitudinal outcome at each scheduled visit averaged over 100 simulations for the missing at random scenario. Similar missing rates were observed under the MNAR mechanism. For each simulated dataset, we applied two approaches to impute the missing data. The two approaches differed only in the imputation of the missing longitudinal data, where the first approach applied the MCMC method by assuming a multivariate normal distribution for the longitudinal outcome given all other variables; and the second approach used the FCS method. We refer these two approaches as "MVN" and "FCS" approaches throughout the remainder of this paper. The number of imputations was 10 for each simulated data set, and Steps 1a, 1b and 2 in Section 3.3 were iterated 10 times for each imputation.

To evaluate the performance of the MI approach, we considered the multi-state model, where the longitudinal and survival outcomes are analyzed jointly by classifying subjects into a set of mutually exclusive states. For the simulation study, we considered 5 states at each time t ,

$$s_t = k, \quad k = 1, \dots, 5, \quad (8)$$

where $k = 1, \dots, 5$ represent the states of no events and $Y(t)$ below 30, between 30 and 60 and above 60, dialysis and deaths, respectively.

We considered the following estimands associated with the multi-state model: (i) the multi-state probabilities $P_{kt} = P(s_t = k)$ for $t \in [0, 120]$; (ii) the areas under the multi-state probability curves, $\int_0^{120} P_{kt} dt$ for $k = 1, \dots, 5$; (iii) transition probabilities between states over designated time intervals, $P(s_t = k/s_u = l)$ with $t > u$. We also considered (iv) the mean intercept and slope of the longitudinal trajectory; (v) the average one-year change of the longitudinal outcome preceding dialysis or death. With a complete imputed dataset, all estimands expect for (iv) can be estimated as simple proportions or means. The mean intercept and slope of the longitudinal data were estimated by using the unweighted least square method [36]. Additional details concerning the estimation procedures are described in the Supplementary Materials.

The estimates from multiple imputed datasets were summarized using Rubin's rule [1]. For any estimand Q , let $\hat{Q}_j$ be the estimate obtained from the j th imputed dataset, $j = 1, \dots, m$. V_j is the variance estimate associated with $\hat{Q}_j$. The multiple imputation estimate for Q is the average of the m estimates

$$\hat{Q} = \frac{\sum_j \hat{Q}_j}{m}$$

with the variance estimate as

$$\hat{V} = \frac{\sum_j V_j}{m} + \left(1 + \frac{1}{m}\right) B,$$

where B is the sample variance of the estimates across m imputed datasets. The relative increase in variance due to missingness is calculated by $r = \frac{(m+1)B}{\sum_j V_j}$ and the relative efficiency is $\frac{m}{m+r}$.

Figure 3 compares the true and estimated multi-state probability curves when data are missing at random. Each curve is an average across 100 simulations. The red lines represent the true probability curves, which are obtained from the complete datasets without missing data. The blue and green lines are from the imputed datasets by applying the MVN and FCS approaches, respectively, which are both very close to the true curves. The black lines are the probability curves estimated from the observed datasets with missing data, which are clearly biased from the true curves. We display the black curves as a frame of reference to illustrate the discrepancies when applying the simple approaches that applied to the imputed data to the datasets with missing data. In practice, it is rare that simple approaches are directly applicable to missing data but more sophisticated approaches, e.g., the multi-state model developed by [9], are needed. Table 1 shows the simulation results for other estimands under MAR mechanism. For the areas under the probability curves and the transition probabilities, the estimates obtained from the imputed data are generally close to the true values. The estimates are somewhat greater than the true values around the tail end of the follow-up period, which may be due to the higher missing rate and smaller number of longitudinal measurements therein. The results of the MVN and FCS approaches are very similar to each other in general. As expected, the direct estimates from the observed data are biased. For the mean intercept and slope of the longitudinal trajectories, all estimates are close to the true values since the unweighted least-square method is known to be unbiased in the case of informative dropout. For the decline of Y one-year prior to dialysis or death, all estimates are also close to the true values but those from the imputed data have smaller variances since more subjects are involved in estimating the declines with imputed data.

Figure 4 and Table 2 show the simulation results when data are not missing at random (MNAR). In this case, we also assumed a Gamma distribution for $\mathbf{b}_1$, the slope of $Y(t)$. The estimates based on the FCS and the MVN approaches again exhibited little evidence of bias, with the exception of small deviations at the tail end of the follow-up period.

4.2. An Application to AASK

According to the scheduled timeline of measuring eGFR and proteinuria, the complete longitudinal data can be arranged chronologically as the vector,

$$\mathbf{Y} = (\text{Prot}_1, \text{eGFR}_1, \dots, \text{Prot}_9, \text{eGFR}_9, \text{eGFR}_{10}, \text{Prot}_{10}, \text{eGFR}_{11}, \text{eGFR}_{12}, \text{Prot}_{11}, \dots, \text{Prot}_{15}, \text{eGFR}_{21})^T,$$

where $eGFR_j$ denotes the j th square-root transformed eGFR measurement and $Prot_k$ denotes the k th log transformed proteinuria measurement. The baseline covariates $\mathbf{Z}$ include age, gender, smoking status, LVH, albumin, urinary acid and serum nitrogen.

Missing data are imputed using the same approaches described in Section 4.1. Figure 5 shows the eGFR trajectories of 4 randomly chosen subjects. Each black dot represents an observed value and a blue point represents an imputed value. Three subjects have observed event times while the bottom-right panel shows imputed event times (dashed orange lines). The top-right panel illustrates a case in which the observed time of dialysis is used to impute prior missing eGFR data. These results are from the MVN approach; similar results are obtained using the FCS approach. Figure 6 illustrates the joint imputation of missing data in the bivariate longitudinal outcome of proteinuria and eGFR measurements for three AASK subjects. Figure 1 and 2 in the Supplementary Materials show the means of the longitudinal variables and the cumulative event probabilities over 100 iterations, which suggest satisfactory convergence after 10 iterations.

The simulation results in Section 4.1 suggest that we can apply simple approaches to imputed datasets to obtain satisfactory estimates. On the other hand, there may exist more sophisticated approaches for estimating the same parameters based on observed dataset with missing values. For example, the multi-state probabilities can be estimated using the approach proposed by [9]. We refer it as the “HLT” approach in this section. More details about the HLT approach can be found in [9]. Table 3 compares the estimated multi-state probabilities by applying the simple approach to the imputed datasets and the HLT approach to the originally observed datasets with missing values, respectively. Bootstrap procedure is needed to obtain standard errors for the HLT approach. The estimates from these two approaches are close to each other, suggesting that simple approach when applied to the “complete” imputed dataset can yield similar results as the more sophisticated HLT approach.

5. Discussion

Missing data are usually inevitable in clinical studies and multiple imputation has been a popular method to handle missing data. In this paper, we discuss multiple imputation for missing longitudinal and survival data simultaneously when analyses are complicated by informative dropout, non-ignorable missingness, etc. Many joint modeling approaches have been proposed to analyze such data over the past three decades. However, applications of joint models in the medical literature have been limited due to their complexity in both implementation and interpretation. The proposed MI approach can serve as an alternative to the joint modeling approaches. With complete imputed datasets, simple and transparent statistical approaches can be applied to obtain satisfactory results, as shown by the simulation studies and an application to the AASK study.

Our approach is easy to implement using mainstream software such as SAS and R. Furthermore, the parametric imputation approach is able to naturally accommodate multivariate longitudinal outcomes and competing-risk survival outcomes. Another

flexibility of our approach is that investigators may choose own preferred imputation models at each step.

Our MI approach is analogous to the chained equation approach [32, 29]. The chained equation approach is a flexible alternative to the common MI methods that require an appropriate multivariate distribution of the variables for imputation, which is often not applicable in the presence of non-normal or categorical variables [37]. However, theoretical properties of the chained equation approach have not been established in general [29]. Our approach shares the same limitation. Our procedure may not guarantee consistent estimation in all situations, but as long as good imputations are created, the proposed method should reduce, to a great extent, any bias due to informative dropout, censoring, and intermittently missingness.

As is the case for multiple imputation in general, it is advisable to include all variables that enter into the analysis of the complete data into the imputation model [31, 38]. Therefore one should include longitudinal outcomes as covariates in the imputation models for survival outcomes and vice versa. In practice, model selection procedures may be used to tailor specific imputation models. For example, one could include in the imputation model all variables that significantly predict the incomplete variable or those that predict the missingness of the incomplete variable. The validity of the imputation models can be checked with standard diagnostic techniques. For example, residual plots might be drawn for each imputation model based on data from the final iteration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was sponsored by grant R01DK090046 from the National Institutes of Health.

References

1. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; New York: 1987.
2. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. Wiley; Hoboken, NJ: 2002.
3. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madrid, Spain). 2009; 18(1):1–43.
4. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica. 2004; 14:809834.
5. Vonesh EF, Greene T, Schluchter MD. Shared parameter models for the joint analysis of longitudinal data and event times. Statistics in Medicine. 2006; 25:143163.
6. Hsieh F, Tseng YK, Wang JL. Joint modeling of survival and longitudinal data: Likelihood approach revisited. Biometrics. 2006; 62:1037–1043. [PubMed: 17156277]
7. Li L, Hu B, Greene T. A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. Biometrics. 2009; 65:737–745. [PubMed: 19173700]
8. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine. 2007; 26:2389–2430. [PubMed: 17031868]

9. Hu B, Li L, Wang X, Greene T. Nonparametric multistate representations of survival and longitudinal data with measurement error. *Statistics in Medicine*. 2012; 31:2303–2317. [PubMed: 22535711]
10. Rizopoulos D, Verbeke G, Lesaffre E. Fully exponential Laplace approximations for the joint modeling of survival and longitudinal data. *Journal of the Royal Statistical Society (Series B)*. 2009; 71:637–654.
11. Rizopoulos D. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Computational Statistics & Data Analysis*. 2012; 56:491–501.
12. Rizopoulos, D. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton: Chapman and Hall/CRC; 2012.
13. Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics*. 2003; 28:353368.
14. Schafer, JL. Multiple imputation with PAN. In: Sayer, AG.; Collins, LM., editors. *New Methods for the Analysis of Change*. Washington, DC: American Psychological Association; 2001. p. 355–377.
15. Yang X, Li J, Shoptaw S. Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Statistics in Medicine*. 2008; 27:2826–49. [PubMed: 18205247]
16. Ali MW, Siddiqui O. Multiple imputation compared with some informative dropout procedures in the estimation and comparison of rates of change in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*. 2000; 10:165–181. [PubMed: 10803723]
17. Paik MC, Tsai WY. On using the Cox proportional hazards model with missing covariates. *Biometrika*. 1997; 84:579593.
18. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 1999; 18:681694.
19. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*. 2009; 28(15):1982–98. [PubMed: 19452569]
20. Taylor JMG, Murray S, Hsu C-H. Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*. 2002; 58(3):221–232.
21. Xiang F, Murray S, Liu X. Analysis of transplant urgency and benefit via multiple imputation. *Statistics in Medicine*. 2014; doi: 10.1002/sim.6250
22. Faucett CL, Schenker N, Taylor JM. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics*. 2002; 58:37–47. [PubMed: 11892691]
23. Hsu CH, Taylor JM, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*. 2006; 25:3503–3517. [PubMed: 16345047]
24. Rizopoulos D, Verbeke G, Molenberghs G. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*. 2010; 66:20–29. [PubMed: 19459832]
25. Appel LJ, Middleton J, Miller ER, Lipkowitz M, Norris K, Agodoa LY, Bakris G, Douglas JG, Charleston J, Gassman J, Greene T, Jamerson K, Kusek JW, Lewis JA, Phillips RA, Rostand SG, Wright JT. The rationale and design of the AASK cohort study. *J Am Soc Nephrol*. 2003; 14(Suppl 2):S166S172.
26. K/DOQI clinical practice guidelines for chronic kidney disease. National Kidney Foundation; 2002.
27. Kurland B, Heagerty P. Directly parameterized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths. *Biostatistics*. 2005; 6:241–258. [PubMed: 15772103]
28. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. 1987; 82(398):528–540.
29. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011; 30:377–399. [PubMed: 21225900]
30. van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*. 2008; 14(4):447463.

31. Schafer, JL. Analysis of Incomplete Multivariate Data. Chapman & Hall; London: 1997.
32. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007; 16(3):219–242. [PubMed: 17621469]
33. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011; 45(3):1–67.
34. Brand, JPL. PhD thesis. Erasmus University; Rotterdam: 1999. Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.
35. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*. 2006; 76(12): 10491064.
36. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*. 1989; 45:939–955. [PubMed: 2486189]
37. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*. 2007; 26:1368–1382. [PubMed: 16810713]
38. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*. 2006; 59:1092–1101. [PubMed: 16980150]

Figure 1.
Percentages of observed and missing data at follow-up visits in AASK

Figure 2.
Average missing rates of longitudinal outcomes in the simulation study (MAR)

Figure 3.
Comparison of true and estimated multi-state probability curves (MAR)

Figure 4.
Comparison of true and estimated multi-state probability curves (MNAR)

Figure 5.
eGFR trajectories of a random sample of 4 AASK subjects

Figure 6.
eGFR and proteinuria trajectories of three AASK subjects

Table 1

Simulation results when data are missing at random (MAR)

Estimand	True		Observed		$\hat{Q}$		SE		SD		$\hat{Q}$		SE		SD		RE		SD		RE	
	Q	$\hat{Q}$	SE	SD	$\hat{Q}$	$\hat{Q}$	SE	SD	SE	SD	$\hat{Q}$	$\hat{Q}$	SE	SD	SE	SD	RE	RE	SE	SD	RE	RE
Area under the multi-state probability curve, C_k																						
$k = 1$	14.459	13.413	1.309	1.554	14.431	1.381	1.531	1.382	1.539	0.997	14.417	1.382	1.539	0.997	14.417	1.382	1.539	0.997	1.382	1.539	0.997	0.997
$k = 2$	49.543	47.430	1.545	1.761	49.322	1.634	1.595	1.632	1.604	0.995	49.275	1.632	1.604	0.995	49.275	1.632	1.604	0.995	1.632	1.604	0.995	0.995
$k = 3$	39.315	35.375	1.234	1.291	39.285	1.330	1.234	1.326	1.239	0.995	39.347	1.326	1.239	0.995	39.347	1.326	1.239	0.995	1.326	1.239	0.995	0.995
$k = 4$	11.772	17.039	1.384	1.447	11.919	0.975	1.067	0.998	1.070	0.998	11.912	0.974	1.070	0.998	11.912	0.974	1.070	0.998	0.974	1.070	0.998	0.998
$k = 5$	4.911	6.742	0.810	0.887	5.042	0.587	0.610	0.588	0.618	0.996	5.048	0.588	0.618	0.996	5.048	0.588	0.618	0.996	0.588	0.618	0.996	0.996
Transition probability																						
$P(s(30) = 2 s(0) = 1)$	0.542	0.387	0.044	0.047	0.543	0.049	0.047	0.985	0.543	0.049	0.047	0.985	0.543	0.049	0.047	0.985	0.543	0.049	0.047	0.985	0.985	0.985
$P(s(30) = 3 s(0) = 2)$	0.323	0.279	0.025	0.026	0.322	0.028	0.027	0.986	0.323	0.028	0.027	0.986	0.323	0.028	0.026	0.985	0.323	0.028	0.026	0.985	0.985	0.985
$P(s(60) = 3 s(30) = 2)$	0.375	0.235	0.025	0.032	0.379	0.032	0.029	0.981	0.380	0.033	0.028	0.978	0.380	0.033	0.028	0.978	0.380	0.033	0.028	0.978	0.978	0.978
$P(s(60) = 4 s(30) = 3)$	0.185	0.141	0.028	0.031	0.183	0.032	0.033	0.998	0.183	0.032	0.033	0.997	0.183	0.032	0.033	0.997	0.183	0.032	0.033	0.997	0.997	0.997
$P(s(90) = 4 s(60) = 3)$	0.289	0.207	0.028	0.035	0.289	0.032	0.030	0.991	0.288	0.032	0.030	0.991	0.288	0.032	0.030	0.991	0.288	0.032	0.030	0.991	0.991	0.991
$P(s(90) = 5 s(60) = 3)$	0.132	0.099	0.021	0.025	0.132	0.024	0.023	0.991	0.132	0.024	0.023	0.991	0.132	0.024	0.022	0.991	0.132	0.024	0.022	0.991	0.991	0.991
$P(s(120) = 4 s(90) = 3)$	0.256	0.224	0.033	0.042	0.266	0.035	0.031	0.984	0.265	0.035	0.031	0.984	0.265	0.035	0.031	0.984	0.265	0.035	0.031	0.984	0.984	0.984
$P(s(120) = 5 s(90) = 3)$	0.252	0.228	0.033	0.041	0.258	0.035	0.030	0.983	0.257	0.035	0.030	0.983	0.257	0.035	0.029	0.983	0.257	0.035	0.029	0.983	0.983	0.983
Intercept and slope of longitudinal trajectory																						
μ_{b0}	6.945	6.945	0.049	0.051	6.944	0.049	0.051	0.998	6.941	0.057	0.051	0.995	6.941	0.057	0.051	0.995	6.941	0.057	0.051	0.995	0.995	0.995
μ_{b1}	-0.025	-0.025	0.001	0.001	-0.025	0.001	0.001	0.992	-0.025	0.001	0.001	0.987	-0.025	0.001	0.001	0.987	-0.025	0.001	0.001	0.987	0.987	0.987
1-yr decline of Y prior to events																						
Dialysis	-0.034	-0.034	0.009	0.008	-0.033	0.005	0.003	0.955	-0.033	0.005	0.003	0.955	-0.033	0.005	0.003	0.955	-0.033	0.005	0.003	0.955	0.955	0.955
Death	-0.030	-0.032	0.011	0.011	-0.030	0.006	0.004	0.948	-0.030	0.006	0.004	0.949	-0.030	0.006	0.004	0.949	-0.030	0.006	0.004	0.949	0.949	0.949

SE: standard error from Rubin's rule; SD: Empirical standard error; RE: relative efficiency

Table 2

Simulation results when data are missing at random (MNAR)

Estimand	True		Observed			MVN			FCS		
	Q	$\hat{Q}$	SE	SD	$\hat{Q}$	SE	SD	RE	$\hat{Q}$	SE	RE
Area under the multi-state probability curve, C_k											
$k = 1$	14.284	12.815	1.248	1.392	14.212	1.372	1.427	0.997	14.194	1.376	0.996
$k = 2$	49.366	46.634	1.496	1.726	49.019	1.630	1.573	0.995	49.012	1.634	0.994
$k = 3$	39.456	35.433	1.232	1.299	39.559	1.325	1.364	0.994	39.586	1.331	0.993
$k = 4$	11.998	18.209	1.443	1.509	12.185	0.985	1.016	0.998	12.182	0.986	0.998
$k = 5$	4.896	6.908	0.831	0.931	5.025	0.589	0.616	0.996	5.026	0.589	0.996
Transition probability											
$P(s(30) = 2 s(0) = 1)$	0.551	0.382	0.044	0.047	0.550	0.049	0.048	0.984	0.551	0.049	0.984
$P(s(30) = 3 s(0) = 2)$	0.322	0.274	0.025	0.031	0.324	0.028	0.029	0.984	0.324	0.028	0.984
$P(s(60) = 3 s(30) = 2)$	0.381	0.221	0.025	0.030	0.387	0.033	0.028	0.976	0.387	0.034	0.975
$P(s(60) = 4 s(30) = 3)$	0.192	0.141	0.028	0.034	0.189	0.032	0.035	0.997	0.189	0.032	0.997
$P(s(90) = 4 s(60) = 3)$	0.290	0.209	0.028	0.036	0.290	0.032	0.032	0.991	0.290	0.032	0.989
$P(s(90) = 5 s(60) = 3)$	0.133	0.095	0.020	0.023	0.131	0.024	0.022	0.991	0.131	0.024	0.990
$P(s(120) = 4 s(90) = 3)$	0.255	0.219	0.033	0.035	0.266	0.035	0.028	0.984	0.266	0.035	0.984
$P(s(120) = 5 s(90) = 3)$	0.250	0.218	0.033	0.046	0.256	0.035	0.033	0.984	0.256	0.035	0.983
Intercept and slope of longitudinal trajectory											
μ_{b0}	6.945	6.945	0.049	0.054	6.945	0.050	0.055	0.997	6.944	0.050	0.997
μ_{b1}	-0.025	-0.025	0.001	0.001	-0.025	0.001	0.001	0.989	-0.025	0.001	0.987
1-yr decline of Y prior to events											
Dialysis	-0.034	-0.033	0.010	0.010	-0.033	0.005	0.003	0.956	-0.032	0.006	0.952
Death	-0.031	-0.034	0.011	0.013	-0.030	0.006	0.004	0.950	-0.030	0.006	0.946

SE: standard error from Rubin's rule; SD: Empirical standard error; RE: relative efficiency

Table 3

Comparison of multi-state probabilities estimated using the simple and HLT approaches

State	t	Simple Approach		HLT Approach	
		$\hat{Q}$	SE	$\hat{Q}$	SE
1	30	0.061	0.007	0.061	0.007
	60	0.168	0.011	0.167	0.011
	90	0.235	0.013	0.238	0.013
	120	0.298	0.015	0.305	0.014
2	30	0.036	0.006	0.037	0.006
	60	0.077	0.008	0.077	0.008
	90	0.125	0.010	0.122	0.010
	120	0.166	0.012	0.162	0.011
3	30	0.175	0.011	0.186	0.013
	60	0.159	0.011	0.171	0.013
	90	0.150	0.011	0.164	0.012
	120	0.116	0.013	0.137	0.012
4	30	0.521	0.016	0.509	0.017
	60	0.469	0.016	0.467	0.016
	90	0.405	0.016	0.399	0.015
	120	0.344	0.018	0.324	0.017
5	30	0.206	0.012	0.207	0.013
	60	0.127	0.010	0.118	0.012
	90	0.084	0.009	0.078	0.008
	120	0.077	0.010	0.082	0.014