

RESEARCH ARTICLE

Open Access



A Bayesian natural cubic B-spline varying coefficient method for non-ignorable dropout

Camille M. Moore^{1,2*} , Samantha MaWhinney², Nichole E. Carlson² and Sarah Kreidler³

Abstract

Background: Dropout is a common problem in longitudinal clinical trials and cohort studies, and is of particular concern when dropout occurs for reasons that may be related to the outcome of interest. This paper reviews common parametric models to account for dropout and introduces a Bayesian semi-parametric varying coefficient model for exponential family longitudinal data with non-ignorable dropout.

Methods: To demonstrate these methods, we present results from a simulation study and estimate the impact of drug use on longitudinal CD4⁺ T cell count and viral load suppression in the Women's Interagency HIV Study. Sensitivity analyses are performed to consider the impact of model assumptions on inference. We compare results between our semi-parametric method and parametric models to account for dropout, including the conditional linear model and a parametric frailty model. We also compare results to analyses that fail to account for dropout.

Results: In simulation studies, we show that semi-parametric methods reduce bias and mean squared error when parametric model assumptions are violated. In analyses of the Women's Interagency HIV Study data, we find important differences in estimates of changes in CD4⁺ T cell count over time in untreated subjects that report drug use between different models used to account for dropout. We find steeper declines over time using our semi-parametric model, which makes fewer assumptions, compared to parametric models. Failing to account for dropout or to meet parametric assumptions of models to account for dropout could lead to underestimation of the impact of hard drug use on CD4⁺ cell count decline in untreated subjects. In analyses of subjects that initiated highly active anti-retroviral treatment, we find that the estimated probability of viral load suppression is lower in models that account for dropout.

Conclusions: Non-ignorable dropout is an important consideration when analyzing data from longitudinal clinical trials and cohort studies. While methods that account for non-ignorable dropout must make some unavoidable assumptions that cannot be verified from the observed data, many methods make additional parametric assumptions. If these assumptions are not met, inferences can be biased, making more flexible methods with minimal assumptions important.

Keywords: Reversible jump Markov chain Monte Carlo, Missing data, Dropout, Varying coefficient model, HIV

*Correspondence: mooreca@njhealth.org

¹Center for Genes, Environment and Health, National Jewish Health, 1400 Jackson St., 80206 Denver, CO, USA

²Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 17th Pl., Mail Stop B119, 80045 Aurora, CO, USA

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Dropout is a common problem in longitudinal clinical trials and cohort studies, and is of particular concern when dropout occurs for reasons that may be related to the outcome of interest. For example, HIV studies are often longitudinal in nature, and it is well documented that many subjects have missing observations due to death or disease progression, leading to concerns of non-ignorable dropout [1]. Dropout is not ignorable when missingness depends on the values of the unobserved outcomes, even after conditioning on the available data [2]. In this scenario, standard longitudinal data analyses can produce biased results.

This work was motivated by the challenges associated with comparing laboratory markers of HIV disease progression and treatment response between drug users and other subjects in the Women's Interagency HIV Study (WIHS). Illicit and recreational drug use has been hypothesized to accelerate HIV disease progression by directly enhancing virus replication and by impairing immune responses. While laboratory in vitro and animal studies suggest that drug and alcohol use impairs immune function and increases HIV replication, results from epidemiological studies have been mixed [3]. These conflicting results may be in part linked to differential dropout between drug users and other subjects. Similar dropout related challenges have been identified in quality of life data from clinical trials of cancer therapies [4], antidepressant clinical trials [5] and studies of smoking cessation programs [6], among others. Considering the potential impact of non-ignorable dropout on the results of statistical analyses is particularly important in this context.

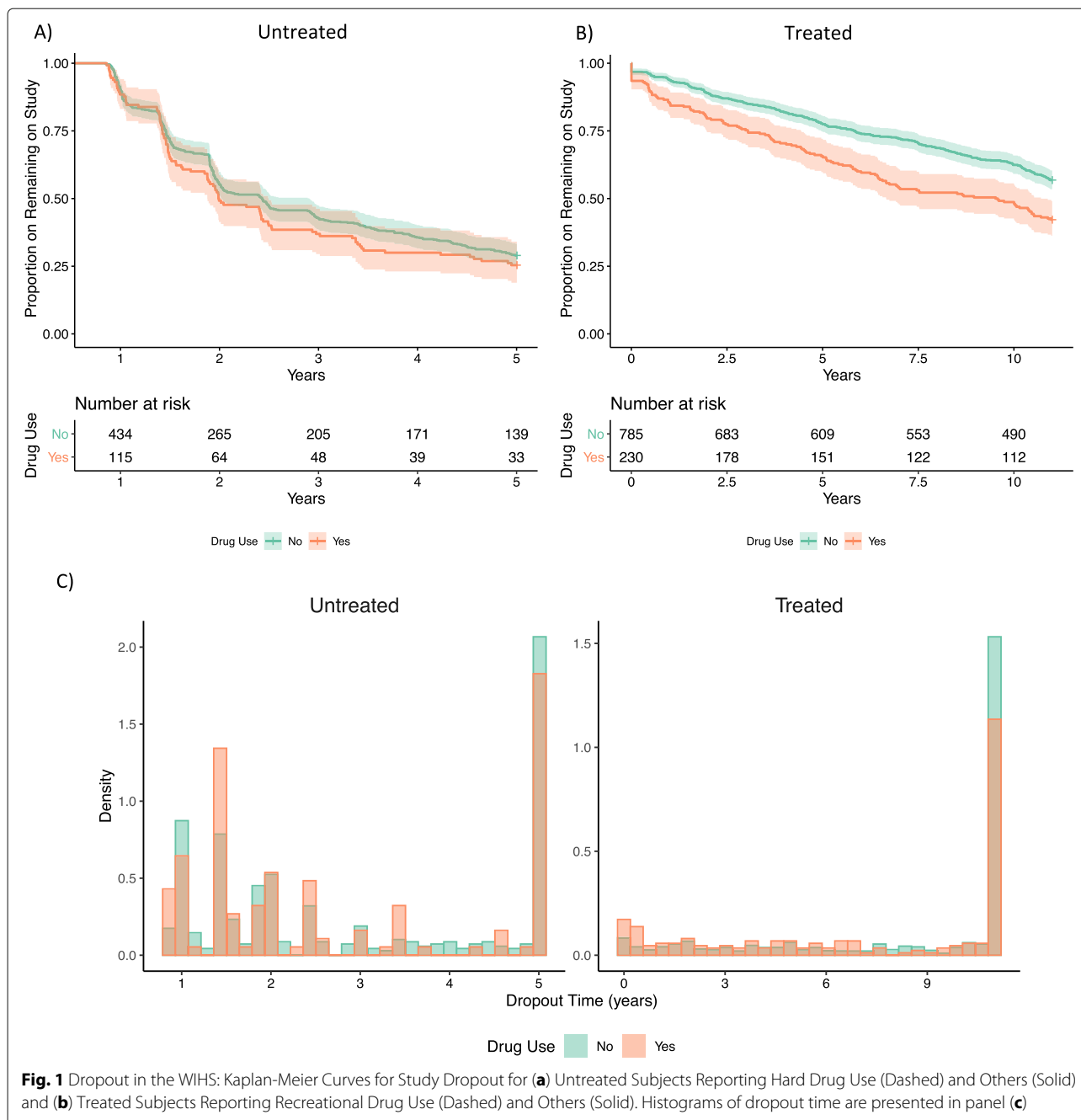
While all methods that account for non-ignorable dropout rely on unavoidable assumptions that cannot be verified from the observed data [7], many methods make additional parametric assumptions about the distribution of dropout times or the functional form of the relationship between regression coefficients and dropout time. This paper reviews common parametric methods to account for non-ignorable dropout and introduces a Bayesian semi-parametric varying coefficient generalized linear mixed model to more flexibly accommodate dropout. This method extends existing frequentist natural cubic B-spline varying coefficient methods to account for dropout in longitudinal studies with a Gaussian outcome [3, 8] to other non-normal outcomes in the exponential family. Fitting the model in a Bayesian framework allows the number and location of spline knots to be jointly modeled with other model parameters, removing dependence on the choice of knots and more accurately characterizing model uncertainty. We illustrate how inference differs between parametric and semi-parametric models to account for dropout in the analysis of longitudinal changes

in CD4⁺ T cell count and viral load suppression in the WIHS.

Background on the WIHS

The WIHS is an ongoing prospective study of the natural and treated histories of HIV infection in women, with behavioral data and specimens collected at semiannual visits by multiple sites since 1994 [9]. In contrast to male populations, HIV and AIDS are more prevalent among women of color exposed through heterosexual partners or intravenous drug use [10, 11]. Two common measures of disease progression measured in the WIHS are CD4⁺ T cell count, a measure of immunologic health, and viral load, a measure of the concentration of HIV-1 RNA in the blood. For HIV⁺ subjects that have initiated highly active antiretroviral therapy (HAART), the primary measure of treatment effectiveness is suppression of viral load (HIV-1 RNA below detection limits). The goal of our analyses is to understand the impact of drug use on disease progression and treatment response. In untreated subjects, we compare longitudinal changes in CD4⁺ T cell count and viral load suppression between subjects that report hard drug use and other subjects in the WIHS, as there is evidence to suggest that hard drug use in particular can dampen immune response and increase virus replication. Rates of treatment initiation among hard drug users are lower or treatment occurs later for a variety of reasons, including provider perceptions that they are unable to keep appointments, are not ready for treatment, have unstable living situations, are unable to fill prescriptions or have limited ability to adhere to treatment. In addition, non-physician providers, are more likely to care for illicit drug users and to resist prescribing HAART [12]. Thus, the number of hard drug users initiating treatment is limited. In addition, any recreational drug use may potentially reduce compliance to HAART regimens. Therefore, for treated subjects in the WIHS, we compare longitudinal viral load suppression between recreational drug users and others.

In our initial investigation into the data, we found several causes for concern. We noted that a large proportion of subjects dropped out of the study early, with half of untreated subjects lost by 2.4 years (median of 4 observations, Fig. 1a) and a quarter of treated subjects lost by 5 years (median of 19 observations, Fig. 1b) after treatment initiation. In addition, drug users tended to drop out of the study earlier than other subjects and were more likely to die within 1 year of their last study observation (Table 2). Due to the prevalence and differential distribution of dropout, missing data could have a large impact on the results of our analysis. Untreated subjects that dropped out of the study had lower mean CD4⁺ at their last visit compared to subjects that remained on study (Fig. 2a), and treated subjects that dropped out of the study early were less likely to have suppressed viral load (Fig. 2b). This



suggests that subjects that dropped out may have done so due to more rapidly deteriorating health, raising concerns of non-ignorable dropout.

Methods

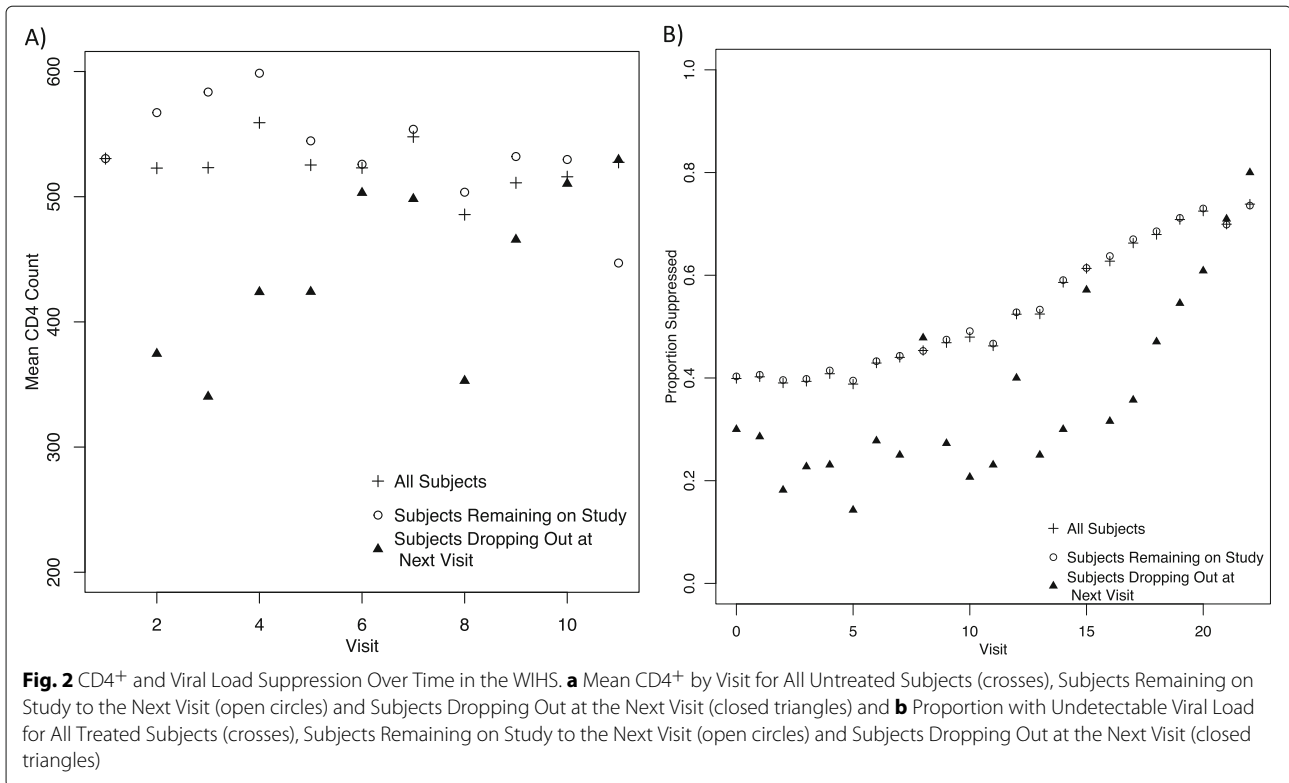
Background on methods to account for missing data

Dropout is not ignorable and data are missing not at random when missingness depends on the values of the unobserved outcomes, even after conditioning on the available data [2]. Selection, frailty and mixture models are likelihood based approaches that can account for data

that are missing not at random. While there are several methods to account for non-ignorable dropout in longitudinal studies with a Gaussian response, methods for non-normal data are less developed [13]. The literature is particularly sparse for addressing non-ignorable dropout in GLMMs in semi-parametric or Bayesian frameworks.

Selection models

Selection models factor the joint distribution of the outcomes, y , which include both observed and missing values, and missing data indicators, r , as $f(y|x)f(r|y,x)$.



Frequentist parametric selection models for Gaussian outcomes have been proposed by several authors [14–16], and parametric selection models for binary outcomes have been proposed by Ibrahim et al.[17] and Wu and Wu [18]. Identification of parameters in selection models can be challenging and relies on distributional assumptions for the outcome and a parametric relationship for how potentially missing outcome data are related to the probability of missingness. In addition, selection models typically require specialized numerical routines for maximizing the likelihood, which can limit practical utility for broad ranges of problems [6].

Frailty models

Frailty models, also called shared parameter models, factor the joint distribution of the outcome and missing data indicators as $\int f(y|x, \eta)f(r|x, \eta)dF(\eta|x)$, where η are the shared parameters or frailties that induce dependence between the outcomes and missing data indicators. Parametric frailty models have been proposed for both Gaussian and non-normal outcomes [19–23]. Identification of parameters in frailty models is driven by the parametric frailty distribution. This choice is often arbitrary, and may influence the validity of results [6]. Another key assumption of frailty models is that the repeated measures are independent of drop-out times conditional on the frailties.

For example, Schluchter proposed a two stage frailty model, which we will consider in our simulation study and application. The first stage assumes that each subject’s responses follow a linear regression with random intercept b_{i0} and slope b_{i1} , which can be written $y_i = X_i b_i + e_i$, where e_i is a vector of independent, normally distributed error terms for subject i . In the second stage, the subject-specific random coefficients and the natural log of dropout time, u_i , are modeled with a joint multivariate normal distribution:

$$\begin{pmatrix} b_i \\ \log(u_i) \end{pmatrix} \sim N \left(\mu = \begin{pmatrix} \mu_b \\ \mu_u \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_b & \sigma'_{bu} \\ \sigma_{bu} & \sigma_u^2 \end{pmatrix} \right)$$

where μ_b is the mean of the random coefficients, μ_u is the mean of the natural log of dropout time, Σ_b is the covariance matrix of the random coefficients, σ_{bu} is a row vector containing the covariances of u and each random coefficient, and σ_u^2 is the variance of the natural log of the dropout times. This model allows the underlying slope and intercept to be associated with dropout time, via the covariance parameters σ_{bu} . If these covariances are zero, then the random coefficients and dropout time are independent and dropout is assumed to be non-informative. In addition to assuming a log-normal distribution of dropout times, the model assumes that there is a linear relationship between the log of dropout time and the dropout time specific intercepts and slopes, since the random coefficients

are related to dropout time through the covariance parameters. Violations of these assumptions can lead to biased estimates and inaccurate inference.

Mixture models

Mixture models factor the joint distribution of the outcome and missing data indicators as $f(y|r,x)f(r|x)$ [19, 24–28]. Pattern mixture models [29] are a popular method to account for non-ignorable missingness when missingness can be categorized into distinct patterns. After classifying data according to missing data patterns, models can be fit to the outcome data within each pattern. Kaciroti et al. have described Bayesian pattern mixture models for binary and count data,[30–32] however, these methods may not be feasible for large numbers of dropout patterns or continuous dropout times. For example, in the WIHS, follow-up visits were intended to occur every 6 months, but the exact timing of visits varies greatly between subjects, so that observation times are not aligned and dropout may occur at any continuous point in time.

Varying coefficient models (VCM) are another mixture model approach that more easily accommodate continuous dropout times. VCMs adjust for dropout by allowing regression coefficients to depend on dropout time. For example, for a Gaussian distributed outcome, the response vector for subject i is modeled using a linear mixed model, with regression coefficients that depend on dropout time, such that $y_i = X_i\beta(u_i) + Z_i\alpha_i + e_i$, where $\beta(u_i)$ are the dropout varying regression coefficients, X_i is the design matrix for the fixed effects, Z_i is the design matrix associated with the random effects, α_i , and e_i is a vector of normally distributed error terms. If the regression coefficients are constant with respect to dropout time, the model reduces to a standard generalized linear mixed model (GLMM). Assuming regression coefficients are linear (or low-order polynomial) functions of dropout time results in Wu and Bailey’s conditional linear model (CLM) [25]. However, if the regression coefficients are not linearly related to dropout time (or the polynomial function is mis-specified) estimates can be biased [8, 33]. For Gaussian outcomes, semi-parametric varying coefficient models that only require that regression coefficients are smooth, continuous functions of dropout time have been proposed, making them more robust [8, 33]. In a Bayesian framework, methods for binary outcomes have utilized marginalized transition models for population level rather than subject-specific inference.

Bayesian varying coefficient models for non-ignorable dropout

We introduce a Bayesian natural cubic B-spline varying coefficient GLMM (BNSV) that can account for dropout in longitudinal studies with exponential family

outcomes, while avoiding assumptions about the distribution of dropout times or the functional form of the relationship between regression coefficients and dropout time, common in parametric frailty and mixture models. This method extends existing frequentist natural cubic B-spline varying coefficient methods to account for dropout in longitudinal studies with a Gaussian outcome[3, 8] to other non-normal outcomes in the exponential family. Similar models have been proposed for Gaussian outcomes using penalized splines [34].

Fitting the semi-parametric varying coefficient model in a Bayesian framework has several advantages. The number and location of spline knots control the smoothness, shape, and flexibility of the spline over the range of dropout times; however, fitting in a frequentist framework, it is unclear how to choose these parameters [35–39]. We utilize a reversible jump Markov chain Monte Carlo (RJMCMC) approach that jointly models the number and location of knots for the spline and does not require the choice of a single set of spline knots to make statistical inference. In addition, there is no need to specify a parametric distribution for the dropout times or to use an extra bootstrap simulation to estimate standard errors, as is required in the frequentist, semi-parametric approach.

VCM for longitudinal exponential family outcomes

Let $y = (y_1 \dots y_m)'$ be the set of outcomes observed on m subjects with n_i observations each at times $t = (t_1 \dots t_m)'$. Let $u = (u_1 \dots u_m)'$ be the set of m observed dropout times. First we describe the conditional model for $y|u$, which allows the change in the outcome over time to depend on dropout time and results in dropout time specific estimates.

For exponential family outcomes, the observation specific conditional VCM is:

$$f(y_{ij}|u_i, \alpha_i, \eta_{ij}) = \exp \left[\frac{y_{ij}\eta_{ij} - b(\eta_{ij})}{\phi} \right] c(y_{ij}, \phi)$$

$$\mu_{ij} = E(y_{ij}|\eta_{ij}) = b'(\eta_{ij})$$

$$g(\mu_{ij}) = \eta_{ij} = \beta_0 + \beta_1(u_i)t_{ij} + C_{ij}\beta_C + Z_{ij}\alpha_i \quad (1)$$

where $g(\cdot)$ is the link function, η_{ij} is the linear predictor, Z_{ij} is the design matrix associated with the random effects, α_i , and ϕ is a scale parameter. For a model with a random intercept and slope, let $\alpha_i = \begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$. β_0 is the intercept, and $\beta_1(u_i)$ is the dropout-varying slope. C_{ij} is the design matrix associated with the covariate effects, β_C , which do not depend on dropout time.

Natural cubic b-splines

The slope, $\beta_1(u_i)$, in Eq. 1 is assumed to be a smooth function of dropout time and is modeled using natural

cubic B-splines [40]. The i th subject’s dropout-time specific slope is $\beta_1(u_i) = \sum_{k=1}^{D+1} \theta_k \tilde{B}(\mathbf{u}, D, \mathbf{l})_{[i,k]}$. Here D is the number of degrees of freedom for the dropout-varying component of the slope and $\tilde{B}(\mathbf{u}, D, \mathbf{l})$ is the matrix of natural cubic B-spline basis functions evaluated at \mathbf{u} with $D + 1$ knots (including 2 boundary knots) at locations $\mathbf{l} = \{l_1, \dots, l_{D+1}\}$, for $D \geq 1$. For $D = 0$, there is no dropout-varying effect and $\tilde{B}(\mathbf{u}, D, \mathbf{l})_{[i,1]} = 1$ for all subjects. $\boldsymbol{\theta} = (\theta_1 \dots \theta_{D+1})$ are the coefficients associated with the basis functions.

Dropout time model and Bayesian bootstrapping

While inference conditional on \mathbf{u} can be made without assumptions about the distribution of \mathbf{u} , it is often of interest to summarize the results with a marginal or “dropout adjusted” estimate of the outcome that does not depend on dropout time, which requires integrating over the distribution of dropout times. We utilize Rubin’s Bayesian bootstrap method [41] to flexibly model the distribution of dropout times, to estimate the proportion of subjects dropping out at each observed dropout time, and to calculate marginal estimates in a straightforward manner [34].

The Bayesian bootstrap repeatedly samples the proportion of subjects dropping out at each of the observed dropout times, rather than re-sampling the observed dropout times themselves, as would be done in a frequentist bootstrap. Define $\mathbf{u}^0 = (u_1^0, \dots, u_R^0)$ as the R unique ordered observed dropout times. Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_R)$ be the vector of probabilities of dropping out at each observed dropout time and $\mathbf{N} = (N_1, \dots, N_R)$ be the number of subjects observed dropping out at each unique dropout time. The likelihood is proportional to $\prod_{r=1}^R \pi_r^{N_r}$. If we assume the prior distribution of $\boldsymbol{\pi}$ is proportional to $\prod_{r=1}^R \pi_r^{-1}$, the posterior distribution of $\boldsymbol{\pi}$ is proportional to $\prod_{r=1}^R \pi_r^{N_r-1}$, which is the kernel of a Dirichlet distribution. The posterior distribution of $\boldsymbol{\pi}$ is then Dirichlet with concentration parameters (N_1, \dots, N_R) .

Calculation of marginal effects

Working on the linear predictor scale, it is possible to calculate a marginal slope, averaged over both the distribution of dropout times and random effects. Note that the calculation of the marginal slope depends on the assumption that subjects continue on the same trajectory after their dropout. The expected value of the linear predictor at time t is:

$$\begin{aligned} E(\eta_{ij}|t, C) &= \int \int \{\beta_0 + \beta_1(u)t + C\boldsymbol{\beta}_C + \mathbf{Z}\boldsymbol{\alpha}\} dF(\boldsymbol{\alpha})dF(u|C) \\ &= \int \{\beta_0 + \beta_1(u)t + C\boldsymbol{\beta}_C\} dF(u|C) \\ &= \beta_0 + C\boldsymbol{\beta}_C + t \int \beta_1(u)dF(u|C) \end{aligned}$$

$E(\eta_{ij}|t, C)$ is also a linear function of time with slope $\beta'_1 = E[\beta_1(u)|C]$. If we assume the distribution of dropout times does not depend on the covariates ($F(u|C) = F(u)$), then $\beta'_1 = E[\beta_1(u)]$, and the marginal slope can be estimated at each iteration of the RJMCMC algorithm in a straightforward manner. At iteration s , $\beta'^{(s)}_1 = (\boldsymbol{\pi}^{(s)})^T \boldsymbol{\beta}_1(\mathbf{u}^0)^{(s)}$.

If the assumption that the distribution of dropout times does not depend on the covariates is inappropriate, it may not always be possible to easily estimate marginal slopes, particularly in more complex cases where the distribution of dropout times may depend on continuous covariates or several different covariates. However in simple cases, for example comparing the change in the outcome over time between treatment or drug use groups, marginal effects can be easily calculated, even if the distribution of dropout times depends on group. Here, the Bayesian bootstrap can be performed separately for each group and group specific marginal slopes can be calculated, as shown in our application to the WIHS.

Prior distributions

D is assumed to have a Poisson(λ) prior distribution [35]. For knot locations, we assume a discrete set of M candidates, such as the order statistics of the observed drop out times. For a given D , all sets of knots are assumed to have the same prior probability, so that $p(l_1 \dots l_{D+1}|D) = \binom{M}{D+1}^{-1} = \frac{(D+1)!(M-D-1)!}{M!}$.

The fixed effect coefficients for the natural B spline basis functions are assumed to have a multivariate normal prior with mean zero, and independent covariance structure, such that $\boldsymbol{\theta} \sim MVN_{D+1}(0, \mathbf{R}_0)$, where $\mathbf{R}_0 = \sigma^2_{\beta} I_{D+1 \times D+1}$. $I_{(D+1) \times (D+1)}$ is a $(D + 1) \times (D + 1)$ identity matrix. In practice, σ^2_{β} is chosen to be large enough to be “non-informative.” We similarly assume $(\beta_0, \boldsymbol{\beta}_C)'$ have a multivariate normal prior, $MVN(0, \sigma^2_{\beta} I)$. In addition, we assume an inverse Wishart prior for the covariance of the random effects, and for a normally distributed outcome, an inverse gamma prior for the variance of the residual error.

Estimation and implementation

A RJMCMC algorithm [42] is used to fit the BNSV model and has been implemented in the InformativeDropout R package available at <https://github.com/kreidles/informativeDropout>. Full details of the sampler and a discussion of implementation issues can be found in Section 1 of the [supplementary material](#).

Simulation study

Methods compared and data simulation

We assess the performance of the BNSV, CLM,[25] and Schluchter’s parametric frailty model[21] in estimation of the marginal slope (expected change in the outcome over

time) as well as dropout time specific slopes. We chose to compare to the CLM and parametric frailty models as these are popular methods that are straight-forward to implement.

Simulated data were generated under four different scenarios, including two normally distributed outcomes and two binary outcomes. In these four scenarios the slopes were related to dropout time by two different dropout mechanisms: (i) a continuous and smooth function meeting assumptions of the BNSV and (ii) a discontinuous step function. In addition, simulations for linear dropout-varying slopes and no dropout effect are presented in Section 2 of the [supplementary material](#) (available online) and illustrate that the BNSV can also fit CLMs and GLMMs.

More specifically, the following form for the data was assumed: $\eta_{ij} = \beta_0 + \beta_1(u_i)t_{ij} + \alpha_{0i} + \alpha_{1i}t_{ij}$, $i = 1..m, j = 1..n_i$ for m subjects with n_i observations for the i th subject, where $(\alpha_{0i}, \alpha_{1i})' \sim N(0, \Sigma_\alpha)$. For the Gaussian simulations, $y_{ij}|\eta_{ij} \sim N(\eta_{ij}, \sigma_\epsilon^2)$, and $\beta_0 = 0$. Dropout times were $u = U/15 \in [0, 1]$, resulting in 16 time points spaced equally from 0 to 1. Uniform dropout was created from a beta-binomial where $p \sim \text{Beta}(1.5, 1.5)$ and $U \sim \text{Bin}(15, p)$. The within-subject variance, σ_ϵ^2 , was set at 0.067. The elements of Σ_α were as follows: $\sigma_0^2 = 0.4$, $\sigma_1^2 = 0.01$ and $\sigma_{01} = -0.01$. These simulation settings were developed in other papers that tested methods for analyzing non-ignorable dropout in a frequentist setting. [8, 33] For the binary simulations, $y_{ij}|\eta_{ij} \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{ij}))$, $\beta_0 = -3$, and for stability, dropout began at the third observation. The elements of Σ_α were as follows: $\sigma_0^2 = 0.4$, $\sigma_1^2 = 0.1$ and $\sigma_{01} = -0.01$. The forms of the dropout-varying slope were: Normal (i): $\beta_1(u) = -3 \exp(-4u)$, Normal (ii): $\beta_1(u) = I_{(u>2/3)}$, Binary (i): $\beta_1(u) = 10\{1 - 2 \exp(-4u)\}$, Binary (ii): $\beta_1(u) = 4 + 61_{(u>2/3)}$ (Fig. 3). The magnitude of the dropout effects in these scenarios were similar to those seen in the WIHS and other typical HIV cohort studies. For each simulation scenario, 1000 datasets with 400 subjects each were created.

Methods of evaluation

The BNSV and Bayesian versions of the CLM and frailty models were fit to each dataset, as well as a naive GLMM that did not account for dropout. The performance of the methods was evaluated graphically and in terms of bias, variance, and mean square error for the marginal slope, estimated by the posterior mean. All analyses were implemented in R using custom MCMC algorithms utilizing the splines, MASS, mvtnorm, MCMCpack, and gtools packages. An R package to implement BNSV models is available at <https://github.com/kreidles/informativeDropout>.

Implementation

For the BNSV, a maximum of 10 degrees of freedom were considered for the dropout-varying component of the slope, for a maximum of 11 total degrees of freedom for the slope. The prior mean for the number of degrees of freedom for the dropout-varying component of the slope was set to 5, and the prior variance for the coefficients was set to 25 for the normal simulations, and 100 for the binary simulations. The prior for Σ_α was $IW(3, I)$ and the prior for $\frac{1}{\sigma_\epsilon^2}$ was $IG(0.001, 0.001)$. The probability of proposing a birth/dimension increase was 0.2. The MCMC chain was initiated with five equally spaced knots (including the 2 boundary knots) and coefficients set to their least squares or WLS estimates. Random effects were started at 0. Chains were run for 40,000 iterations with a burn in of 10,000 without thinning.

Results

Model performance was quantified in terms of bias, variance, and mean squared error (MSE) for the marginal slope (Table 1). The GLMM had the lowest variance, likely because the method makes unmet assumptions that simplify the model and also has the fewest parameters. The BNSV method had the lowest bias and MSE for the marginal slope in all scenarios. Graphs of the predicted BNSV, frailty model, and CLM slopes at each dropout time are presented in Fig. 3. The BNSV method was able to more accurately describe the relationship between dropout time and the slope compared to both the frailty model and the CLM, which always fits a linear relationship. While in some cases the CLM had low bias or MSE for the marginal slope, it had poor model fit and did not perform well in the estimation of dropout time specific slopes. For example, in the Binary (ii) simulation, the CLM under-estimates slopes at early dropout times and over-estimates slopes at later dropout times (Fig. 3), such that these errors are averaged out in the marginal slope calculation, despite the poor model fit.

Analysis of the WIHS data

We applied the BNSV method to investigate the impact of drug use on longitudinal HIV outcomes in the WIHS. For untreated subjects, we hypothesized that hard drug users would have steeper declines in CD4⁺ T cell count compared to other untreated subjects in the cohort. For HIV+ subjects that have initiated highly active antiretroviral therapy (HAART), the primary measure of treatment effectiveness is suppression of viral load (HIV-1 RNA below detection limits). We hypothesized that recreational drug users would have slower increases in the odds of viral load suppression compared to other subjects in the cohort.

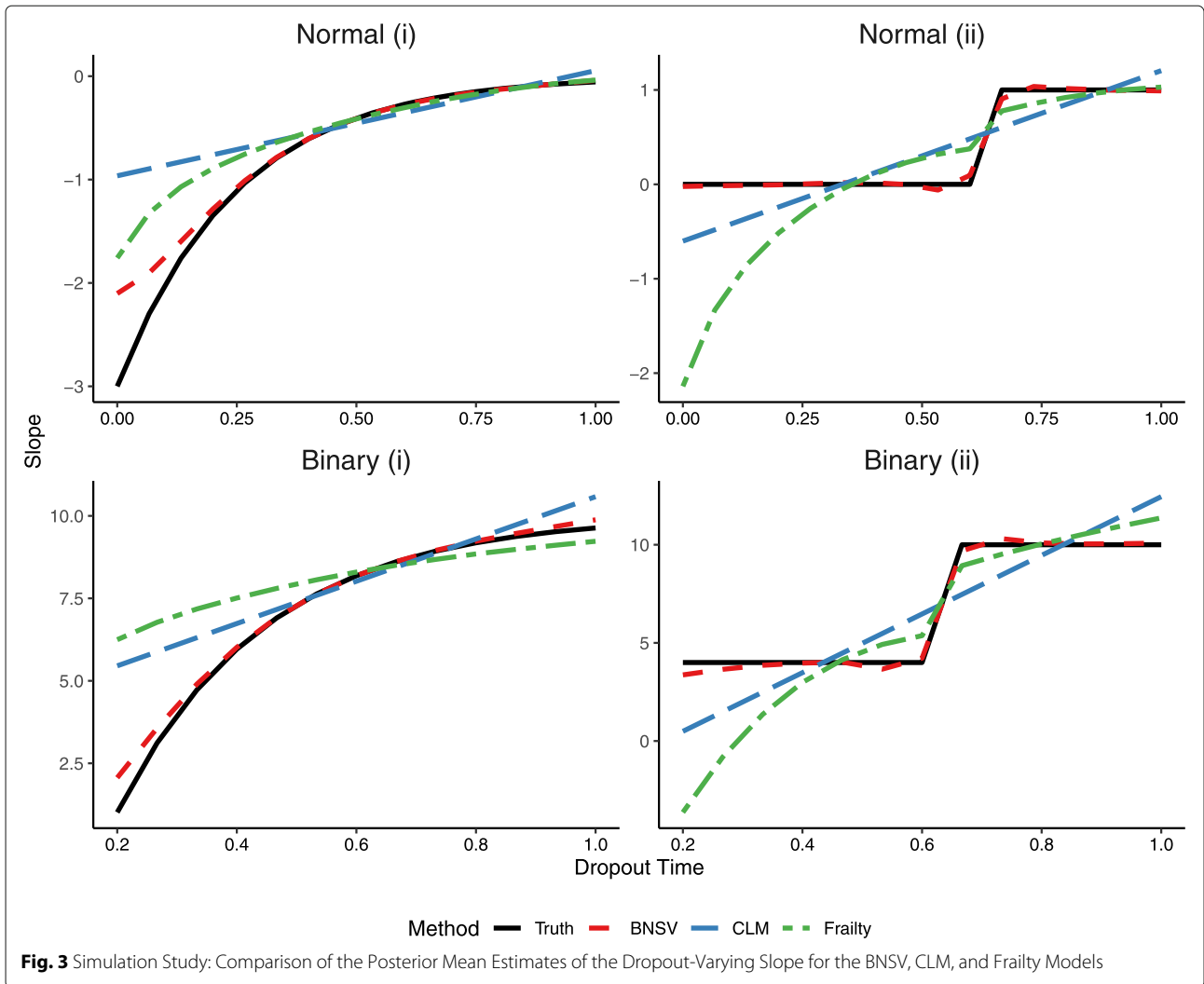


Table 1 Simulation Study Comparing BNSV, CLM, Frailty and GLMM Models: Posterior Mean Marginal Slope Estimates, Bias (Relative Bias), Variance and Mean Squared Error (MSE)

Method	Normal (i)				Normal (ii)			
	Slope	Bias	Var.	MSE	Slope	Bias	Var.	MSE
BNSV	-0.64	0.06(9%)	0.01	0.02	0.35	-0.002 (-1%)	0.01	0.01
CLM	-0.45	0.25 (36%)	0.003	0.06	0.30	-0.05 (-14%)	0.03	0.03
Frailty	-0.52	0.19 (27%)	0.004	0.04	0.16	-0.19 (-54%)	0.01	0.04
GLMM	-0.21	0.49 (70%)	0.001	0.24	0.65	0.31(89%)	0.002	0.10
Method	Binary (i)				Binary (ii)			
	Slope	Bias	Var.	MSE	Slope	Bias	Var.	MSE
BNSV	7.49	0.13 (2%)	0.19	0.20	6.67	-0.05 (-1%)	0.17	0.18
CLM	8.02	0.66 (9%)	0.11	0.54	6.47	-0.24 (-4%)	0.14	0.20
Frailty	8.14	0.78 (11%)	0.14	0.75	6.04	-0.96 (-14%)	0.18	1.11
GLMM	9.09	1.74 (24%)	0.09	3.11	8.56	1.85 (28%)	0.14	3.55

Methods

We utilized the BNSV method to compare longitudinal changes in CD4⁺ count between consistent hard drug users and other untreated subjects and to compare viral load suppression between consistent recreational drug users and other treated subjects in the WIHS while accounting for dropout. Subjects were classified as consistent hard drug users if they reported injection or non-injection use of cocaine, opiate or amphetamine use at 50% or more of visits combined with use within the last year before dropout. Subjects were classified as consistent recreational drug users if they reported marijuana, or use of cocaine, opiate, amphetamine, or other drugs at 50% or more of visits combined with use within the last year before dropout. Dropout time was calculated as the day of the last visit + 1. Descriptive statistics are presented in Table 2.

Ln(CD4⁺) was modeled for untreated subjects from the initial WIHS cohort (first recruitment period) for the first 5 years of the study, beyond which many of the subjects had missing data. If a subject remained on study for longer than 5 years a dropout time of 5 years +1 day (1826 days) was assigned. In addition to hard drug use, baseline ln(CD4⁺) and its interaction with time were included as covariates in the model. Viral load suppression was modeled for subjects that initiated treatment between 1995 and 2000 for all visits up to 11 years after initial treatment initiation, when many subjects no longer had available data. Again, if a subject remained on study for longer than 11 years a dropout time of 11 years +1 day was assigned. Since detection limits of viral load assays changed over time, viral loads under 400 copies/mL were considered “undetectable”. Baseline ln(CD4⁺) and log₁₀(viral load) (measurements preceding treatment initiation) and their interactions with time were included as covariates in the model.

Different dropout-varying slopes and dropout time distributions were allowed for drug users and other

subjects. The RJMCMC chains were run for 200,000 iterations, with a burn in of 50,000 iterations. A Poisson prior with a mean of 5 was used for the number of knots in the model. Spline coefficients and covariates were updated in separate blocks. Normal distributions with mean 0 and variance of 100 were used as priors for the coefficients to be “non-informative.” Slopes on the linear predictor scale, averaged over dropout time, were calculated using the Bayesian bootstrap method. For comparison, CLMs and Schluchter’s frailty models, as well as GLMMs that did not account for dropout, were fit to the data using a similar MCMC estimation algorithm.

Results

Longitudinal CD4⁺ count

Consistent hard drug users tended to dropout of the study earlier and were more likely to dropout due to death (Table 2). Analyses accounting for dropout with the BNSV show that overall, hard drug users had more rapid declines in CD4⁺ count than those who did not use hard drugs (Fig. 4a). Assuming a baseline CD4⁺ count of 478.5 (median), hard drug users CD4⁺ counts declined by 33.5% per year (95% CI: 25.0–41.2) compared to 17.8% (95% CI: 14.9–20.7) for others in the WIHS (Table 3). Comparing these results to a linear mixed-effects model, declines in CD4⁺ were steeper and the magnitude of the difference between hard drug users and non-users was larger in the BNSV model (Fig. 4a). Using a linear mixed model, hard drug users were found to have 22.4% declines in CD4⁺ count per year (95% CI: 17.8–26.8) compared to 14.6% (95% CI: 12.1–17.0) in others. For subjects that did not report hard drug use, the changes in CD4⁺ count per year estimated using the CLM and frailty models were similar to the BNSV; however for subjects that did report hard drug use, the BNSV estimated larger declines in CD4⁺ count than the CLM or frailty model. For the CLM, this difference can be explained by the larger declines pre-

Table 2 Demographic Characteristics of Untreated Subjects with HIV Disease in the WIHS by Drug Use Group. Median (Interquartile Range) or Percent (N)

	Untreated		HAART	
	Other (N=566)	Drug User (N=248)	Other (N=785)	Drug User (N=230)
Age (years)	35.1 (30.2–40.4)	37.8 (33.0–42.5)	38.7 (33.7–43.6)	40.5(34.9–45.7)
Baseline CD4 ⁺	499 (320–688)	417 (272–594)	271 (150–424)	256 (126–423)
Baseline log ₁₀ (Viral Load)	4.0 (3.1–4.7)	4.3 (3.8–4.9)	4.1 (3.3–4.9)	4.3 (3.4–5.0)
Dropout Time (days)	713.5 (394.2–1818.0)	404.5 (208.8–882.5)	4015.0 (2136.0–4015.0)	3485.0 (1112.0–4015.0)
Minority	84.1%(476)	81.0% (201)	84.3% (662)	76.1% (175)
Died within 1 year of dropout	4.2% (24)	11.3% (28)	24.2% (150)	34.8% (80)

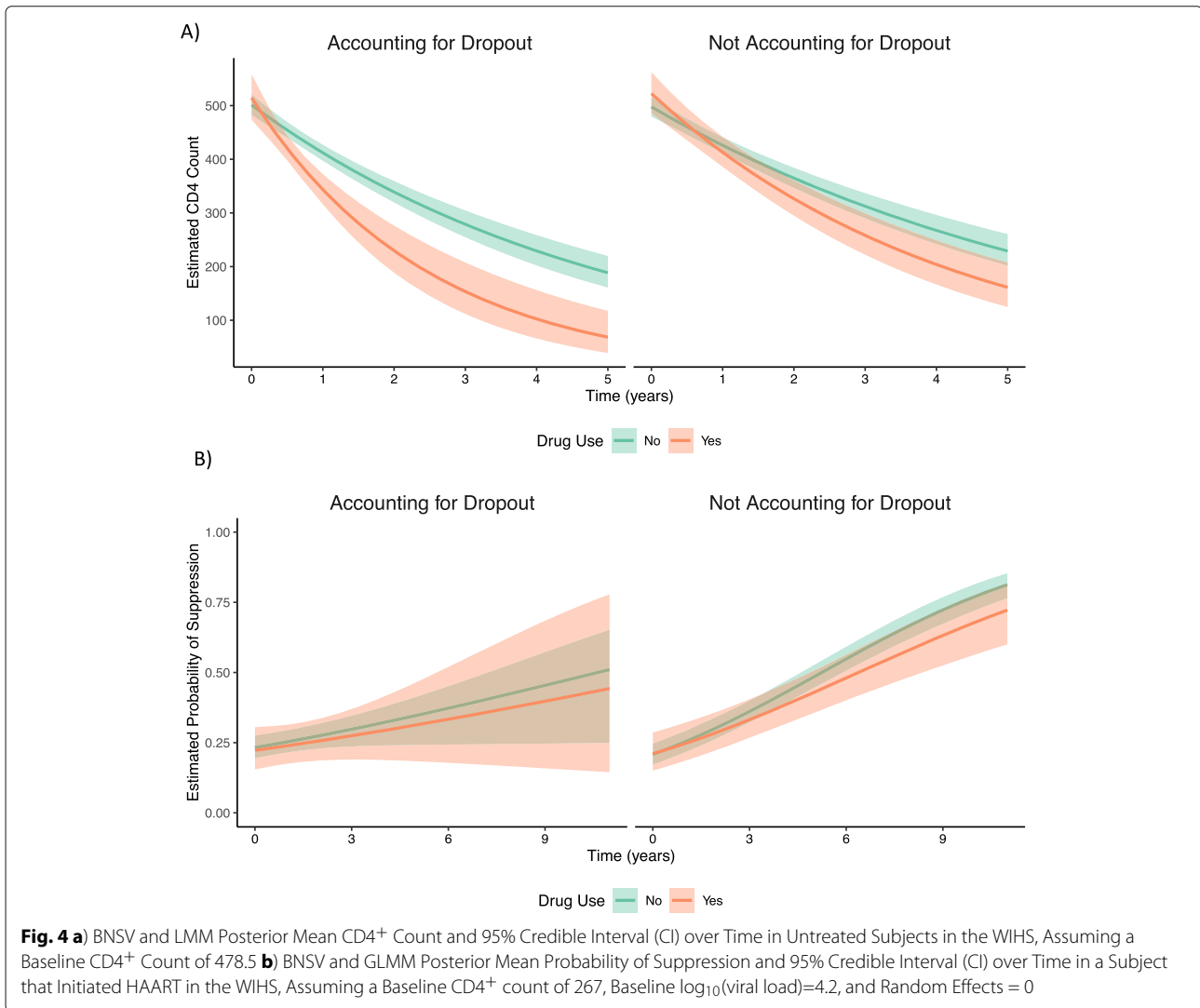


Table 3 Estimated Changes in ln(CD4⁺) and ln(Odds of Viral Load Suppression) per Year for Untreated Subjects in the WIHS Using the BNSV and GLMM Methods. Changes in ln(CD4⁺) assume a baseline CD4⁺ of 478.5. Changes in ln(odds) assume baseline CD4⁺ of 267 and baseline log₁₀(viral load)=4.2. PM=posterior mean, CI=credible interval, PP=posterior probability of a difference < 0, indicating steeper declines in CD4⁺ and less rapid increases in odds of viral load suppression among drug users

	Others		Drug users		Difference		PP
	PM	95% CI	PM	95% CI	PM	95% CI	
A) Untreated: Δ ln(CD4⁺)							
BNSV	-0.20	(-0.23, -0.16)	-0.41	(-0.53, -0.29)	-0.21	(-0.34,-0.09)	0.9996
Frailty Model	-0.19	(-0.22, -0.15)	-0.27	(-0.33, -0.22)	-0.08	(-0.15, -0.02)	0.99
CLM	-0.20	(-0.22, -0.17)	-0.27	(-0.32, -0.22)	-0.07	(-0.13, -0.02)	0.99
GLMM	-0.16	(-0.18, -0.13)	-0.24	(-0.29, -0.18)	-0.08	(-0.14, -0.02)	0.99
B) HAART: Δ ln(Odds of Suppression)							
BNSV	0.11	(0.005, 0.17)	0.09	(-0.06, 0.25)	-0.02	(-0.18,0.15)	0.61
Frailty Model	0.18	(0.14, 0.23)	0.15	(0.05, 0.25)	-0.03	(-0.14, 0.07)	0.74
CLM	0.12	(0.06, 0.17)	0.08	(-0.04, 0.19)	-0.04	(-0.16, 0.09)	0.73
GLMM	0.26	(0.22,0.29)	0.21	(0.14, 0.28)	-0.05	(-0.13, 0.03)	0.90

dicted by the BNSV for recreational drug users with early dropout times (Fig. 5a). For the frailty model, this is likely due to the lack of fit of the lognormal distribution for dropout times.

Longitudinal viral load suppression

Consistent recreational drug users also tended to dropout of the study earlier and were more likely to dropout due to death than other subjects that initiated HAART (Table 2). The average change in the log odds of viral load suppression per year assuming a median baseline CD4⁺ count of 267 and log₁₀(viral load) of 4.2 are presented in Table 3. The probability of suppression for a subject with the average slope, baseline CD4⁺ count of 267 and log(viral load) of 4.2 are shown in Fig. 4b. For both recreational drug users and other subjects that initiated HAART, the estimated probability of viral load suppression as well as the change in the odds of suppression over time were reduced using the BNSV method to account for

dropout compared to a standard GLMM; however, estimated differences in the change in odds of suppression over time between drug users and others are similar for the two models. For a recreational drug user with baseline CD4⁺ count of 267, log(viral load) of 4.2, and random effects of 0, the odds of viral load suppression increased by 1.07 times per year (95% CI: 0.93 to 1.21), compared to 1.12 times per year (95% CI: 1.06 to 1.19) for a subject with the same covariates that did not use recreational drugs. Using a standard GLMM that did not account for dropout, these estimates were 1.23 (95% CI: 1.15 to 1.32) and 1.29 (95% CI: 1.24 to 1.34) respectively. While recreational drug users had smaller increases in the odds of suppression per year, this difference was not statistically significant (Table 3). The CLM showed similar results to the BNSV, likely because a linear relationship between dropout time and changes in the log(odds of suppression) fit the data well for both drug users and others (Fig. 5b). The frailty model showed increases in the odds of

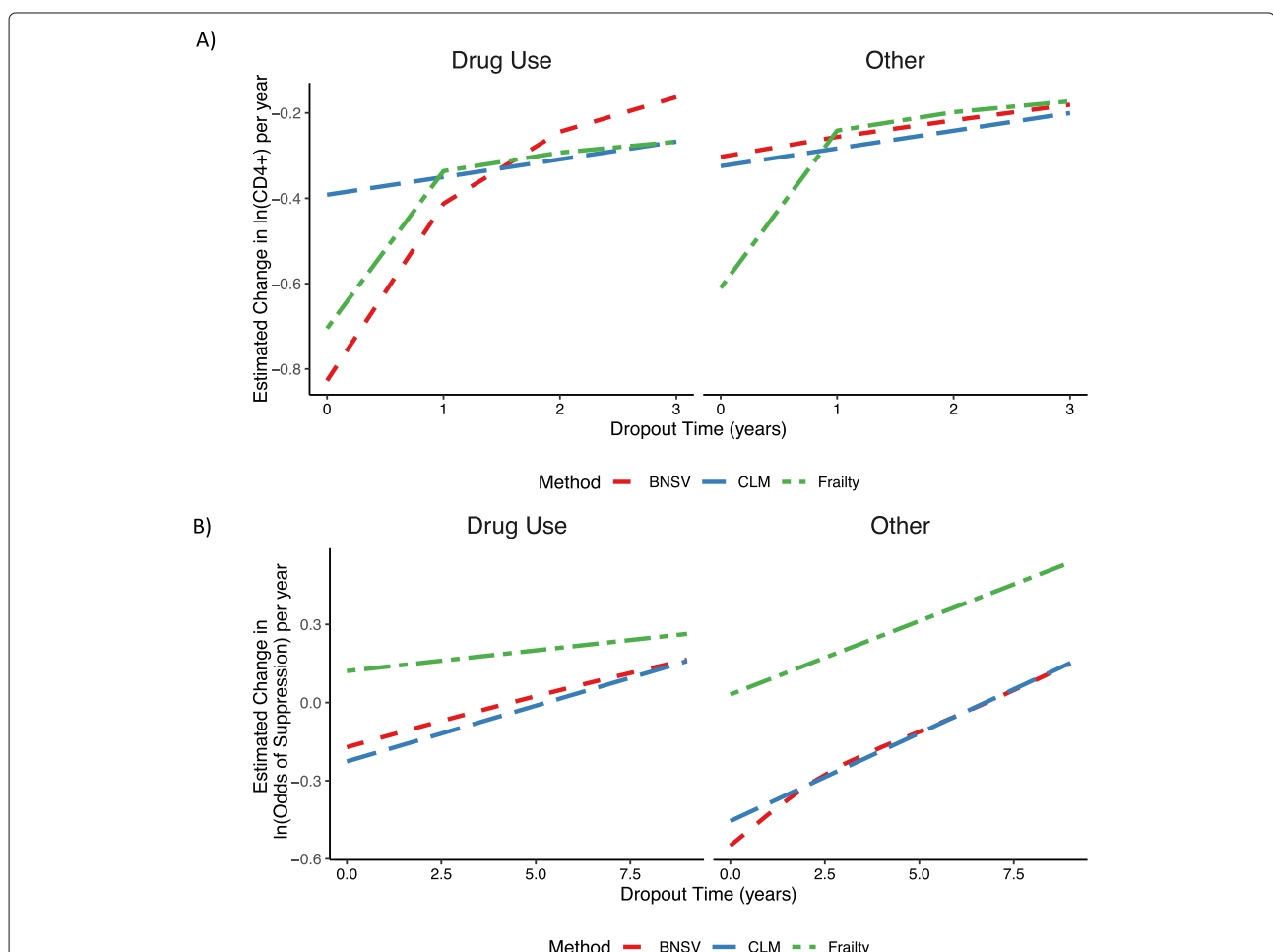


Fig. 5 a) BNSV, CLM and Frailty Model Posterior Mean Estimated Changes in CD4⁺ Count per Year by Dropout Time for Untreated Subjects in the WIHS, Assuming a Baseline CD4⁺ Count of 478.5 **b)** BNSV, CLM and Frailty Model Posterior Mean Estimated Changes in ln(Odds of Suppression) per Year for Subjects that Initiated HAART in the WIHS, Assuming a Baseline CD4⁺ count of 267, Baseline log₁₀(viral load)=4.2, and Random Effects = 0

suppression intermediate between the GLMM and BNSV. Again, this is likely due to lack of fit of the lognormal distribution for dropout times, which in turn influences the estimated dropout time-specific slopes via to covariance parameter, σ_{bu} .

Sensitivity analysis

The results of these analyses rely on the assumption that subjects continue on the same linear trajectory after their dropout. We test the sensitivity of our results to this assumption by considering a proportional attenuation of the slope after a subject's drop out, such that after dropping out, a subject's slope becomes $\delta\beta_1(u_i)$ (Section 3 of the [supplementary material](#)). For CD4⁺ declines, while the estimates of the differences between drug users and others are reduced assuming, $\delta = 0.25, 0.5, 0.75$, hard drug users still have significantly lower CD4⁺ counts at years 1 to 4 than other untreated subjects. For viral load suppression, the odds of suppression remain lower for consistent recreational drug users compared to other subjects that initiated HAART for $\delta = 0, 0.25, 0.5, 0.75$, however, as in the primary analysis, differences between drug users and others were not statistically significant.

Discussion

Potentially non-ignorable dropout is an important consideration when analyzing data from longitudinal clinical trials and cohort studies. While methods that account for non-ignorable dropout must make some unavoidable assumptions that cannot be verified from the observed data [7], many methods make additional parametric assumptions about the distribution of dropout times or the functional form of the relationship between regression coefficients and dropout time. If these assumptions are not met, inferences can be biased, making flexible methods with minimal assumptions important. In our simulation studies, we showed that the BNSV method, which non-parametrically models this distribution of dropout times with a Bayesian bootstrap and flexibly models the relationship between regression coefficients and dropout time with natural cubic B-splines, has reduced bias and mean squared error for the marginal slope and more accurately captures the dropout time varying slope than other methods, such as the CLM and parametric frailty models, which make additional parametric assumptions. These improvements are important since dropout time distributions and relationships between dropout time and changes in outcomes may not always follow simple parametric distributions or polynomial forms in real world analyses.

In our application to the WIHS, we find important differences in estimates of changes in CD4⁺ T cell count over time in untreated subjects that report hard drug use between different models used to account for dropout. We find steeper declines over time using the BNSV model,

which makes fewer assumptions, compared to the CLM and frailty models. Failing to account for dropout or to meet parametric assumptions of models to account for dropout could lead to underestimation of the impact of hard drug use on CD4⁺ T cell count decline in untreated subjects. In our analyses of viral load suppression in subjects that initiated treatment, accounting for dropout using the BNSV showed smaller increases in viral load suppression over time compared to the frailty model and GLMM that did not account for dropout. The relationship between dropout time and the change in log(odds of suppression) was approximately linear, so that the CLM produced similar results to the BNSV. While we did not find significant differences in the odds of suppression between drug users and others in any of our analyses, the probability of suppression was lower when accounting for dropout using the BNSV or CLM. Failing to appropriately account for dropout could lead to over-estimation of the probability of viral load suppression. These low levels of suppression are concerning and require further investigation into methods to help subjects with treatment compliance and affordability of medications.

One drawback of the BNSV method is that the RJMCMC algorithm is computationally intensive; however, we did not find that computational times were prohibitive in either our simulation study or the WIHS data analysis. For the Normal (i) and Binary (i) simulations (400 subjects, 3000-4000 observations), the BNSV took 8.8 and 19.1 minutes, respectively, to complete 40,000 iterations using a MacBook Pro with 3.5 GHz Intel Core i7 processor and 16 GB of RAM. For the WIHS analyses, the analysis of CD4⁺ T cell count (814 subjects, 3,196 observations) took 1.2 hours to complete 200,000 iterations; the analysis of viral load suppression took 6.2 hours to complete 200,000 iterations, due to the larger sample size (1,015 subjects / 15,909 observations) and because Metropolis Hastings steps must be used to estimate the random effects in models with a binary outcome. For comparison, the CLM took 41 minutes and 4.6 hours, and the frailty model took to 22 minutes and 3.3 hours to run the same number of iterations for the CD4⁺ T cell and viral load analyses respectively ([Supplementary Materials](#), Table 5).

Conclusions

We propose a flexible, semi-parametric natural cubic B-spline varying coefficient method to account for dropout in a Bayesian framework. The BNSV extends existing frequentist natural cubic B-spline varying coefficient methods to account for dropout in longitudinal studies with a Gaussian outcome [3, 8] to other non-normal outcomes in the exponential family, while also allowing the number and location of spline knots to be jointly modeled with other model parameters, removing dependence on the

choice of spline knots and more accurately characterizing model uncertainty. The BNSV allows for dropout occurring at any continuous point in time and avoids making parametric assumptions about the distribution of dropout times or the functional form of dropout-varying slope. Results of our simulation studies show that the BNSV reduces bias and mean squared error for the marginal slope compared to parametric frailty models, CLMs and standard GLMMs when non-ignorable dropout is present. The BNSV can also accurately fit models with a linear dropout-varying effect or no dropout-varying effect.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12874-020-01135-3>.

Additional file 1: Supplementary Material.

Abbreviations

BNSV: Bayesian natural cubic B-spline varying coefficient GLMM; CLM: Conditional linear model; GLMM: Generalized linear mixed model; HAART: Highly active antiretroviral therapy; HIV: Human immunodeficiency virus; IG: Inverse gamma; IW: Inverse Wishart; MCMC: Markov chain Monte Carlo; MSE: Mean squared error; RJMCMC: Reversible jump Markov chain Monte Carlo; VCM: Varying coefficient model; WIHS: Women's interagency HIV study

Acknowledgements

The authors would like to thank Nancy Hessol for providing her expertise on the WIHS dataset. Data in this manuscript were collected by the Women's Interagency HIV Study (WIHS). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). WIHS (Principal Investigators): Bronx WIHS (Kathryn Anastos), U01-AI-035004; Brooklyn WIHS (Howard Minkoff and Deborah Gustafson), U01-AI-031834; Chicago WIHS (Mardge Cohen and Audrey French), U01-AI-034993; Metropolitan Washington WIHS (Seble Kassaye), U01-AI-034994; Connie Wofsy Women's HIV Study, Northern California (Ruth Greenblatt, Bradley Aouizerat, and Phyllis Tien), U01-AI-034989; WIHS Data Management and Analysis Center (Stephen Gange and Elizabeth Golub), U01-AI-042590. The WIHS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the National Cancer Institute (NCI), the National Institute on Drug Abuse (NIDA), and the National Institute on Mental Health (NIMH). Targeted supplemental funding for specific projects is also provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute on Alcohol Abuse and Alcoholism (NIAAA), the National Institute on Deafness and other Communication Disorders (NIDCD), and the NIH Office of Research on Women's Health. WIHS data collection is also supported by UL1-TR000004 (UCSF CTSA) and UL1-TR000454 (Atlanta CTSA).

Authors' contributions

CMM developed the BNSV method, performed the simulation studies and WIHS analyses, and wrote the first draft of the manuscript. SM and NEC supervised this work, including BNSV methods development, simulation studies and applications and participated in editing and writing of the manuscript. SEK developed the InformativeDropout R package. All authors have read and approved the manuscript.

Funding

This work was supported by the National Institutes of Health, National Institute on Drug Abuse [DA037778, DA030495]. The National Institute on Drug Abuse did not participate in the study design, analysis, interpretation of data or in writing this manuscript.

Availability of data and materials

A minimal dataset supporting the conclusions of this article and code to implement BNSV models are available as part of the InformativeDropout R package: <https://github.com/kreidles/informativeDropout>. Full data from the WIHS can be obtained from <https://www.niaid.nih.gov/research/wihs-public-dataset>.

Ethics approval and consent to participate

Ethics approval and consent was waived by the Colorado Multiple Institutional Review Board (COMIRB), as this study makes use of only publicly available data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Center for Genes, Environment and Health, National Jewish Health, 1400 Jackson St., 80206 Denver, CO, USA. ²Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 17th Pl., Mail Stop B119, 80045 Aurora, CO, USA. ³Sunrun Inc., 717 17th St., Suite 200, 80202 Denver, CO, USA.

Received: 29 April 2020 Accepted: 25 September 2020

Published online: 07 October 2020

References

- Lanoya E, Mary-Krause M, Tattevin P, Dray-Spirac R, Duvivierd C, Fischere P, Obadiaf Y, Lert F. Predictors identified for losses to follow-up among HIV-seropositive patients. *J Clin Epidemiol*. 2006;59:829–35.
- Little J, Rubin D. *Statistical Analysis with Missing Data*, Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2002, pp. 312–40.
- Moore C, MaWhinney S, Forster J, Carlson N, Allshouse A, Wang X, Routy J-P, Conway B, Connick E. Accounting for dropout reason in longitudinal studies with nonignorable dropout. *Stat Methods Med Res*. 2017;26(4):1854–66.
- Fairclough D, Peterson H, Chang V. Why are missing quality of life data a problem in clinical trials of cancer therapy? *Stat Med*. 1998;17:667–77.
- Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward M, Mallinckrodt C, Carroll R. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004;5:445–64.
- Hogan J, Roy J, Korkontzelou C. Tutorial in biostatistics, handling drop-out in longitudinal studies. *Stat Med*. 2004;23:1455–97.
- Daniels M, Hogan J. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton: Chapman and Hall/CRC; 2008, pp. 85–114 and 165–215.
- Forster J, MaWhinney S, Ball E, Fairclough D. A varying-coefficient method for analyzing longitudinal clinical trials data with nonignorable dropout. *Contemp Clin Trials*. 2012;33:378–85.
- Barkan S, Melnick S, Preston-Martin S, Weber K, Kalish L, Miotti P, Young M, Greenblatt R, Sacks H, Feldman J. The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology*. 1998;9(2):117–25.
- Bacon M, von Wyl V, Alden C, Sharp G, Robison E, Hessol N, Gange S, Barranday Y, Holman S, Weber K, Young M. The Women's Interagency HIV Study: an observational cohort brings clinical sciences to the bench. *Clin Diagn Lab Immunol*. 2005;12(9):1013–9. <https://doi.org/10.1128/CDLI.12.9.1013-1019.2005>.
- Center for Disease Control and Prevention. *Cases of HIV infection and AIDS in the United States in 2002. HIV/AIDS Surveillance Report*, vol. 14. Atlanta: Department of Health and Human Services, Public Health Service; 2002. p. 1–40. Department of Health and Human Services, Public Health Service, Atlanta.
- Loughlin A, Metsch L, Gardner L, Anderson-Mahoney P, Barrigan M, Strathdee S. Provider barriers to prescribing haart to medically-eligible hiv-infected drug users. *Aids Care*. 2004;16(4):485–500.
- Ibrahim J, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test*. 2009;18:1–43.

14. Heckman J. Sample selection bias as a specification error. *Econometrica*. 1979;47:153–61.
15. Diggle P, Kenward MG. Informative drop-out in longitudinal data-analysis. *Appl Stat J R Stat Soc C*. 1994;43:49–93. <https://doi.org/10.2307/2986113>.
16. Heckman J. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Stat Med*. 1998;17:2723–32.
17. Ibrahim J, Chen M, Lipsitz S. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*. 2001;88:551–64.
18. Wu K, Wu L. Generalized linear mixed models with informative dropouts and missing covariates. *Metrika*. 2007;66:1–18.
19. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995;51(1):151–68.
20. Albert P, Follmann D. Modeling repeated count data subject to informative dropout. *Biometrics*. 2000;56:667–77.
21. Schluchter M. Methods for the analysis of informatively censored longitudinal data. *Stat Med*. 1992;11:1861–70.
22. Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV-positive patients. *J Am Stat Assoc*. 1998;93:46–53.
23. Ten Have T, Kunselman A, Pulkstenis E, Landis J. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics*. 1998;54:367–83.
24. Rubin D. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J Am Stat Assoc*. 1977;72:538–43.
25. Wu M, Bailey K. Estimation and comparison of changes in the presence of informative right censoring; Conditional linear model. *Biometrics*. 1989;45:939–55.
26. Wu M, Bailey K. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Stat Med*. 1988;7:337–46.
27. Ekholm A, Skinner C. The Muscatine children's obesity data reanalysed using pattern mixture models. *J Appl Stat*. 1998;47:251–63.
28. Fitzmaurice G, Laird N, Shneyer L. An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Stat Med*. 2001;20:1009–21.
29. Pauler D, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Stat Med*. 2003;22:795–809.
30. Kaciroti N, Schork M, Raghunathan T, Julius S. A Bayesian sensitivity model for intention-to-treat analysis on binary outcomes with dropouts. *Stat Med*. 2009;28:572–85.
31. Kaciroti N, Raghunathan T, Schork M, Clark N. A Bayesian model for longitudinal count data with non-ignorable dropout. *J R Stat Soc Ser C Appl Stat*. 2008;57:521–34.
32. Kaciroti N, Raghunathan T, Taylor J, Julius S. A Bayesian model for time-to-event data with informative censoring. *Biostatistics*. 2012;13:341–54.
33. Hogan J, Lin X, Herman B. Mixtures of varying-coefficient models for longitudinal data with discrete or continuous nonignorable dropout. *Biometrics*. 2004;60:854–64.
34. Su L, Hogan J. Varying-coefficient models for longitudinal processes with continuous-time informative dropout. *Biostatistics*. 2010;11(1):93–110.
35. Biller C. Bayesian Regression Splines in Semiparametric Generalized Linear Models. *J Comput Graph Stat*. 2000;9(1):122–40.
36. Biller C, Fahrmeir L. Bayesian varying-coefficient models using adaptive regression splines. *Stat Model*. 2001;1:195–211.
37. Eubank R. *Nonparametric Regression and Spline Smoothing*. Second Edition. New York: CRC Press; 2002, pp. 277–308.
38. Denison D, Mallick B, Smith A. *J R Stat Soc Ser B Stat Methodol*. 1998;60(2):333–50.
39. Friedman J, Silverman B. Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*. 1989;31(1):3–21.
40. de Boor C. *A Practical Guide to Splines*. Revised Edition. New York: Springer-Verlag; 2001, pp. 87–142.
41. Rubin D. The Bayesian bootstrap. *Ann Statist*. 1981;9:130–4.
42. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995;82:711–32.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

