STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1455–1497 (DOI: 10.1002/sim.1728)

TUTORIAL IN BIOSTATISTICS
Handling drop-out in longitudinal studies

Joseph W. Hogan[1,*,†], Jason Roy[2,‡] and Christina Korkontzelou[1]

[1]*Center for Statistical Sciences, Department of Community Health, Brown University, Providence, RI 02912, U.S.A.*
[2]*Department of Biostatistics and Computational Biology, University of Rochester Medical Centre, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, U.S.A.*

SUMMARY

Drop-out is a prevalent complication in the analysis of data from longitudinal studies, and remains an active area of research for statisticians and other quantitative methodologists. This tutorial is designed to synthesize and illustrate the broad array of techniques that are used to address outcome-related drop-out, with emphasis on regression-based methods. We begin with a review of important assumptions underlying likelihood-based and semi-parametric models, followed by an overview of models and methods used to draw inferences from incomplete longitudinal data. The majority of the tutorial is devoted to detailed analysis of two studies with substantial rates of drop-out, designed to illustrate the use of effective methods that are relatively easy to apply: in the first example, we use both semi-parametric and fully parametric models to analyse repeated binary responses from a clinical trial of smoking cessation interventions; in the second, pattern mixture models are used to analyse longitudinal CD4 counts from an observational cohort study of HIV-infected women. In each example, we describe exploratory analyses, model formulation, estimation methodology and interpretation of results. Analyses of incomplete data requires making unverifiable assumptions, and these are discussed in detail within the context of each application. Relevant SAS code is provided. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: missing data; pattern-mixture model; inverse probability weighting; semiparametric regression; repeated measures; smoking cessation

*Correspondence to: Joseph W. Hogan, Center for Statistical Sciences, Department of Community Health, Brown University, Providence, RI 02912, U.S.A.
†E-mail: jwh@brown.edu
‡The work of this paper was completed while Dr Roy was an Assistant Professor in the Center for Statistical Sciences at Brown University.

Contract/grant sponsor: NIH; contract/grant number: R01-AI50505
Contract/grant sponsor: NIH; contract/grant number: P30-AI42853
Contract/grant sponsor: NIH; contract/grant number: R01-CA77249
Contract/grant sponsor: CDC; contract/grant number: U64-CCU106795
Contract/grant sponsor: CDC; contract/grant number: U64-CCU206798
Contract/grant sponsor: CDC; contract/grant number: U64-CCU306802
Contract/grant sponsor: CDC; contract/grant number: U64-CCU506831

Received June 2002
Accepted September 2003

1. INTRODUCTION

Longitudinal studies and the data they generate occupy a central place in public health and clinical research. When undertaking confirmatory or model-based analyses of longitudinal data, investigators frequently are confronted with having to assess and address potential biases introduced by missing data. Causes for missing data vary and depend in large measure on the type of study being conducted, but include intermittent missed visits by study participants, discontinued participation or other loss to follow-up (possibly initiated by either the participant or the study investigators), lack of effectiveness of a treatment or intervention and mortality. See Reference [1] for a comprehensive listing as it applies to clinical trials.

This paper is concerned specifically with *drop-out* in longitudinal studies, and focuses on regression-based modes of analysis. New developments for handling drop-out have proliferated in the literature; see References [2–9] for reviews. In the interest of maintaining focus, our tutorial does not emphasize distinctions about reason for drop-out; however, several recent papers have addressed this issue, particularly as it relates to treatment discontinuation versus loss to follow-up [10–15].

In the data examples, our tutorial draws from two studies where drop-out is an important issue: the HER Study (HERS) [16], a 7-year follow-up study of HIV-infected women and those at high risk for HIV infection, and the Commit to Quit Study (CTQ) [17, 18], a longitudinal clinical trial designed to assess the effect of vigorous exercise on smoking cessation in women. In each of these, the goal is to draw inferences about one or more covariates on a response variable of interest; drop-out poses a problem if those who drop out are systematically different from those who do not, particularly if these differences cannot be explained by observed response or covariate data. This notion is formalized in Section 3.

Our tutorial has four objectives:

1. To characterize various types of drop-out mechanisms (Sections 3 and 4).
2. To review likelihood-based and moment-based estimation methods in the absence of missing data (Section 3) and in the presence of drop-out (Section 4).
3. To illustrate by example two methods for handling drop-out, pattern mixture models and semi-parametric selection models, and to provide a detailed description of relevant exploratory analyses, underlying assumptions for modelling (and sensitivities to violations of these) and interpretation of model-based inferences (Section 5).
4. To provide software code that will enable readers to implement these methods, with minor modifications.

We confine our attention to regression-based analyses in the presence of drop-out, but certainly non-regression approaches can be used [2, 19]. Methods described herein are also related to non-compliance and causal inference, in the sense that causal inference based on counterfactuals or potential outcomes can be viewed as a missing data problem [20–24]. Although a complete treatment is beyond the scope of the tutorial, ideas from causal inference do play a role in interpretation of treatment effects in the CTQ study.

The remainder of our tutorial is arranged in the following manner: Section 2 provides descriptions of the HERS and CTQ studies; Section 3 introduces relevant notation, distinguishes *full* from *observed* data and briefly reviews regression methods for longitudinal data; Section 4 reviews Little and Rubin's [25] missing data taxonomy, describes its specialization to longitudinal data and provides an overview of methods for adjusting for drop-out in both

the likelihood-based and semi-parametric modelling frameworks; Section 5 demonstrates the application of these methods and Section 6 concludes with a discussion of other important issues in this field of research.

2. TWO LONGITUDINAL STUDIES WITH DROP-OUT

2.1. Commit to quit study (CTQ)

The data for our first example come from CTQ, a randomized controlled clinical trial designed to compare cognitive-behavioural smoking cessation plus supervised vigorous exercise (hereafter referred to as 'exercise') and cognitive-behavioural smoking cessation plus equal staff contact time (hereafter 'control') among healthy but sedentary women [17, 18]. The study enrolled 281 women and scheduled them to participate in 12 weekly sessions of a cognitive behavioural programme tailored for women. Subjects in the exercise condition participated in three sessions of vigorous exercise per week for the 12 weeks of the programme. Those assigned to the control condition participated in three 45–60 min educational sessions per week for the 12 weeks of the programme. As with many smoking cessation trials, the CTQ study has a 4-week 'run-in' period prior to the target quit date, between the fourth and fifth weeks. By week 12, only 93 of 134 (69 per cent) and 96 of 147 (65 per cent) had a 7-day cessation recorded at week 12 in the exercise and contact control arms, respectively. Although net drop-out rates are similar at the 12th week, drop-out in the contact control arm is higher following the target quit date (week 4), whereas drop-out in the exercise arm is higher during the period from weeks 5 to 12.

Drop-out is a notorious problem in smoking cessation studies, and both theoretical and empirical research supports the notion that the act of dropping out is highly correlated both with covariate information (e.g. number and duration of previous cessation attempts) and with response history in the particular study; see Reference [26] for a review and Reference [27] for an analysis of drop-out pattern from the CTQ study. Our analysis in Section 5.1 illustrates the use of likelihood-based and semi-parametric regression to draw inference about treatment effects when drop-out may depend on baseline covariates, longitudinal responses, or both.

2.2. HIV epidemiology research study (HERS)

The HERS is a CDC-sponsored 7-year longitudinal cohort study that enrolled 1310 women in all, 871 of whom were HIV-positive. For full details related to study design and patient recruitment, see Smith *et al.* [16]. Women were scheduled for follow-up every 6 months, at which time a variety of clinical, behavioural and demographic information was collected. One substudy of the HERS is concerned with studying the role of baseline patient characteristics on variations in longitudinal CD4 cell count among women infected with HIV (for a more detailed account, see Reference [28]). CD4 cell count is an important marker of disease progression because HIV directly attacks this lymphocyte; it also is a useful surrogate marker of treatment effects [29].

Attrition rate in the HERS, like many other long-term follow-up studies, is appreciable; Table I shows, for the HERS, the number and proportion observed and dropped out at each study visit following baseline. Moreover, it is quite plausible that drop-out is closely related to CD4 counts that are missing because of drop-out; i.e. that the unobserved CD4 counts among those who drop out are systematically lower than those who continue follow-up, even after adjusting for observed covariates and CD4 counts. Under some specific but largely

Table I. Patterns of observing CD4 in the HER Study, by visit. Missing data at each visit is a combination of intermittent missingness and drop-out.

Visit	In follow-up (proportion)	CD4 observed (proportion)	Dropped out
1	871 (1.00)	850 (0.98)	56
2	815 (0.94)	706 (0.81)	28
3	787 (0.90)	692 (0.79)	39
4	748 (0.86)	665 (0.76)	46
5	702 (0.81)	617 (0.71)	34
6	668 (0.77)	587 (0.67)	32
7	636 (0.73)	576 (0.66)	22
8	614 (0.70)	547 (0.63)	19
9	595 (0.68)	522 (0.60)	24
10	571 (0.66)	506 (0.58)	18
11	553 (0.63)	492 (0.56)	104
12	449 (0.52)	405 (0.46)	133

untestable assumptions, pattern mixture models can be used to model dependence between missing responses and drop-out. Our analysis in Section 5.2 demonstrates the use of pattern mixture modelling to estimate the effects of the following covariates on mean CD4 and on change in CD4 over time: HIV-RNA (viral load), HIV symptom severity, antiviral treatment status and number of years aware of HIV infection.

3. FULL-DATA REGRESSION MODELS FOR LONGITUDINAL RESPONSES

Since our focus is on regression-based approaches, a review of regression concepts and notation is needed. This section reviews basic regression for settings where data are fully observed. Many of the models for incomplete data, described in Section 4, are extensions or generalizations of the regression models described here. We assume throughout that parameters indexing the full-data distribution are of primary interest; however, this may not always be the case in practice. For example, certain subpopulations such as treatment compliers or those surviving to a specific time point may be of primary interest.

We assume that the full data consist of T serial observations taken on n individuals at a fixed set of time points $t_1,\ldots,t_T$. The number and timing of measurements for the full data is assumed to be equal for all subjects, although in practice this can typically be relaxed without loss of generality for the methods we describe herein. Hence, for individual i, where $i=1,\ldots,n$, *full* data consists of the vector $Y_i=(Y_{i1},\ldots,Y_{iT})^{\mathrm{T}}$.

The vector Y_i also can be viewed as the realization of a stochastic process $\{Y_i(t)\}$ at fixed time points $t_1,\ldots,t_T$, and is sometimes modelled this way [30, 31]. More generally, the measurement times may be unequally spaced or even random variables themselves, possibly depending on the response process itself [30]. The full-data regression techniques discussed here apply in principle whenever the measurement times are independent of $\{Y_i(t)\}$, but for handling drop-out, irregular measurement times may introduce non-trivial complications. These are discussed in Section 4.

In addition to the full Y data, denote the full collection of covariates as $V_i=(V_{i1}^{\mathrm{T}},\ldots,V_{iT}^{\mathrm{T}})^{\mathrm{T}}$, where each V_{it} is a $1\times S$ vector of covariates linked to Y_{it}. We further partition V_i as

$V_i = (X_i, W_i)$ to distinguish covariates whose association with Y_i is of direct interest (X_i) from those of an auxiliary nature (W_i), because even when scientific interest is in $f(y \mid x)$, information on W may be useful in accounting for the effects of missing data. We assume that X_i and W_i are always fully observed on all individuals. See References [32–36] for methods to handle missing covariates in regression settings.

Turning now to regression models, the primary objective is to estimate $E(Y_{it} \mid X_i)$, the mean of Y given X. As before, let $Y_i = (Y_{i1}, \ldots, Y_{iT})^{\mathrm{T}}$. Define the associated $T \times P$ matrix of covariates $X_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{iT}^{\mathrm{T}})^{\mathrm{T}}$, where $X_{it} = (X_{it1}, \ldots, X_{itP})$ is the $1 \times P$ covariate vector associated with Y_{it}. Finally, define the $T \times 1$ mean vector $\mu_i = (\mu_{i1}, \ldots, \mu_{iT})^{\mathrm{T}} = E(Y_i \mid X_i)$. The covariates X_{it} are assumed to be exogenous in the sense that they meet at least one of the following three conditions [37]: (i) they are fixed by design (such as planned measurement times), (ii) they are deterministic functions of baseline covariates or (iii) they are *external* to the measurement process in the sense that response Y_{it} is not predictive of covariates $X_{i,t+1}$. An example of a covariate that violates (i) is a measurement time dictated by underlying condition, as when sicker individuals may have more doctor visits and hence more frequently recorded outcomes; condition (iii) would be violated by a covariate such as time-varying therapy used to treat a condition measured by Y. In an HIV study, for example, if X_{it} is a binary indicator of receiving antiviral therapy and Y_{it} is the average number of HIV-1 RNA copies per ml of plasma, then X_{it} may violate (iii) if it is preferentially given to those with greater viral burden. Violations of the exogeneity condition can arise from multiple sources, including measurement error and confounding; see References [38–41] for a detailed discussion.

3.1. Marginal models

Marginal regression models are specified in terms of $\mu_i = E(Y_i \mid X_i)$, which measures marginal covariate effects. Semi-parametric methods that exploit connections to generalized linear models have become popular for fitting marginal models because they avoid having to specify a multivariate joint distribution and because both continuous and discrete responses can be handled in a unified framework. A widely used estimation procedure for semi-parametric inference from full data is generalized estimating equations, or GEE [42, 43], which requires specification of only the first two moments of Y as opposed to its entire distribution. The mean of Y is assumed to be related to covariates through a smooth, monotone link function $g : \mathbb{R}^T \to \mathbb{R}^T$ that is linear in covariates, i.e.

$$g(\mu_i) = (g(\mu_{i1}), \ldots, g(\mu_{iT}))^{\mathrm{T}} = X_i\beta$$

where β is a $P \times 1$ vector of regression coefficients. A consistent and semi-parametric estimator of β is the solution to the $P \times 1$ system of estimating equations

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} \Sigma_i^{-1} (Y_i - \mu_i) = 0 \tag{1}$$

where $D_i = D_i(\beta) = (\partial \mu_i / \partial \beta)^{\mathrm{T}}$ is a $T \times P$ matrix of derivatives, $\mu_i = \mu_i(\beta) = g^{-1}(X_i\beta)$ is the $T \times 1$ mean vector, and $\Sigma_i = \mathrm{var}(Y_i \mid X_i)$ is symmetric, positive definite $T \times T$ matrix of variances and covariances. The variance matrix Σ_i can be decomposed as $\Sigma_i(\phi, \alpha, \beta) = \phi A_i(\beta)^{1/2} C_i(\alpha) A_i(\beta)^{1/2}$, where $A_i(\beta)$ is a $T \times T$ diagonal matrix with elements $a(\mu_{i1}), \ldots, a(\mu_{iT})$ (for some variance function $a(\cdot)$), $C_i = C_i(\alpha)$ is a $T \times T$ working correlation matrix, and ϕ is a scale parameter. Liang and Zeger [43] show that when data are fully observed, the solution

to (1) is consistent and asymptotically normal even when $C_i(\alpha)$ is incorrectly specified, so long as the model for the mean is correct. For correctly specified Σ_i, (1) coincides with the likelihood score equation for correlated multivariate normal data when identity link is used (because $D_i^{\mathrm{T}} = X_i^{\mathrm{T}}$).

It is also possible to estimate marginal regression models based on the likelihood of the joint distribution $f(y_{i1}, \ldots, y_{iT} \mid x)$. A fully likelihood-based inference for correlated data may be desired if association parameters are of direct interest, or in the case where data are missing. When data are missing, proper specification of the likelihood—including covariances—takes on added importance. Outside of the multivariate normal distribution, modelling association structures can be unwieldy. For example, if responses are binary, one can potentially specify all associations up to order T (there are $\sum_{j=2}^{T} \binom{T}{j}$ in all), although in many cases the higher-order associations can reasonably be assumed to be zero [44]. See Reference [45] for a comprehensive review. In our first example (Section 5.1), we use a model for the joint distribution of repeated binary outcomes, where associations are modelled in terms of longitudinal transition probabilities [46].

3.2. *Random effects models*

A likelihood-based alternative to specifying the full marginal distribution of $Y_{i1}, \ldots, Y_{iT}$ is to structure correlation using individual-specific random effects. The usual approach is to specify a regression model of responses conditionally on the random effects and assume that within subject, the $Y_{i1}, \ldots, Y_{iT}$ are independent given random effects. Integrating out the random effects yields marginal correlation between the Y_{it} within subject. See References [47, 48] for normal-error formulations, and Reference [49] for a more general review in the context of longitudinal data.

At the first level of the model, the Y_{it}'s are assumed to be independent, conditional on subject-specific random effects η_i. Typically the conditional mean will take the form

$$g\{E(Y_i \mid \eta_i)\} = X_i\beta + Z_i\eta_i$$

where $g : \mathbb{R}^T \to \mathbb{R}^T$ is the (vector-valued) link function and Z_i is a design matrix for the subject-specific random effects. This representation of the conditional mean motivates the term 'mixed-effects model' because the coefficients quantify both population-level (β) and individual-level (η_i) effects. At the second level, the η_i follow some distribution, typically multivariate normal. The marginal joint distribution is obtained by integrating over η_i,

$$f(y_{i1}, \ldots, y_{iT}) = \int f(y_{i1}, \ldots, y_{iT} \mid \eta_i)\, \mathrm{d}F(\eta_i)$$

Note that the structure at the first level follows a generalized linear model conditional on η. For example, when the response is normally distributed we may assume a normal error linear regression model for the conditional distribution of Y given η. For binary responses we may assume that a logistic regression model applies, conditional on η. Marginally, correlation is induced through the shared random effects η. The marginal mean of Y_i is obtained by integrating the conditional mean over the distribution of the random effects, i.e.

$$\mu_i = \int g^{-1}(X_i\beta + Z_i\eta_i)\, \mathrm{d}F(\eta_i)$$

The parameters β have marginal interpretations only for certain link functions (e.g. identity link); however in general, the marginal mean $E(Y_i \mid X_i)$ will no longer be related linearly to X_i through g.

4. CHARACTERIZING DROP-OUT IN LONGITUDINAL STUDIES

Full-data models fit to incomplete data are inherently non-identifiable. Specifically, a full-data distribution $f_\theta(y \mid x)$ cannot be identified non-parametrically from an incomplete sample of the Y's; strategies for modelling and estimation therefore rest on a number of assumptions and restrictions. These can take a number of forms; for example, *distributional* or *parametric* assumptions involve assigning a specific model to the full-data distribution $f_\theta(y \mid x)$, such as multivariate normal. One can also impose *structural* assumptions on the full-data distribution. In longitudinal data with drop-out, this can be used as a basis for extrapolating missing data from observed data; for example, if $x_{it} = (1, t)$, then $E\{Y_{it} \mid x_{it}\} = \alpha_0 + \alpha_1 t$ is a structural assumption because it specifies the mean of both observed and missing responses.

Our focus in this section is on the use of assumptions about the missing data process and its relation to covariates and to observed and unobserved responses in the full data. Specifically, we apply Rubin's [50] MCAR-MAR taxonomy to the case of longitudinal data with drop-out (see also References [3, 8]). For longitudinal data, the formulation, interpretation and implications of these assumptions are not always immediately obvious. In what follows, we begin by stating our inferential objective in fairly general terms, then list specific assumptions under which likelihood-based or moment-based inference can proceed. We do not attempt to be comprehensive, but rather focus on those assumptions that tend to be qualitatively evaluable in practical settings and that lead to analyses which are relatively easy to implement. More comprehensive and technical accounts can be found in Chapters 8–10 of Little and Rubin [25], who give an overview for continuous and discrete data problems; Little [3], who classifies modelling strategies for handling drop-out; Diggle and Kenward [8], who apply the MCAR-MAR taxonomy to parametric models for longitudinal data; and Robins *et al.* [37], who focus on semi-parametric analyses.

Recall that the full-data response vector is $Y_i = (Y_{i1}, \ldots, Y_{iT})^{\mathrm{T}}$. To discuss missing data generally, we introduce indicator variables $R_i = (R_{i1}, \ldots, R_{iT})^{\mathrm{T}}$, where $R_{it} = 1$ if Y_{it} is observed and $= 0$ otherwise. Associated with Y_{it} is a $T \times P$ matrix of covariates $X_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{iT}^{\mathrm{T}})^{\mathrm{T}}$, and interest is in estimating parameters from the full-data distribution $f_\theta(y \mid x)$; in a regression context, the objective is to model $E(Y_{it} \mid X_i)$. Finally, there may exist auxiliary measured variables $W_{it} = (W_{it1}, \ldots, W_{itQ})^{\mathrm{T}}$ that are not included in the regression model but can be used to gain information about the missing data mechanism, missing response values or both. Although are framing the problem in terms of discrete indicators for missing data, many of the ideas apply more generally to continuous-time processes, e.g. $Y_i(t)$, $R_i(t)$ and $W_i(t)$. However, the technical aspects of model fitting may be significantly more complicated. See Reference [51] for a comprehensive overview.

4.1. Missing completely at random

Working from Rubin [50] and from Little and Rubin [25, pp. 14–17], we distinguish the following types of missing data mechanisms. Partition Y_i into its observed and missing components, such that $Y_i = (Y_{i,\mathrm{obs}}^{\mathrm{T}}, Y_{i,\mathrm{mis}}^{\mathrm{T}})^{\mathrm{T}}$. Throughout, we assume X_i is fully observed, and

that the individuals with one or more measurements constitutes a random sample from the population of interest (i.e. there are no Y vectors that are completely unobserved). Recall that the full set of covariates V_i consists of model covariates X_i and auxiliary variables W_i, such that $V_i = (X_i, W_i)$.

Definition MCAR (*missing completely at random*): Unobserved components of Y are MCAR if the probability of non-response depends neither on V nor on Y; i.e. if $R \perp\!\!\!\perp (Y, V)$, or equivalently $f(r \mid v, y) = f(r)$.

An example of MCAR is when the number of follow-up visits differs by individual due to staggered entry and administrative censoring at a fixed calendar time.

A slightly weaker assumption allows missingness to depend on model covariates X, but requires that missingness is independent of Y, given X. Diggle and Kenward [8] consider this to be a version of MCAR, while Little [3] to classify it separately as covariate-dependent missingness. Since the missing data mechanism is allowed to depend on observed data, Little [3] refers to this as covariate-dependent missingness. When inference is based on a model, the key distinction is that if missingness depends on X, it may depend on a specific functional form of X and hence the model must be correctly specified. We adopt the term 'covariate-dependent MCAR' and describe it here for completeness; however, because our focus is on regression modelling, we only distinguish between MCAR and covariate-dependent MCAR when it is necessary for clarity.

Definition Covariate-dependent MCAR: An MCAR missingness process is is covariate-dependent if the probability of non-response depends only on model covariates X, and conditionally on X, does not depend on Y or W; i.e. $R \perp\!\!\!\perp (Y, W) \mid X$, or equivalently $f(r \mid v, y) = f(r \mid x)$.

Covariate-dependent MCAR states that within levels of X, observed elements $Y_{i,\text{obs}}$ constitute a random sample from the full-data vector Y_i. It is an appropriate assumption, for example, for analysing data from a multicentre study where centre indicator is a model covariate, drop-out rate differs by centre and drop-out within centre does not further depend on excluded covariates, observed responses or missing responses.

4.2. *Missing at random (MAR)*

For longitudinal data, it is uncommon for drop-out and missingness to be independent of responses. Conditional on model covariates X, MAR allows missingness to depend on observable Y values, auxiliary covariates W or both. The type of MAR being assumed may have bearing on the choice of analysis because some analyses may require further assumptions. For example, if missingness depends on auxiliary variables, then likelihood-based methods may require the analyst to model or make assumptions about their marginal joint distribution. In this subsection, we describe versions of MAR that allow dependence on Y_{obs} and W, and differentiate a sequential formulation of MAR that is used for semi-parametric inference.

Definition Missing at random (*MAR*): Unobserved components of Y are MAR if the probability of non-response may depend on X, W and Y_{obs}, but conditionally on these is independent of Y_{mis}. Specifically, $R \perp\!\!\!\perp Y_{\text{mis}} \mid (X, W, Y_{\text{obs}})$, or equivalently $f(r \mid x, w, y) = f(r \mid x, w, y_{\text{obs}})$.

Missingness will be MAR in a longitudinal study if, for example, among participants with the same covariate profile, those who are *observed* to be sicker (via their values of Y_{obs}) are more likely to have missing values, so long as their missingness probability does not further depend on their missing responses.

The version of MAR that we have stated here is rather general because we allow missingness to depend on auxiliary covariates W. For likelihood-based inference this may present complications if W is associated with responses, conditionally on model covariates X; in particular incorporation of information on W may require specification of $f(w)$ and subsequent integration over w. See References [52, 53] for examples, and Section 4.4.3 for further discussion. By contrast, MAR mechanisms that depend on auxiliary covariates can be handled using multiple imputation methods [54–57] and semi-parametric methods that employ inverse weighting [37, 58] because the distribution of auxiliary covariates does not need to be modelled.

4.3. *Sequential MAR*

Following Robins *et al.* [37], we describe a sequential version of MAR that is used to justify semi-parametric inference using inverse probability weighted (IPW) estimators. We describe its use for monotone missing data patterns, although it can also be used in more general settings [58].

Under sequential MAR, drop-out at t may depend on observable data up to t, including observed responses, model covariates and auxiliary covariates. It is helpful to define variable histories using script notation as follows:

$$\mathscr{X}_{it} = \{X_{i1}, \ldots, X_{it}\}, \quad \mathscr{Y}_{it} = \{Y_{i1}, \ldots, Y_{it}\}, \quad \mathscr{W}_{it} = \{W_{i1}, \ldots, W_{it}\}$$

Let $\mathscr{F}_{it} = \{\mathscr{X}_{it}, \mathscr{W}_{it}, \mathscr{Y}_{i,t-1}\}$ denote covariate history up to and including time t and response history up to but *not* including time t. A monotone pattern of missingness is assumed such that $R_{i1} = 1$ and for $t = 2, \ldots, T$, $R_{it} = 1 \Rightarrow R_{i,t-1} = 1$. Sequential MAR is defined as follows [37, 59].

Definition S-MAR (*Sequential missingness at random*):[‡] Conditionally on past history $\mathscr{F}_{it}$ and the full-data response vector Y_i, drop-out at time t does not depend on current or future response data $Y_{it}, \ldots, Y_{iT}$; formally,

$$\text{pr}(R_{it} = 0 \mid R_{i,t-1} = 1, \mathscr{F}_{it}, Y_i) = \text{pr}(R_{it} = 0 \mid R_{i,t-1} = 1, \mathscr{F}_{it})$$

S-MAR is analogous to the assumption of 'non-informative censoring' in survival analysis (cf. Reference [60, p. 100]). Non-informative censoring is the condition under which standard counting process statistics retain their martingale properties (e.g. logrank test, partial likelihood score statistics, etc.), and holds if the hazard of an uncensored failure time S is equivalent its crude hazard under censoring; i.e. if

$$\text{pr}(t \leqslant S \leqslant t + \text{d}t \mid S \geqslant t) = \text{pr}(t \leqslant S \leqslant t + \text{d}t \mid S \geqslant t, C \geqslant t)$$

where C is a censoring time. This property essentially says that, given $S \geqslant t$ and $C \geqslant t$, S and C are independent given the past. The S-MAR assumption states that given $R_{i,t-1} = 1$, the future but potentially incomplete responses $(Y_{it}, \ldots, Y_{iT})^{\text{T}}$ (analogous to S) are independent of R_{it} (analogous to C), given the past.

[‡]This is Assumption 2a in Robins *et al.* [37].

Even under monotone missingness, MAR and S-MAR are different assumptions. Under MAR, drop-out at t (R_{it}) can depend on elements of X, W and $Y_{i,\text{obs}}$ observed *before*, *at*, and *after* t. Under S-MAR, drop-out at time t can depend only on those elements of X, W and $Y_{i,\text{obs}}$ observed *before* t (or at t, for X and W).

4.4. Modes of inference under MCAR and MAR

4.4.1. Likelihood-based inference. Define $f_\psi(r \mid y,x)$ to be the distribution of missing data indicators. Assuming a correctly specified parametric full-data model $f_\theta(y \mid x)$, likelihood-based inference for full-data parameter θ can be based on the likelihood of the observed data. This requires both the MAR assumption and a *separable parameters* assumption, meaning the parameter spaces of θ and ψ are non-overlapping. The combination of MAR and separable parameters assumptions constitutes the *ignorability* condition, which derives its name from the fact that the missing data model can be left unspecified, or *ignored.* Specifically, if (i) one is willing to assume a model for $f_\theta(y \mid x)$, (ii) θ and ψ are separable and (iii) MAR holds conditionally on X, then the joint distribution of Y_{obs} and R—and hence the observed-data likelihood contribution for an individual—factors over θ and ψ as follows:

$$\begin{aligned} L_i^{\text{O}}(\theta,\psi) &\propto f_{\theta,\psi}(Y_{\text{obs},i}, R_i, X_i) \\ &= \int f_\theta(Y_{\text{obs},i}, Y_{\text{mis},i} \mid X_i) f_\psi(R_i \mid Y_{\text{obs},i}, Y_{\text{mis},i}, X_i)\, \mathrm{d}Y_{\text{mis},i} \\ &= f_\psi(R_i \mid Y_{\text{obs},i}, X_i) \int f_\theta(Y_{\text{obs},i}, Y_{\text{mis},i} \mid X_i)\, \mathrm{d}Y_{\text{mis},i} \\ &= f_\psi(R_i \mid Y_{\text{obs},i}, X_i) f_\theta(Y_{\text{obs},i} \mid X_i) \end{aligned}$$

Recall that θ indexes the full-data model $f_\theta(y \mid x)$. The implication of ignorability is that likelihood-based inference about θ can be based on any function proportional in θ to $\prod_{i=1}^n f_\theta(Y_{\text{obs},i} \mid X_i)$. Reference [25] provides full details. Although this is an appealing consequence of MAR, the form of the observed-data likelihood may not possess the simplicity of the full-data likelihood and hence optimization may require specialized algorithms such as EM [61].

4.4.2. Semi-parametric inference. Under MCAR or MAR, unbiased moment-based estimating equations that use only observed response data are easily constructed. Let $\Delta_i = \text{diag}(R_{i1}, \ldots, R_{iT})$ be the $T \times T$ diagonal matrix with zeroes on the off-diagonal, and let $K(X_i, \beta)$ denote a $p \times T$ matrix of weights that can be any function of X_i and β. Typically (e.g. for GEE), this is chosen to be $K_i = K(X_i, \beta) = D_i(X_i, \beta)\, \Sigma(X_i, \beta)^{-1}$. Then the estimating equations for β, based only on observed data, are

$$U_1(\beta) = \sum_{i=1}^n K(X_i, \beta) \Delta_i \{Y_i - \mu_i(X_i, \beta)\} = 0 \tag{2}$$

MCAR implies that $\Delta_i \perp\!\!\!\perp Y_i \mid X_i$, so the expectation of each summand is zero, the estimating equations are unbiased at the true value of β, and their solution $\hat{\beta}$ is consistent for the true β.

See Reference [37, Section 3] for general discussion, and Reference [43, Theorem 2] for details on properties of $\hat{\beta}$ when the weight function K_i is allowed to depend on covariance parameters.

If missingness depends either on Y_{obs} or W, it is no longer true that the estimating equations (2) will yield a consistent estimator of β because of the dependence between Δ_i and Y_i. If S-MAR holds, then consistent estimates of β can be obtained by solving a properly weighted version of (2). The following two assumptions must hold in addition to S-MAR:

Assumption 1 (*Non-zero probability of remaining in study*)
Given past history $\mathscr{F}_{it}$, the probability of being observed at t is bounded away from zero; i.e. $\text{pr}(R_{it}=1 \mid R_{i,t-1}=1, \mathscr{F}_{it}) > \xi > 0$.

Assumption 2 (*Correct specification of drop-out model*)
The functional form of the hazard of drop-out at t is known up to a vector α of parameters; i.e. $\text{pr}(R_{it}=0 \mid R_{i,t-1}=1, \mathscr{F}_{it}) = \lambda_{it}(\alpha)$, where λ is a known function and α is an unknown finite-dimensional parameter.

Under monotone missingness, it follows immediately that the marginal response probabilities are

$$\pi_{it}(\alpha) = \text{pr}(R_{it}=1 \mid \mathscr{F}_{it}) = \prod_{j=1}^{t} \{1 - \lambda_{it}(\alpha)\}$$

Re-define $\Delta_i(\alpha) = \text{diag}\{R_{i1}/\pi_{i1}(\alpha), \ldots, R_{iT}/\pi_{iT}(\alpha)\}$. If Assumptions 1 and 2 hold, and if drop-out occurs according to S-MAR, then subject to some regularity conditions, the solution $\hat{\beta}$ to the weighted estimating equation

$$U_2(\beta, \hat{\alpha}) = \sum_{i=1}^{n} K(X_i, \beta) \Delta_i(\hat{\alpha}) \{Y_i - g^{-1}(X_i, \beta)\} = 0 \tag{3}$$

is consistent for β. In (3), $\hat{\alpha}$ is a consistent estimator of α under a correctly specified model $\lambda_{it}(\alpha)$ [37, Theorem 1].

4.4.3. Likelihood versus semi-parametric inference under MAR. There are several comparisons worth making between semi-parametric estimation from (3) under S-MAR and likelihood-based estimation under MAR, at least the way we have described them here. First, the likelihood-based methods tend to treat longitudinal data as clustered data that happen to be temporally aligned, assuming that $Y_{\text{mis}} \perp\!\!\!\perp R \mid (Y_{\text{obs}}, X)$, regardless of where drop-out occurs, whereas with semi-parametric inference from weighted estimating equations, the S-MAR assumption conditions only on elements of Y_{obs} realized prior to a fixed time. Second, the semi-parametric approach uses auxiliary covariates W_i without having to make any assumptions about their distribution or their relationship to either Y_i or X_i. In principle, likelihood-based methods can make use of auxiliary variables to help explain the missing data process, but in general this requires the analyst to specify the joint distribution $f(y, r, w \mid x)$ and then integrate over w. For example, if the joint distribution is factored as

$$f_{\psi,\theta,\gamma}(y, r, w \mid x) = f_{\psi}(r \mid y, w, x) f_{\theta}(y \mid w, x) f_{\gamma}(w \mid x)$$

then even under ignorability, specification of $f_\theta(y \mid w,x)$ and $f_\gamma(w \mid x)$ is needed. An exception, mentioned in Section 4.2, is when W is independent of Y, given X, which implies $f_\theta(y \mid w,x) = f_\theta(y \mid x)$. In simple cases, where w has a small number of support points (e.g. a single binary covariate), or where the distributional form of $f(y,r,w \mid x)$ is known with high confidence, use of an auxiliary variable can be straightforward. Otherwise it can involve extra modelling assumptions and non-trivial computation. The third and final point is that under MAR, the likelihood-based approach requires correct specification of the full-data model $f_\theta(y \mid x)$, but no further modelling of the missing data mechanism is necessary. The semi-parametric approach requires correct specification of the full-data mean function (e.g. $g\{E(Y_i \mid X_i)\} = X_i\beta$), but also requires that the hazard of drop-out $\lambda_{it}(\alpha)$ be correctly specified. Hence, the need to specify $\lambda_{it}(\alpha)$ is the price the analyst must pay in order to avoid making distributional assumptions about $f_\theta(y \mid x)$. We note here that an active area of research concerns development of semi-parametric estimators that are 'doubly robust' in the sense that, under some structural assumptions about the joint distribution of Y and R, consistent estimates of β can be had if *either* $E(Y_i \mid X_i)$ or $\lambda_{it}(\alpha)$ is correctly specified (see Reference [62, Section 1.2.6]).

4.5. *Missingness not at random*

Missingness and drop-out are not at random if the MAR or S-MAR assumptions are violated. The majority of analytic approaches under MNAR are based on models for the joint distribution of the response and the drop-out mechanism, and can be classified according to how the joint distribution $f(Y,R)$ is factored. We review selection models, mixture models and frailty models, illustrating with examples to highlight model interpretation and key assumptions being made for identification.

These models and methods can be viewed in the broader context of jointly modelling repeated measures and an event time, and in fact some of the models for dealing with drop-out have been used to model jointly evolving processes such as CD4 trajectory and HIV-related death [51, 53, 63–65] and multivariate correlated processes [31, 66].

4.5.1. Likelihood-based methods. Likelihood-based methods are common for handling drop-out that is not MAR. (This is sometimes called *informative drop-out*, although the term 'informative' is imprecise. The term 'non-ignorable' is more precise because it refers to drop-out processes that violate the ignorability condition defined in Section 4.4.1.). Our brief survey illustrates likelihood-based approaches by describing three modelling strategies: selection modelling, where the joint distribution is factored as the product of the full-data model and a selection model; mixture modelling, where the full data is modelled as a mixture over drop-out times or patterns; and frailty models, where a latent frailty term captures dependence between drop-out and the response process. The respective factorizations are listed here:

$$f(y,r \mid x) = f(y \mid x) f(r \mid y,x) \tag{4}$$

$$f(y,r \mid x) = f(y \mid r,x) f(r \mid x) \tag{5}$$

$$f(y,r \mid x) = \int f(y \mid \eta,x) f(r \mid \eta,x)\, \mathrm{d}F(\eta \mid x) \tag{6}$$

Selection models [67] require the user to specify a model $f_\theta(y_1,\ldots,y_T)$ for the full data and a selection model that characterizes the drop-out probability as a function of covariates and the full data $y_1,\ldots,y_T$. For longitudinal data with drop-out, it is convenient to specify the selection model in terms of the hazard function associated with drop-out

$$\lambda_{it} = \mathrm{pr}(R_{it} = 0 \mid R_{i,t-1} = 1, Y_{i1},\ldots,Y_{iT})$$

A prototypical example is the model used by Diggle and Kenward [8], wherein the full-data model follows a multivariate normal regression:

$$Y_{it} = X_{it}\beta + e_{it}, \quad t = 1,\ldots,T$$

with $e_i = (e_{i1},\ldots,e_{iT})^{\mathrm{T}} \sim \mathrm{N}(0,\Sigma)$. The companion selection model allows drop-out to depend on covariates and on both observed and missing responses; for example,

$$\mathrm{logit}(\lambda_{it}) = \alpha_t + X_{it}\gamma + Y_{i,t-1}\psi_1 + Y_{it}\psi_2$$

An appealing feature of this model is that it generalizes to longitudinal data, in a natural way, the Little and Rubin MCAR-MAR taxonomy; i.e. Y_{it} is missing when drop-out occurs at t, so drop-out is MAR when $\psi_2 = 0$ and MNAR otherwise.

The assumption of multivariate normality for $f(y)$ is sufficient to identify ψ_2, which can be viewed as either a blessing or a curse. When subject matter justification exists for assuming normality, then it is sensible to use this knowledge to inform an otherwise unidentifiable parameter [41]; however, these models must be applied with caution and if possible subject to sensitivity analysis [68], because it is impossible to critique the assumption of normality when responses are missing.

Identification of parameters relies on two key features of this model: normality for the response distribution and linear dependence between $\mathrm{logit}(\lambda_{it})$ and (the possibly missing) Y_{it} in the selection model. It is not possible to distinguish between violations of these two critical assumptions, which means parametric selection models need to be applied with caution. Kenward [68] provides some practical strategies for examining sensitivity to normality. Other authors have looked at local departures from MAR under the assumption that both the response and selection models are correctly specified [69].

Selection models in general require specialized numerical routines for maximising the likelihood, limiting practical utility for broad ranges of problems. Moreover, in spite of parametric assumptions, it may still be the case that the likelihood is very flat with respect to parameters like ψ_2 that characterize outcome-dependent non-MAR selection, leading to numerical instabilities. See References [8, 70] for discussion of estimation issues related to normal full-data models; see References [71–74] for discussion of selection models with binary data. For settings with multivariate repeated measures, see Reference [36]. The models described above can be fit using OSWALD software [75] and in some circumstances using BUGS [76].

Mixture models treat the full-data distribution as a mixture over drop-out times or patterns, and in that sense regard drop-out as a source of heterogeneity. Early examples of mixture models for drop-out and missing data include Rubin [77] and Wu and Bailey [78, 79]. Little coined the term 'pattern mixture model' for multivariate data where missingness can be categorized into distinct patterns. Applied to longitudinal studies, pattern-mixture models are well suited to analyses where the number of drop-out times is small [80–83].

As an illustrative example, we refer to an application of pattern mixture models to longitudinal data with drop-out by Little and Wang [82], who use the models to analyse longitudinal

data from a depression study. The full data consist of $Y_i = (Y_{i1}, \ldots, Y_{iT})^{\mathrm{T}}$, with drop-out leading to missingness at time T. Hence, the drop-out variable R_i is Bernoulli with parameter $\psi = \mathrm{pr}(R_i = 1)$. The mixture model is specified in two parts; the first gives the conditional distribution of Y given R, and the second specifies the marginal distribution of R. For this example, the conditional distribution $f(y \mid r, x)$ follows:

$$(Y_i \mid R_i = r) \sim \mathrm{N}(\mu^{(r)}, \Sigma^{(r)})$$

where for pattern $r \in \{0, 1\}$, $\mu^{(r)}$ is a $T \times 1$ mean vector and $\Sigma^{(r)}$ is the $T \times T$ error variance matrix. When R is Bernoulli as in our example, the marginal distribution of Y_i is a mixture of multivariate normal distributions with mean and variance given by

$$E(Y) = \psi\mu^{(1)} + (1 - \psi)\mu^{(0)}$$
$$\mathrm{var}(Y) = \psi(1 - \psi)(\mu^{(1)} - \mu^{(0)})^2 + \psi\Sigma^{(1)} + (1 - \psi)\Sigma^{(0)}$$

It is immediately clear that pattern mixture models are under-identified. If, as in our illustrative example, some data are missing at T, then there are no data to identify either the Tth component of $\mu^{(0)}$ or the Tth row and column of $\Sigma^{(0)}$. Little [80, 81] and Molenberghs *et al.* [84] discuss a variety of parameter constraints for identifying these parameters. For the unstructured version of pattern mixture models, Rubin [77] and Daniels and Hogan [83] propose methods for sensitivity analysis based on between-pattern differences in the mean and variance–covariance parameters.

For longitudinal data, structuring the pattern-specific means as functions of time and other covariates leads to natural and explicit extrapolations. For example, if we assume $\mu_t^{(0)} = \beta^{(0)}t$, and if data are available for at least two distinct values of t among those with $R = 0$, then $\mu_t^{(0)}$ is identified for t by virtue of this assumed structure (e.g. see References [85, 86]). Identification in the presence of covariates is considered in more detail in our application to the CD4 data.

In most cases, mixture models inherently represent non-MAR mechanisms because drop-out can be shown to depend on possibly missing Y's. Again referring to the mixture-of-normals model above, some algebra shows that

$$\mathrm{logit}\{\mathrm{pr}(R_i = 1 \mid Y_{i1}, \ldots, Y_{iT})\} \propto \det\Sigma^{(0)} - \det\Sigma^{(1)} + \{Y_i - \mu^{(1)}\}^{\mathrm{T}}\{\Sigma^{(1)}\}^{-1}\{Y_i - \mu^{(1)}\} - \{Y_i - \mu^{(0)}\}^{\mathrm{T}}\{\Sigma^{(0)}\}^{-1}\{Y_i - \mu^{(0)}\}$$

which is a linear function of Y_{it} and Y_{it}^2 for all t, and all first-order cross-products $Y_{is}Y_{it}$ for all $t \neq s$. Under the constraint that $\Sigma^{(1)} = \Sigma^{(0)}$, logit of the selection probability is linear in the Y_{it}'s.

Frailty models use latent frailties or random effects to induce dependence between the responses Y and the missing data indicators R. Some versions of frailty models also have been referred to as 'shared parameter models' [87, 88], or models with 'random-coefficient-dependent drop-out' [3]. A key feature of these models is that they are specified conditionally on the frailty term as in (6), with the assumption that repeated measures are independent of drop-out times conditional on the frailties.

An early example is given by Wu and Carroll [89], who assume the full-data distribution of the repeated measures follows a linear random effects model where, conditional on random effects η,

$$Y_{it} = \eta_{0i} + \eta_{1i}t + e_{it}$$

where the random effects (η_{0i}, η_{1i}) follow a bivariate normal distribution $\mathrm{N}(\beta, \Omega)$ and the errors e_{it} are iid $N(0, \sigma^2)$. The hazard of drop-out depends not directly on Y's but on the random effects; e.g.

$$\text{logit}\, \lambda_{it} = \psi_0 + \psi_1\eta_{0i} + \psi_2\eta_{1i}$$

Thus, the conditional joint distribution of Y and R given η is structured as $f(y \mid \eta)f(r \mid \eta)$, and the marginal joint distribution is obtained by integrating against $F(\eta)$.

Here we have loosely called η a frailty, though it may not conform to conventional formulations; in particular, for the Wu and Carroll [89] model, η is technically a random coefficient in the response model and a covariate in the hazard model. Importantly, however, data from both Y and R is used to identify the distribution of η. A more traditional parameterization of a frailty model for joint distribution of repeated measures and drop-out is found in Reference [90].

In some cases, the marginal joint distribution derived from a frailty model takes a closed form as either a selection or mixture model; for example, Schluchter [91] assumes a trivariate normal distribution for η_{0i}, η_{1i} and log drop-out time. This can be viewed either as a frailty model where drop-out time is independent of the repeated measures, conditional on η or marginally as either a selection or mixture model. Other examples using multivariate normal distributions include Mori *et al.* [92] and Wu and Bailey [78, 79].

A key property of these models is that identification is driven by the frailty distribution, which usually is assumed to be parametric (e.g. gamma, normal); this choice is usually arbitrary, and may affect validity of results. Another key assumption is conditional independence between Y and R, given η. Under the assumption that other parts of the model specification are correct, this can be tested (e.g.) by including Y's in the selection model [93].

4.5.2. Semi-parametric methods. Theory and methodology for semi-parametric *selection models* under MNAR drop-out is extensively developed in two recent papers by Robins *et al.* [58, 94]. The underlying principle is to assume the selection (drop-out hazard) model can be decomposed as

$$\text{logit}\, \lambda_{it} = h(\mathscr{F}_{it}; \psi) + q(Y_{it}, \ldots, Y_{iT}; \tau)$$

where h and q are known up to their respective parameters ψ and τ. The parameter ψ quantifies how hazard of drop-out depends on known information up to time t, and τ quantifies dependence upon possibly missing response data, conditionally on $\mathscr{F}_{it}$. For example, if $q(Y_{it}, \ldots, Y_{iT}; \tau) = \tau Y_{iT}$, then τ is the difference in log odds of drop-out per unit difference in Y_{iT}, which may be missing for some individuals. Thus if $\tau \neq 0$, S-MAR is violated. See References [58, 94] for discussion of sensitivity analysis strategies when $\tau \neq 0$.

Recently Fitzmaurice and Laird [86] proposed using GEE to fit semi-parametric *mixture models*. The model specification is very similar to parametric models, but only the first moment

is specified. Following the example from Reference [82], the semi-parametric approach would specify

$$E(Y_i \mid X_i, R_i = r) = X_i \beta^{(r)}$$

where X_i is an $n_i \times p$ covariate matrix, and $\beta^{(r)}$ is a $p \times 1$ vector of regression parameters for those with drop-out time r. An additional assumption is that $\text{var}(Y \mid X, R)$ can depend on drop-out time R only through the mean function. Once the drop-out-specific coefficients have been estimated, the effect on the marginal mean can be estimated using the weighted average

$$\hat{\beta} = \sum_r \hat{\pi}_r \hat{\beta}^{(r)}$$

where $\hat{\pi}_r$ is the sample proportion of drop-outs at time r. As with all mixture models that are mixtures of regressions, when the identity link is used, then $E(Y_i \mid X_i) = X_i \beta$, but this does not hold for non-linear link functions such as logit and log. See Reference [86] for details.

4.5.3. Other approaches. Readers are referred to Wu *et al.* [2] for a review of semi-parametric and non-parametric methods, to Yao *et al.* [19] and Glidden *et al.* [95] for estimating location-scale models from incomplete repeated measures and to Lipsitz *et al.* [96] for discussion of quantile regression under MAR.

5. WORKED EXAMPLES

5.1. Analysis of CTQ study using semi-parametric selection models

Our first illustration uses data from the smoking cessation trial described in Section 2 [18]. We present both a semi-parametric and a likelihood-based analysis of these data; the semi-parametric analysis uses inverse probability weighting under S-MAR, and the likelihood-based analysis uses marginalized transition models, recently introduced by Heagerty [46]. A more detailed version of the semi-parametric analysis, including assessment of sensitivity to missingness not at random, can be found in Reference [97].

The CTQ study enrolled 281 healthy but sedentary women in clinical trial to investigate the effect of vigorous exercise as an aid to smoking cessation; 134 were randomized to behavioural therapy only, and 147 to behavioural therapy plus supervised vigorous exercise. The response data consist of binary indicators of cessation status, recorded at baseline and weekly thereafter for 12 weeks. Participants are expected to quit smoking prior to week 5, defined as the *target quit week*. A number of covariates also was measured at baseline, including physiologic measures such as body mass index, behavioural indicators such as stage of change for cessation behaviour, nicotine dependence measures and variables related to prior smoking history and attempts at cessation (Table II).

In the following analyses, we are interested in drawing inferences about assignment to exercise (i.e. the intention-to-treat effect) for the full sample, assuming that smoking cessation data had been available on every subject throughout the trial. For comparison to the regression approaches, we include familiar *ad hoc* methods such as last value carried forward.

Table II. Covariates included in selection model for CTQ analysis.

Covariate	Description	Unit of measure
Z_{1i}	Treatment group	1 = exercise, 0 = control
Z_{2it}	Cessation status at $t-1$ ($Y_{i,t-1}$)	1 = yes, 0 = no
Z_{3i}	Baseline smoking rate	cigs/day (0–60)
Z_{4i}	Baseline nicotine dependence	Fagerstrom score (1–10)
Z_{5i}	Baseline depression	CESD score (0–49)
Z_{6i}	Baseline weight	pounds
Z_{7i}	Duration of longest previous cessation	days
$Z_{8it},\ldots,Z_{16,it}$	Week indicators	binary (0,1)
$Z_{1i}\times(Z_{8it},\ldots,Z_{16,it})$	Treatment × week	
$Z_{1i}\times Z_{3i}$	Treatment × smoking rate	
$Z_{1i}\times Z_{4i}$	Treatment × nicotine dependence	
$Z_{1i}\times Z_{5i}$	Treatment × depression	
$Z_{1i}\times Z_{7i}$	Treatment × longest prev. cessation	
$Z_{1i}\times Z_{2it}$	Treatment × $Y_{i,t-1}$	

5.1.1. Exploratory analyses and ad hoc adjustments for missing data. Figure 1 shows two important processes for the cessation study: the top and middle panels depict proportion quit by week, stratified by treatment group. The difference between these is that the top panel uses number observed as the denominator, whereas the bottom panel uses number randomized, counting drop-outs as smokers. These plots characterize the true underlying cessation status under two different assumptions; the top panel gives a valid estimate under the MCAR assumption (observed data is a random draw from the full data), and the bottom panel is valid under the assumption that all drop-outs are quitters. Although neither of these assumptions can be validated, MCAR can be tested (see Section 5.1.2 for details). As is typical in smoking cessation trials, the proportion quit is nearly zero for the initial 'run-in' period (4 weeks for this trial), then jumps abruptly at the target quit week. In both plots, the proportion quit following the target quit week is appreciably higher among those randomized to exercise.

The bottom panel shows proportion remaining in follow-up as a function of week. Although the drop-out proportion is similar at the end of the trial, the pattern of drop-out is noticeably different: the majority of drop-outs on the control arm drop-out directly following the quit week. Approximately 20 per cent ($\frac{29}{147}$) on the control arm but only 5 per cent ($\frac{7}{134}$) on exercise dropped out before the target quit date (week 5); however, by the end of the trial, 31 per cent on exercise and 35 per cent on control had dropped out. There are several ways to investigate the relationship between drop-out and observed smoking status. Perhaps the simplest is to compare cessation rates between drop-outs and completers, shown in Figure 2 (top panel), which indicates substantial differences.

Various *ad hoc* approaches can be applied, but do not necessarily follow sound principles for inference. For the sake of comparison, we present three of these, with treatment effect defined in terms of the odds ratio for 7-day cessation at week 12: (i) completers-only analysis, (ii) last value carried forward (LVCF) and (iii) all drop-outs counted as smokers. Analysis (i) is valid under MCAR (only), but is inefficient. Analysis (ii) is popular in the reporting of clinical trials, but has two disadvantages: it is a single-imputation procedure, and hence may lead to under-estimated standard errors [25], and it ignores all previous outcome data,

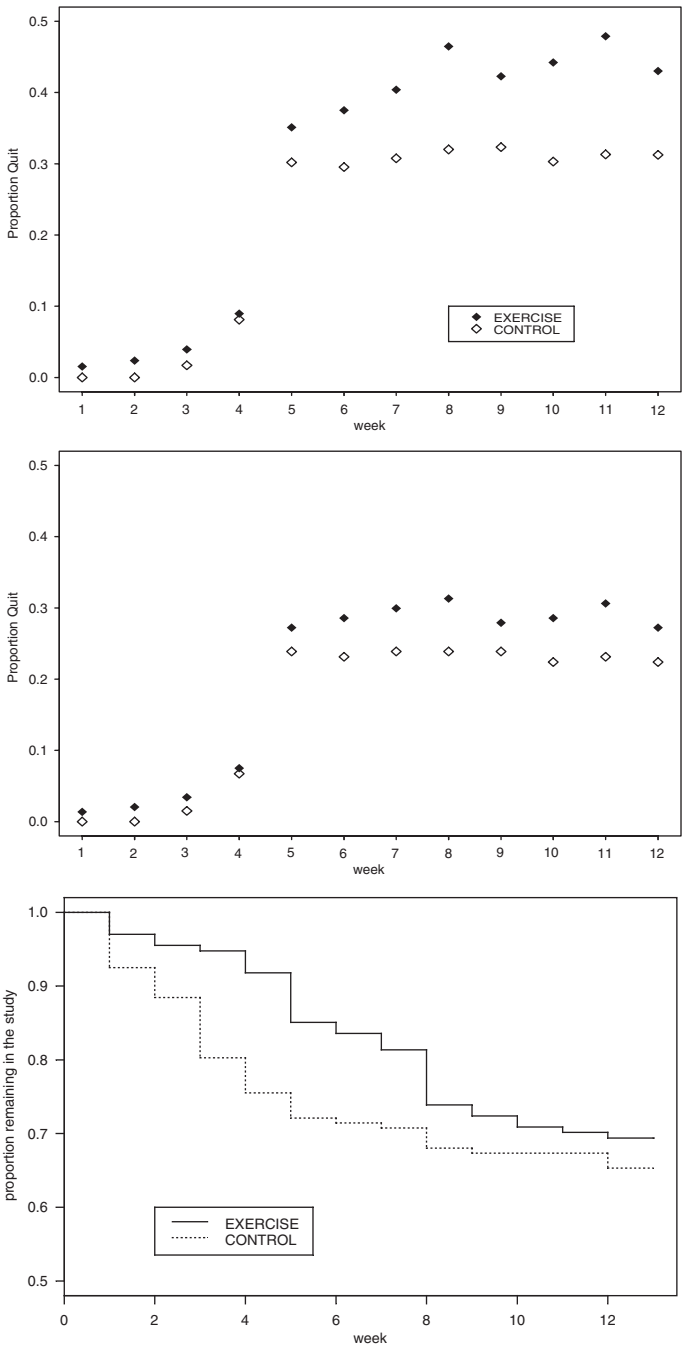

Figure 1. Proportion quit (top and middle panels) and proportion remaining in the study (bottom panel), by treatment group, for the smoking cessation trial. Top panel reflects only observed responses at each week; middle panel treats drop-outs as not quit.

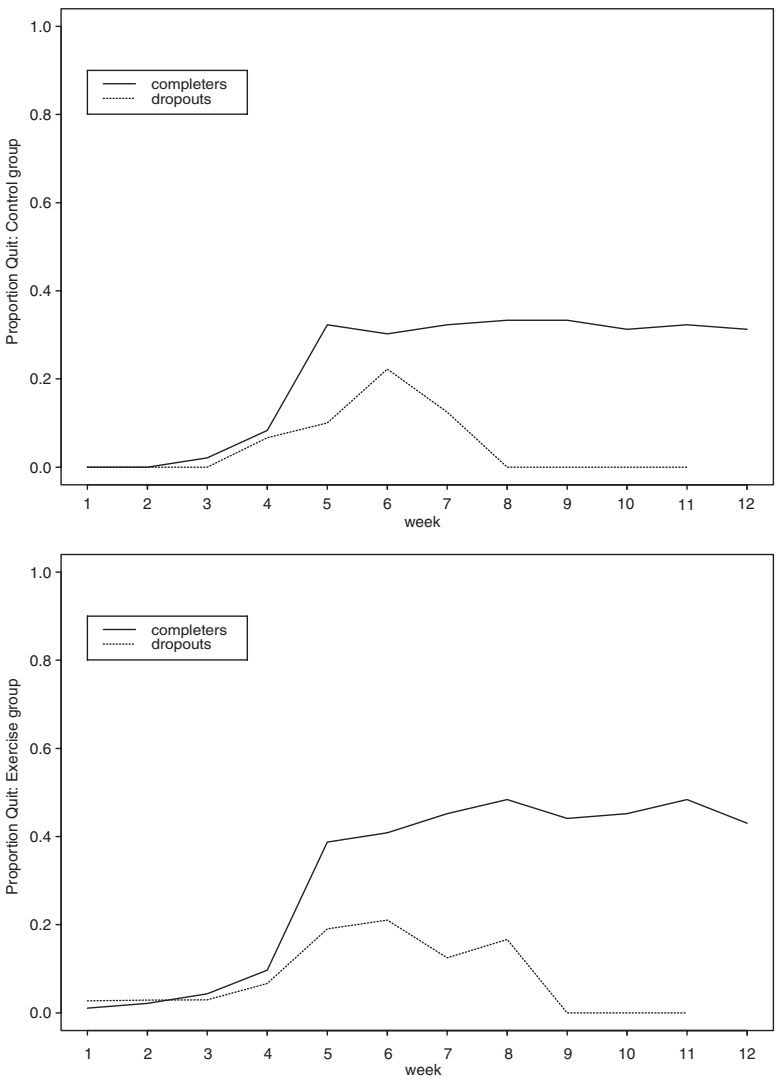

Figure 2. Cessation rate, stratified by study completion status, for each treatment arm in the smoking cessation trial.

which leads to bias if missing smoking status depends on values other than the last one. Analysis (iii) has face-validity for smoking cessation trials [98], but also ignores possible uncertainties about the cessation behaviour of drop-outs. The results from these analyses are reported in Table III, which shows—as expected—that estimated cessation rate is highest when using the completers-only analysis and lowest when counting drop-outs as smokers. Prevalence estimates under LVCF are indeed higher than under the 'drop-outs as smokers' analysis, suggesting that an appreciable number of participants drop-out following a week

Table III. Analyses of smoking cessation data using *ad hoc* methods.

	7-day cessation rate at week 12		
Method	Control	Exercise	Odds ratio (95 % c.i.)
Completers only	0.31	0.43	1.7 (0.9, 3.0)
Last value carried forward	0.24	0.32	1.6 (0.9, 2.7)
Count drop-outs as smokers	0.20	0.30	1.7 (1.0, 2.9)

of cessation. The finding that cessation rates on exercise are higher is robust across these analyses.

5.1.2. Semi-parametric regression under various missing data assumptions. In this section, we describe regression-based estimation of weekly cessation rates and treatment effects under both MCAR and S-MAR. For two candidate models of cessation rate as a function of week, we show how to calculate weights for inverse probability weighting, give the weighted estimating equations used for estimation, and finally summarize the findings.

Response model specification: The first candidate model for the mean response assumes cessation probability is constant for the 4 week period leading up to the target quit week, then allows weekly variations thereafter (9 parameters per treatment group$\times$2 treatments $=18$ parameters total). The second model assumes weekly cessation rate within treatment group is constant over time following the target quit week, thereby reducing the number of mean parameters from 18 to 4. The first model is specified as follows. For $i=1,\ldots,n$ and $t=1,\ldots,12$, let Y_{it} denote cessation status at week t for participant i, where $Y_{it}=1$ if the participant had quit smoking for the previous 7 days, and $Y_{it}=0$ if she had not. Treatment assignment is denoted by X_i ($=1$ if exercise, $=0$ if control). Our model is

$$\text{logit}\{\text{pr}(Y_{it}=1\,|\,X_i)\}=\alpha_t+\beta_t X_i,\quad t=1,\ldots,12 \tag{7}$$

with the constraints $\alpha_1=\cdots=\alpha_4$ and $\beta_1=\cdots=\beta_4$ (constant cessation rate in each treatment arm during the first 4 weeks of follow-up). An equivalent representation that makes the design matrix more transparent is

$$\text{logit}\{\text{pr}(Y_{it}=1\,|\,X_i)\}=I(t\leqslant 4)(\alpha_4+X_i\beta_4)+\sum_{j=5}^{12} I(t=j)(\alpha_j+X_i\beta_j)$$

The β_t parameters are log odds ratios for contrasting treatment arms at each week; treatment-arm-specific cessation rates at each week are easily obtained using the inverse logit function $\text{pr}(Y_{it}=1\,|\,X_i=x)=\exp(\alpha_t+\beta_t x)/\{1+\exp(\alpha_t+\beta_t x)\}$.

The second model further constrains the parameters such that weekly cessation rate is constant within treatment from weeks 5 to 12. This yields

$$\text{logit}\{\text{pr}(Y_{it}=1\,|\,X_i)\}=I(t\leqslant 4)(\alpha_0^*+X_i\beta_0^*)+I(t>4)(\alpha_1^*+X_i\beta_1^*) \tag{8}$$

so that prior to the target quit date, α_0^* and $\alpha_0^*+\beta_0^*$ quantify log odds of weekly cessation in the control and exercise groups, respectively, with α_1^* and $\alpha_1^*+\beta_1^*$ quantifying the same quantities

Table IV. GEE-based estimates of intercept (α_{12}) and log odds ratio for treatment effect (β_{12}) at week 12, using model (7) under MCAR, for various working correlation structures.

Working correlation	$\hat{\alpha}_{12}$ (s.e.)	$\hat{\beta}_{12}$ (s.e.)
Independence	−0.79 (0.22)	0.51 (0.30)
Exchangeable	−0.86 (0.22)	0.39 (0.30)
AR-1	−0.83 (0.22)	0.41 (0.30)
Unstructured	−0.83 (0.22)	0.33 (0.30)

Table V. Number and proportion of transitions from status at $t-1$ to t, from weeks 5 to 12, stratified by treatment. Row totals of proportions sum to one.

		Status at t			
Treatment	Status at $t-1$	Quit	Not quit	Dropped out	Total
Exercise	Quit	286 (0.92)	23 (0.07)	2 (0.01)	311
	Not quit	54 (0.11)	450 (0.85)	28 (0.05)	532
Control	Quit	195 (0.86)	32 (0.14)	2 (0.01)	229
	Not quit	55 (0.09)	526 (0.89)	13 (0.02)	594

following the target quit date. Hence, β_1^* is the log odds ratio that captures time-averaged treatment effect over weeks 5 to 12.

For either model, semi-parametric analysis under MCAR can proceed by fitting the model to observed data using GEE under the independence working correlation assumption, using robust standard errors to adjust for within-subject correlation. When missing data are MCAR and the mean model is correctly specified, GEE will yield consistent estimates and standard errors regardless of the assumed working correlation. Consistency is not guaranteed under MAR however [43, 73], suggesting that for model (7)—which is nearly saturated in the week-by-treatment means—a crude diagnostic for checking the MCAR assumption is to compare point estimates of regression parameters under different working correlations. Large deviations in regression parameters indicates possible departures from MCAR (see Reference [99] for a description of more formal checks of MCAR). Table IV shows point estimates and standard errors for α_{12} and β_{12} when using GEE estimation under model (7); the point estimates vary considerably depending on assumed covariance, suggesting that the MCAR assumption may not be tenable.

Testing the MCAR null hypothesis: Formal tests of the MCAR null hypothesis under monotone drop-out are relatively easy to construct, and for binary data are most naturally formulated in terms of transition probabilities. In the CTQ study, at each week, participants can be in one of three states: quit, not quit or dropped out. Since drop-out is a monotone process, the last of these is an absorbing state. Table V shows transition rates by treatment, aggregated over weeks 5–12.

The MCAR null hypothesis can be stated and tested in terms of parameters from a regression model of the transition probabilities. Specifically, let $R_{it}=1$ if subject i has observed cessation status at time t, and $=0$ otherwise; denote the hazard of drop-out at t by $\lambda_{it}=\text{pr}(R_{it}=0\,|\,R_{it}=1)$. Recall that under MCAR, λ_{it} can be a function of full-data model covariates (in this case week and treatment), but conditional on these, must be independent both of previous cessation outcomes and of excluded covariates. Consider testing the MCAR null hypothesis under full-data model (8), where the covariates are the time variables $I(t\leqslant 4)$ and $I(t>4)$, treatment indicator X_i, and time-by-treatment interactions. Assume that, in addition to the model covariates, hazard of drop-out may possibly depend further on previous cessation outcome $Y_{i,t-1}$ according to the model

$$\text{logit}(\lambda_{it})=I(t\leqslant 4)(\gamma_0+\psi_0 X_i)+I(t>4)(\gamma_1+\psi_1 X_i)+\theta Y_{i,t-1},\quad t=2,\ldots,12 \tag{9}$$

under which the MCAR null hypothesis is $H_0:\theta=0$ and can be tested against S-MAR alternatives using output from a fitted logistic regression. For the CTQ data, $\hat{\theta}=-1.72$ with robust standard error (s.e.) 0.54; the Wald Z statistic is 3.21, strongly suggesting that MCAR is violated. Naturally, the validity of this test requires that (9) is correctly specified, and other types of departures from MCAR are possible; however, (9) is reasonably consistent with the empirical transition rates from Table V, and in our view provides sufficient cause to believe MCAR does not hold.

Weight calculation for S-MAR analysis: Semi-parametric regression under S-MAR can be implemented as a two-step procedure. The first step involves fitting a model for probability of drop-out at each week, from which sampling weights are estimated, and the second step involves fitting model (7) to the observed data, with each observation weighted by the inverse probabilities estimated in step 1. The inferences we describe under S-MAR also require Assumptions 1 and 2.

Let $R_{it}=1$ if Y_{it} is observed for participant i at time t, and $R_{it}=0$ otherwise. Denote the hazard of drop-out as

$$\lambda_{it}=\text{pr}(R_{it}=0\,|\,R_{i,t-1}=1,Z_{it})$$

where $Z_{it}=Z(\mathcal{F}_{it})=(Z_{1it},\ldots,Z_{Pit})$ is a P-dimensional vector of covariates drawn from the individual's data history. From the definition of $\mathcal{F}_{it}$, Z_{it} can include any variables observed up to and including time t, with the exception of Y_{it}. Most of the covariates available in this study are measured only at baseline. Determining which variables to include in the model is somewhat subjective and ideally is done in consultation with subject-matter experts. Our colleagues suggested from prior experience that participants with poor prior history of smoking cessation, or with long history of intensive smoking behaviour, may be more likely to drop-out. Exploratory analyses can also be useful. A simple (but by no means comprehensive) method to screen candidate variables is to compare covariate means between completers and drop-outs.

We estimate the probabilities λ_{it} using logistic regression; in principle, any method that leads to consistent estimates can be used. The model takes the form

$$\text{logit}(\lambda_{it})=Z_{it}\gamma \tag{10}$$

The components of Z_{it} are listed in Table II, and include week, treatment, previous cessation outcome and a variety of individual-specific baseline covariates that are typically related to

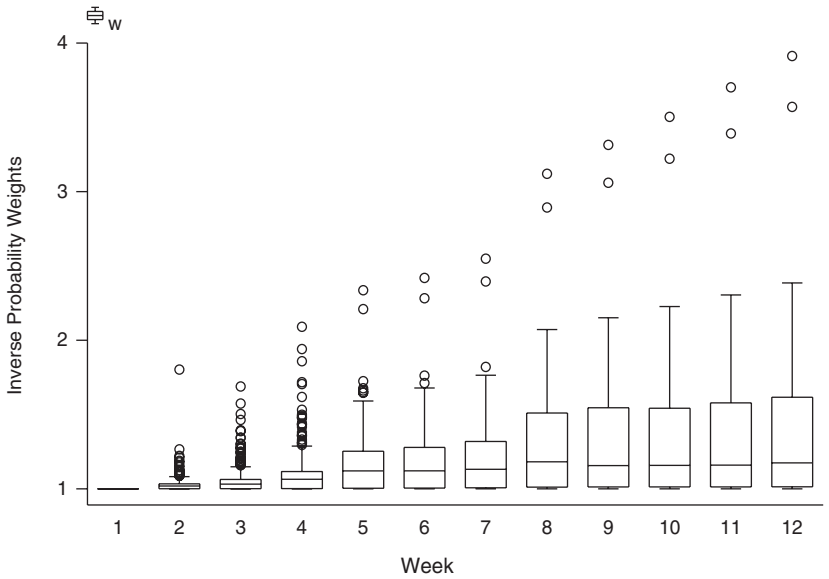

Figure 3. Box plots of inverse probability weights $\hat{\omega}_{it}$ for $t = 5, \ldots, 12$.

success in smoking cessation programs (e.g. depression level, nicotine dependence, duration of previous cessation attempts, etc.). Correct specification of the selection model is a critical component to semi-parametric regression under S-MAR because consistent estimates of the regression parameters in the response model depend on the selection probabilities being consistently estimated. Although the S-MAR null hypothesis cannot be tested, the selection model itself can be critiqued under the assumption that S-MAR holds. Two aspects of model critique are of particular importance: lack of fit and distribution of inverse probability weights (discussed below). For assessing lack of fit, standard methods such as Hosmer and Lemeshow's deciles of risk statistic [100] can be used. For our data, this statistic is 13.2; when referred to chi-square distribution on 8 d.f., $p = 0.10$, indicating satisfactory fit.

Once the selection model is fitted, we obtain the marginal probability that $R_{it} = 1$ using standard conditioning arguments:

$$\hat{\lambda}_{it} = \frac{1}{1 + \exp\{Z_{it}\hat{\gamma}\}}$$

and $\hat{\pi}_{it} = \widehat{\text{pr}}(R_{it} = 1 \mid \mathscr{F}_{it}) = \prod_{j=1}^{t} (1 - \hat{\lambda}_{it})$. For those observations with $R_{it} = 1$, the inverse sampling weights are $\hat{\omega}_{it} = 1/\hat{\pi}_{it}$, which can be interpreted as the number of data points being represented by the observed Y_{it}. Estimation of the response model under inverse weighting can become unstable when the sampling probabilities $\hat{\pi}_{it}$ are very close to zero, leading to outsized weights and attributing undue influence to individual observations; therefore, the distribution of weights should be checked, with particular attention given to extremes in the right-hand tail. Figure 3 shows box plots of $\hat{\omega}_{it}$ at each time point. The variation in weights gets larger with time, but most weights take values between 1 and 2 (the maximum is around 4). Therefore, the estimated π_{it} are bounded well away from zero, with minimum value around 0.25,

and large sample inference can be expected to be stable. The influence of individuals with large weights can be checked using DFBETA-type statistics [101]; see Reference [102] for an illustration.

Estimating full-data regression model via inverse weighting: Once the selection model has been fitted and critiqued, and the weight distribution has been examined, the parameter estimates for model (7) are found as the root of the weighted estimating equations

$$U(\alpha,\beta)=\sum_{i=1}^{n}\sum_{j=1}^{\mathrm{T}} R_{it}\hat{\omega}_{it}K_{it}(X_i^*,\alpha,\beta)\{Y_{it}-\mu_{it}(\alpha,\beta)\}=0$$

where $\alpha=(\alpha_4,\ldots,\alpha_{12})^{\mathrm{T}}$, $\beta=(\beta_4,\ldots,\beta_{12})^{\mathrm{T}}$, $K_{it}(X_i^*,\alpha,\beta)=X_i^{*\mathrm{T}}[\mu_{it}(\alpha,\beta)\{(1-\mu_{it}(\alpha,\beta)\}]^{-1}$, and X_i^* is the 1×18 design matrix needed to parameterize the mean of (7).[§] Estimates for the constrained model (8) follow similarly, using the same weights. Note that $U(\alpha,\beta)$ is a weighted GEE under the working independence assumption; if drop-out is S-MAR, and the selection model has been correctly specified, it is possible to use other working variances and still obtain consistent estimates [73]. Also, it is possible to use augmented estimating equations to increase efficiency [37, 58], but these cannot generally be implemented using standard software and we do not consider them further.

There are several options for standard error estimation. Rotnitzky *et al.* [58] derive an expression that will provide consistent estimates of standard errors; this requires fitting another regression to the score residuals and carrying out some basic matrix computations; another option is the bootstrap [103], wherein entire subject histories are drawn with replacement (i.e. subject is the resampling unit) and the two-step procedure is applied to each bootstrap sample. In a randomized trial, samples should be drawn separately within treatment arm. A third option is to calculate standard errors from the simultaneous system of estimating equations for both outcome and sampling weights. Since each of these options requires extra programming on the part of the user, we recommend following the suggestion of Hernán *et al.* [39], treat the weights as known, and use robust standard error estimates applied to the response model. This will tend to give conservative estimates of standard error.

A comparative summary of cessation rates and treatment effects is shown in Table VI. Clearly the S-MAR estimates of cessation are adjusted downward, a manifestation of the association between smoking status and eventual drop-out. For reporting success rates under either treatment, the MCAR estimates clearly will be overly optimistic. The relative decrease in estimated odds of cessation is about the same in both treatment arms; hence, the odds ratio is the same under both S-MAR and MCAR. The odds ratio confidence interval is slightly more narrow under MAR, but it is worth pointing out that it was computed using a conservative standard error calculation. The added efficiency gained when estimated weights are used (as opposed to treating them as known) would be properly reflected by the consistent estimator described in Reference [58].

5.1.3. Likelihood-based analysis using marginalized transition models. There exist several options for likelihood-based analysis of repeated binary data, including random effects regression models (e.g. logistic-normal [104–106]) and marginal models [44–46, 107]. A potential

[§]There are $12-4+1=9$ parameters for each of two treatment arms because log odds of smoking is assumed constant for the first 4 weeks.

Table VI. Summary of treatment effect estimates from model (8) under MCAR and S-MAR. OR = odds ratio.

			Cessation rate for weeks 5–12		
Assumption	Method/model	Loglik. (params.)	Exercise	Control	OR (95% CI)
MCAR	*GEE	—	0.42	0.31	1.6 (1.0, 2.7)
S-MAR	*GEE-IPW	—	0.33	0.24	1.6 (1.0, 2.6)
MAR	†MTM-1	−610.1 (5)	0.36	0.28	1.5 (0.9, 2.4)
	†MTM-2	−585.7 (6)	0.35	0.27	1.4 (0.9, 2.3)
	†MTM-3	−581.6 (8)	0.36	0.28	1.4 (0.9, 2.3)

*Implemented using working independence assumption with robust standard errors.
†MTM-1 assumes lag 1 correlation only. MTM-2 assumes both lag 1 and lag 2 correlation. MTM-3 allows both lag-1 and lag-2 correlations to depend on time, specifying different values at the target quit date (week 5).

drawback to using random effects models for marginal inferences is the need to integrate over the random effects distribution [49]. For direct comparison to the semi-parametric marginal model, we opted to use a marginalized transition model, or MTM [46], which allows separate specification of the marginal mean and the serial correlation. Details on likelihood-based estimation are found in Reference [46].

The MTM calls for specifying the marginal mean $\mu_{it} = E(Y_{it} \mid X_{it})$ and the conditional mean $\mu^*_{it} = E(Y_{it} \mid X_{it}, \bar{Y}_{i,t-1})$. Parameters in the conditional mean quantify serial correlation. For the marginal mean we use models (7) and (8). The serial correlation is specified by assuming

$$\text{logit}\, \mu^*_{it} = \phi_0 + \phi_{1t} Y_{i,t-1} + \phi_{2t} Y_{i,t-2}$$

so that for $l = 1, 2$, serial association at lag l is

$$\phi_{lt} = \log\{\text{odds}(Y_{it} = 1 \mid Y_{i,t-l} = 1)/\text{odds}(Y_{it} = 1 \mid Y_{i,t-l} = 0)\}$$

Three formulations of the serial correlation are considered:

$$\begin{aligned}
&\text{MTM 1} && \phi_{1t} = \alpha_1 \\
&&& \phi_{2t} = 0 && \text{(lag-1 correlation only, constant for all } t\text{)} \\
&\text{MTM 2} && \phi_{1t} = \alpha_1 \\
&&& \phi_{2t} = \alpha_2 && \text{(lag-1 and lag-2 correlation, constant for all } t\text{)} \\
&\text{MTM3} && \phi_{1t} = \alpha_1 I(t \neq 5) + \alpha_1' I(t = 5) \\
&&& \phi_{2t} = \alpha_2 I(t \neq 5) + \alpha_2' I(t = 5) && \text{(serial correlation differs at week 5)}
\end{aligned}$$

The justification for including Model 3 is that transitions in smoking behaviour are likely to be much different at the target quit date, week 5. To avoid bias in standard error estimation, in principle the correlation should be modelled with sufficient complexity. We considered models where serial correlation depends on treatment, but the treatment effect was very weak and its inclusion did not affect standard errors.

Results from the fitted MTM appear in Table VI and in Figure 4. Comparing likelihoods suggests that Model 3 is most appropriate. Estimated cessation probabilities lay roughly at the mid-point between those obtained from GEE under MCAR and GEE-IPW under S-MAR; the odds ratio for treatment effect, 1.4, is nearly the same as for the other two methods, but the confidence intervals are more narrow, as would be expected when specifying the full joint distribution.

5.2. Analysis of HIV data from an observational cohort study

Our second illustration is based on the HIV data described in Section 2.2. The goal in this analysis is to characterize changes in mean CD4 count over time as a function of baseline covariates. We begin with some exploratory analyses, investigate several regression models and compare the results in the context of each model's assumptions about the reasons for subject drop-out.

5.2.1. Exploratory analyses. For these analyses, we use data from the 871 women who were HIV positive at baseline. We excluded 21 women from the analyses because they were missing CD4 data at baseline. There was a substantial amount of missing data in this study. Table I shows the number of subjects still in follow-up at each visit (i.e. who have not yet dropped out), the number with CD4 count observed and the number that dropped out before the next visit. For example, at visit 1 there were 850 subjects with CD4 count observed, and 52 of them dropped out before the next visit. At visit 2,798 women remained in the study, but only 706 of them had an observed CD4 count. The other missing values resulted from intermittent missingness, for reasons other than drop-out. (For our regression analysis, data from another five women were excluded for missing one or more covariates.)

Figure 5 shows the mean CD4 count at each visit for observed data (horizontal marks), which indicates a decrease over time for the first five visits, followed by a levelling off. This cannot necessarily be interpreted as the population trend for the women who were enrolled at baseline because a substantial number of women dropped out of the study before the 12th visit. Figure 5 also stratifies the mean at each visit by whether or not the subject drops out at the next visit. In every case, the mean was lower for subjects who drop-out at the next visit. That women with lower CD4 count are more likely to drop-out suggests that drop-out depends at least on observed CD4 counts, and implies that model-based means are likely to be quite different from observed means. Handling intermittent missingness also is a concern, which we address below.

The primary objective is to characterize covariate effects on the mean. We assume that for the full data

$$E(Y_{it}) = X_{it}\beta, \;\; t = 1, \ldots, 12 \tag{11}$$

where Y_{it} is the square root of CD4 count at visit t and X_{it} is the vector of covariates at visit t, which consists of: visit number (time); levels of HIV-1 RNA in the plasma (copies/ml); HIV symptomatology (presence of HIV-related symptoms on a scale from 0–5); indicator of antiretroviral therapy (ART) at baseline; and the number of years the subject was aware of her HIV status at enrolment (range 0–8 years). In addition, we consider interactions between visit and the baseline covariates. The square root transformation on CD4 is used as the response to reduce skewness. Based on exploratory analyses, a linear time trend seems reasonable

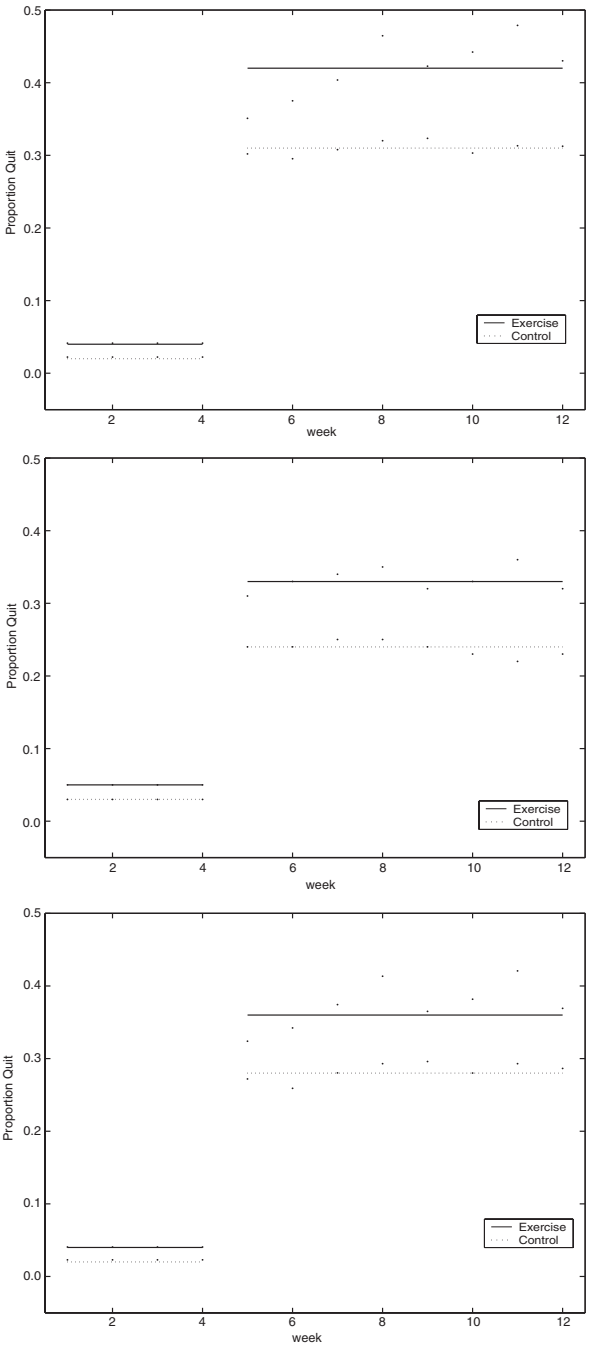

Figure 4. Fitted models under MCAR (top panel), S-MAR (middle panel) and MAR (bottom panel). Points represent estimated cessation rates under model (7); lines represent rates from model (8).

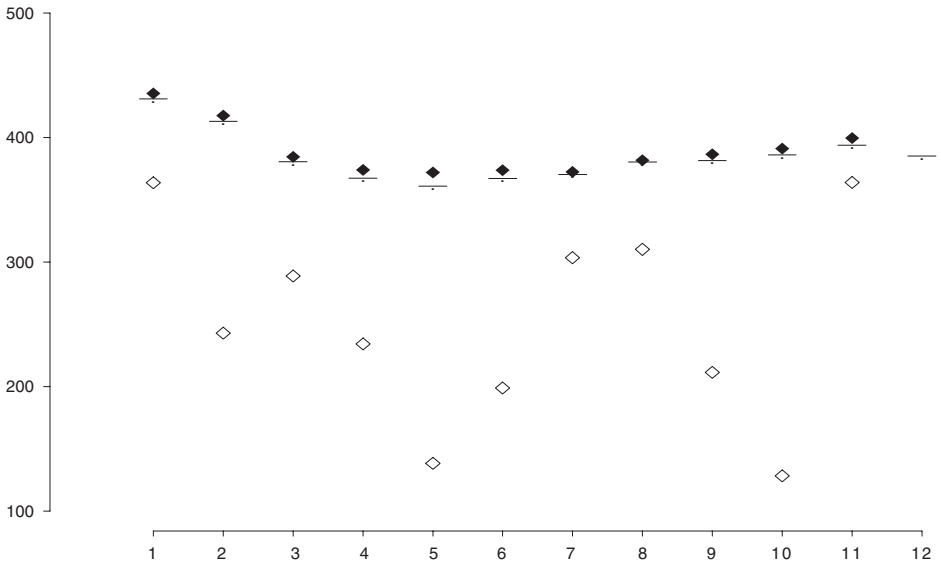

Figure 5. Sample mean of CD4 count at each visit. The horizontal lines are the overall means; the closed diamonds are the means for subjects that do not drop out at the subsequent visit; the open diamonds are the means for subjects that drop out at the subsequent visit.

(i.e. treating visit as a continuous covariate). Since the distribution of viral load tends to be highly skewed, four indicator variables were created corresponding to viral load intervals. In the next subsections, we apply regression models for estimating β that make progressively weaker assumptions about missing data.

5.2.2. Analysis under the MCAR assumption. For purpose of illustration, we fit regression model (11) under the assumption that all missingness is MCAR. The variance of Y_{it} given X_{it} is assumed to be a constant, σ^2. We use the ordinary least squares (OLS) method for estimation, which provides consistent estimates of β under MCAR. Standard errors are obtained using the sandwich estimator [43]. Although technically we have specified only the first two moments and not a full likelihood, the OLS estimator is equivalent to the maximum likelihood estimator under the assumption that $Y_i \sim \mathrm{N}(X_i\beta, \sigma^2 I)$. Estimates are obtained using SAS PROC MIXED; results are given in Table VII under the MCAR heading.

The main findings from this model are that lower values of viral load are associated with higher CD4 counts, but mean CD4 count tends to decline more rapidly for the group with the lowest viral load; ART therapy at baseline also is associated with lower CD4 counts (probably because those on therapy have more advanced HIV disease). The CD4 slope, however, is significantly greater for those on ART at baseline. Not surprisingly, mean CD4 count is lower for those with longer awareness of their HIV status.

5.2.3. Analysis assuming ignorability. As previously indicated, the MCAR assumption is probably implausible for these data. One way to weaken the MCAR is to use likelihood-based inference. We specify the distribution for the response using a random effects model,

wherein we assume that each woman has individual-specific intercept and slope (random effects), which are normally distributed and may be correlated. We also assume the within-subject errors are normally distributed but uncorrelated given the random effects. As we show here, the marginal likelihood of the full data follows a multivariate normal distribution.

At the first level, we assume that conditional on random effects η_i, the full data follows

$$Y_i \sim \mathrm{N}(X_i\beta + Z_i\eta_i,\ \sigma^2 I)$$

where Y_i is the 12-by-1 vector of fully observed (square root) CD4 counts, X_i is the fixed effects design matrix, Z_i is a 12×2 design matrix whose first column is a vector of ones and second column is the measurement times, η_i is the 2×1 vector consisting of individual-specific deviations from the population intercept and slope. At the second level, we assume that $\eta_i \sim \mathrm{N}(0, \Psi)$, where Ψ is a 2-by-2 variance matrix. The marginal distribution of Y given X is a normal mixture of normal distributions; therefore

$$Y_i \sim \mathrm{N}(X_i\beta, Z_i\Psi Z_i^{\mathrm{T}} + \sigma^2 I)$$

and $E(Y_{it} \mid X_{it})$ retains the form given in (11).

Consistent estimates of the parameters from this model can be obtained by maximizing the observed data likelihood under ignorability (MAR plus the separable parameters assumption), treating both drop-outs and intermittently missing values as arising from the same mechanism. By specifying a distribution for the response (e.g. normal) and basing inference on the observed data likelihood, we admit a more plausible assumption about the missing data than the previous analysis, but at the expense of having to justify the distributional assumptions. Estimates from this model also were obtained using SAS PROC MIXED; with results given in Table VII under the MAR heading.

The primary difference between the results from the models under the MCAR and MAR assumptions is that the estimated slope coefficient—which is average rate of change per visit among those with covariate values equal to zero or equal to the reference category—has changed drastically from 0.03 (s.e. 0.17) to –0.69 (s.e. 0.19). Therefore, by accounting for the fact that drop-out may be related to observed values of CD4 count, we now conclude that CD4 count decreases over time for this population, because none of the interaction terms exceeds 0.69. This shift in estimated slope makes intuitive sense given in Figure 5, where it is clear that subjects who drop-out consistently have lower CD4 counts, on average. Under the MCAR assumption, subjects who drop-out sooner contribute fewer observations to the population estimate of the slope, which results in the slope estimate being largely determined by those that remain in the study the longest. By contrast, the random effects model estimates the population slope as a weighted average of individual-specific slopes, which themselves are averages of individual-specific OLS slope estimates and the MCAR population slope estimate [47]. We therefore would expect the slope estimates from the analysis assuming MCAR to overestimate of the rate of change in CD4 count over time. This is manifestation of selection bias, where sicker patients are getting ‘selected’ out of the sample.

As expected, the coefficients of the interaction terms between viral load and time are quite a bit different under models that do not require MCAR. Under the MAR assumption, the change in CD4 count over time was estimated to decline the most for subjects with the highest viral load ($>30\,000$copies/ml) at baseline. This is the opposite of what was found under the MCAR assumption, where those with least viral load (0–500 copies/ml) at baseline were estimated to

Table VII. Results from random effects models for HIV data under various assumptions about the missing data.

Covariate	MCAR est (se)	MAR est (se)	PMM (2 cat) est (se)	PMM (3 cat) est (se)
Intercept	14.25 (0.83)	15.11 (0.77)	15.73 (0.86)	15.92 (0.85)
Time (visit)	0.03 (0.17)	−0.69 (0.19)	−0.79 (0.15)	−0.87 (0.16)
Viral load				
0–500	11.31 (0.82)	10.62 (0.77)	9.99 (0.89)	9.70 (0.88)
500–5k	7.61 (0.79)	6.91 (0.73)	6.22 (0.83)	6.05 (0.84)
5k–30k	3.30 (0.90)	2.98 (0.83)	2.9 (0.92)	2.76 (0.91)
> 30k	ref	ref	ref	ref
HIV symptoms	−0.01 (0.21)	−0.04 (0.2)	−0.04 (0.19)	−0.05 (0.19)
Art at baseline	−4.63 (0.41)	−4.49 (0.39)	−4.26 (0.39)	−4.31 (0.39)
Years aware of HIV	−0.37 (0.10)	−0.39 (0.09)	−0.40 (0.09)	−0.39 (0.09)
Time*viral load				
0–500	−0.47 (0.16)	0.26 (0.19)	0.37 (0.16)	0.60 (0.17)
500–5k	−0.33 (0.16)	0.25 (0.19)	0.24 (0.15)	0.27 (0.15)
5k–30k	−0.05 (0.18)	0.20 (0.20)	0.17 (0.16)	0.14 (0.16)
> 30k	ref	ref	ref	ref
Time*HIV symptoms	−0.03 (0.03)	−0.03 (0.03)	−0.02 (0.03)	−0.01 (0.03)
Time*art at baseline	0.22 (0.06)	0.14 (0.06)	0.13 (0.07)	0.17 (0.08)
Time*years aware of HIV	0.02 (0.01)	0.04 (0.01)	0.04 (0.02)	0.03 (0.02)

have the steepest decline in CD4 count. Again, the difference in the estimates is a result of selection bias being ignored in the MCAR analysis: because subjects with higher viral load tend to have lower CD4 counts, and those with lower CD4 drop-out earlier, we expect biases due to the MCAR assumption to be most pronounced for the subpopulation of subjects with the highest viral load.

5.2.4. A pattern mixture model analysis. The foregoing analyses are only valid under MCAR and MAR, respectively. However, it is not hard to imagine that propensity for drop-out may be related to an individual's unobserved CD4 count, even after conditioning on their previous CD4 counts and baseline covariates. For example, consider two subjects who had the same CD4 and covariates at baseline. Imagine that one subject had a substantial decline in CD4 during the 6 months between visits and the other subject's CD4 stayed relatively stable. The subject whose CD4 declined may be more likely to drop-out of the study, although this cannot be confirmed from observed data. To allow for the possibility that missing data are MNAR we use a pattern mixture model.

As described in Section 5, pattern mixture models assume that the distribution of Y is a mixture over the patterns of missing data. As can be seen from the sample sizes listed in Table I, one challenge with these data is the number of subjects in a given pattern may be quite small; for example, only 19 subjects dropped out after the eighth visit. With small drop-out strata, it may not be practical to assume unique coefficients for every drop-out pattern. As a result, we group the drop-out times, and assume that the distribution of Y is a mixture over groups of drop-out times.

The first challenge is to determine how the drop-out times should be grouped. To investigate this, we fit a linear mixed model with all of the covariates interacted with drop-out indicators.

This led to separate estimates of the covariate effects for every drop-out time. We then created the plots for each covariate, where the vertical axis represented the value of the coefficient and the horizontal axis represented the drop-out times. The idea is to use these plots to identify natural groupings for the patterns. For example, if the estimated coefficients are relatively constant for the first three drop-out times, but increase at the fourth drop-out time, then the first three drop-out times would be grouped together. Upon inspection, grouping the drop-out times into the following three categories seemed reasonable: pattern 1 if the subject's final visit was between visit 1 and visit 5; pattern 2 if the subject's final visit was between visit 6 and visit 10; pattern 3 if the subject's final visit was after visit 10 (includes completers). This method for choosing groups is subjective, and therefore sensitivity of the results to different groupings is examined.

We fit a pattern-mixture model with unique coefficients for each of the three groups of drop-out times. Specifically, let $G_i = j$ if subject i's drop-out time was in pattern j ($j = 1, 2, 3$). We assume

$$[Y_i \mid X_i, \eta_i, R_i, G_i = j] \sim \mathrm{N}(X_i \alpha^{(j)} + Z_i \eta_i, \sigma^2 \mathrm{I})$$

where X_i, Z_i and η_i were defined previously, $\alpha^{(j)}$ is the vector of regression coefficients for the jth grouped pattern. At the second level we assume

$$[\eta_i \mid R_i, G_i = j] \sim \mathrm{N}(0, \Psi)$$

Note that in addition to assuming normality within pattern, this model makes several structural assumptions. First, the conditional distribution of Y is assumed to depend on the drop-out *time* R only through the drop-out *category* (or *pattern*) G. That is, Y is independent of R within pattern. Second, intermittent missingness is assumed to be MAR, given G (within pattern). Third, we assume $\mathrm{var}(Y_i \mid X_i, \eta_i, R_i, G_i) = \mathrm{var}(Y_i \mid X_i, \eta_i)$ and $\mathrm{var}(\eta_i \mid X_i, R_i, G_i) = \mathrm{var}(\eta_i \mid X_i)$, which means variances are constant across patterns (this assumption can easily be relaxed). Fourth, an important identifying assumption implicit in the above model is that covariate effects are the same for missing and observed data within drop-out pattern G; i.e.

$$E(Y_{\mathrm{obs},i} \mid G_i, \eta_i, X_i) = E(Y_{\mathrm{mis},i} \mid G_i, \eta_i, X_i)$$

which clearly is an untestable assumption. It should be noted that while these four assumptions are strong, they can be viewed as more general than the MAR assumptions. In particular, the constraint $\alpha^{(1)} = \alpha^{(2)} = \alpha^{(3)} = \beta$ yields the standard random effects model for which likelihood-based inferences assume MAR. A fifth and final assumption is a structural one concerning covariate distributions. The structure of the mean allows $E(Y \mid X, \eta, G)$ to depend on drop-out group G, but we assume covariates are equally distributed across G, or equivalently that $\mathrm{pr}(G_i = j \mid X_i) = \mathrm{pr}(G_i = j)$. This facilitates averaging regression coefficients over pattern to obtain marginal covariate effects.

Table VIII shows the estimated coefficients and standard errors for each drop-out group. In addition, p-values for Wald tests of the hypothesis that each of the covariates have a constant effect across patterns are presented in Table VIII. Note that this test can be viewed as a test of the MAR null hypothesis, but only under the (untestable) assumption that the pattern mixture model is correctly specified. All of the coefficients do vary significantly across patterns, except for the coefficient of HIV symptomatology. The estimated main effect of time was negative for each pattern, but decreased in magnitude as the drop-out time increases. Lower levels of

Table VIII. Estimates and standard errors of coefficients from the 3 category pattern mixture model. The 3 patterns consist of subjects whose last visit was from 1 to 5, 6 to 10 and 11 to 12, respectively. *P*-values for tests of equality of the coefficients across groups are given.

Covariate	Pattern 1 est (se)	Pattern 2 est (se)	Pattern 3 est (se)	*p*-value
Intercept	14.66 (1.23)	17.05 (1.91)	16.14 (1.20)	<0.001
Time	−2.53 (0.40)	−1.68 (0.32)	−0.10 (0.19)	<0.001
Viral load				
0–500	10.29 (1.41)	8.82 (2.01)	9.67 (1.22)	
500–5k	5.83 (1.16)	3.62 (1.64)	6.63 (1.20)	<0.001
5k–30k	1.34 (1.35)	0.87 (1.90)	3.67 (1.29)	
$> 30k$	ref	ref	ref	
HIV symptoms	−0.08 (0.36)	−0.90 (0.52)	0.14 (0.24)	0.33
Art at baseline	−3.12 (0.94)	−4.52 (1.03)	−4.71 (0.46)	<0.001
Years aware of HIV	−0.42 (0.21)	−0.07 (0.21)	−0.45 (0.11)	<0.001
Time*viral load				
0–500	2.89 (0.48)	1.12 (0.34)	−0.35 (0.18)	
500–5k	1.44 (0.39)	0.89 (0.31)	−0.29 (0.18)	<0.001
5k–30k	0.57 (0.39)	0.48 (0.34)	−0.09 (0.19)	
$> 30k$	ref	ref	ref	
Time*HIV symptoms	0.09 (0.10)	0.09 (0.10)	−0.06 (0.03)	0.05
Time*art at baseline	0.09 (0.27)	−0.07 (0.20)	0.25 (0.06)	<0.001
Time*years aware of HIV	0.01 (0.06)	0.03 (0.04)	0.04 (0.01)	0.03

viral load at baseline were associated with increases in CD4 count over time for subjects that dropped out before visit 11, but were associated with decreases in CD4 count over time for subjects whose last visit was 11 or 12.

While the estimates in Table VIII provide a great deal of insight into the differences between subjects that drop-out at different times, inference about the CD4 count that is unconditional on drop-out times are generally of interest. We therefore calculated the estimated marginal effect of the covariates by taking a weighted average of the conditional effects over the distribution of drop-out times

$$\hat{\beta} = \sum_{j=1}^{3} \hat{\alpha}^{(j)} \hat{\pi}_j$$

where $\hat{\alpha}^{(j)}$ are the MLEs of $\alpha^{(j)}$ from the pattern mixture model (Table VIII), and $\hat{\pi}_j$ is the estimated proportion of subjects with $G_i = j$. The results are given in Table VII under the PMM (3 cat) heading.

The estimated main effect of time is −0.87 (s.e. 0.16), which is larger in magnitude than the estimate under MAR, and substantially different than the estimate under MCAR (standard errors are calculated using the delta method; see Reference [85]). The primary difference between the MAR and pattern mixture analyses is that MAR model assumes the slope of CD4 count over time is the same across patterns; however, from Table VIII it is clear that subjects who dropped out early tended to have a sharper decline in CD4 count over time. As a result, the slope under MAR may be overestimated. Similarly, the mixture model estimates show a larger difference in the slope of CD4 count within viral load categories. In fact, the

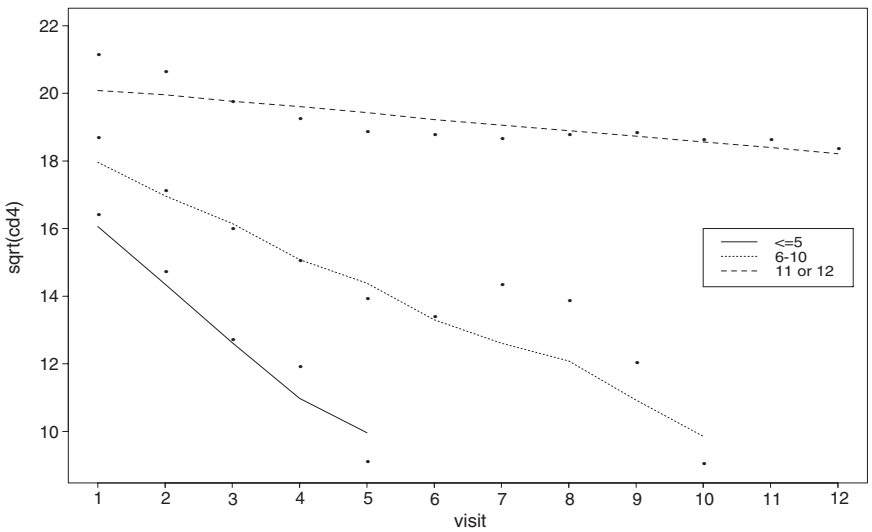

Figure 6. Plot of the predicted means (lines) of CD4 count versus the observed means (points), stratified by drop-out categories, at each visit.

slope for the group with the lowest viral load is significantly different (larger) than the slope for the other viral load groups.

5.2.5. Model diagnostics. We checked the assumptions and fit of the pattern mixture model by (i) looking at the sensitivity of inference to the grouping of drop-out times; (ii) seeing how well the predicted means of CD4 count match the observed means for each drop-out category; (iii) looking at residual plots to identify outliers and departures from the model.

To investigate the sensitivity of inferences to the choice of these specific drop-out categories, we fit a model with two drop-out categories (pattern 1 if last visit prior to visit 11, pattern 2 otherwise). That is, we assumed the marginal distribution of Y is a mixture over these two patterns. The results, after marginalizing over the distribution of the patterns, are given in Table VII under the PMM (2 cat) heading. Comparing the results from the two and three category models, we see that most of the estimates and standard errors are quite similar. The primary difference is for the estimates of the effect of the interaction between time and viral load categories. Both the Bayesian information criterion (BIC) and Akaike's information criterion (AIC) are larger for the model with three categories.

We next investigated how well the model-based visit-specific means fit with the observed data, within pattern. Model-based means were calculated as the sample average of the visit-specific predicted values among all subjects in a given pattern. Figure 6 compares these predicted means to the corresponding sample means from the observed data. For subjects in the second drop-out category, the sample means are slightly higher than predicted by the model for visits 7–9. Overall, though, the predicted and sample means corresponded quite well with one another.

Finally, a plot of the residuals against the predicted values at each visit is given in Figure 7. The residuals generally seem to be randomly dispersed about zero with constant variance. From

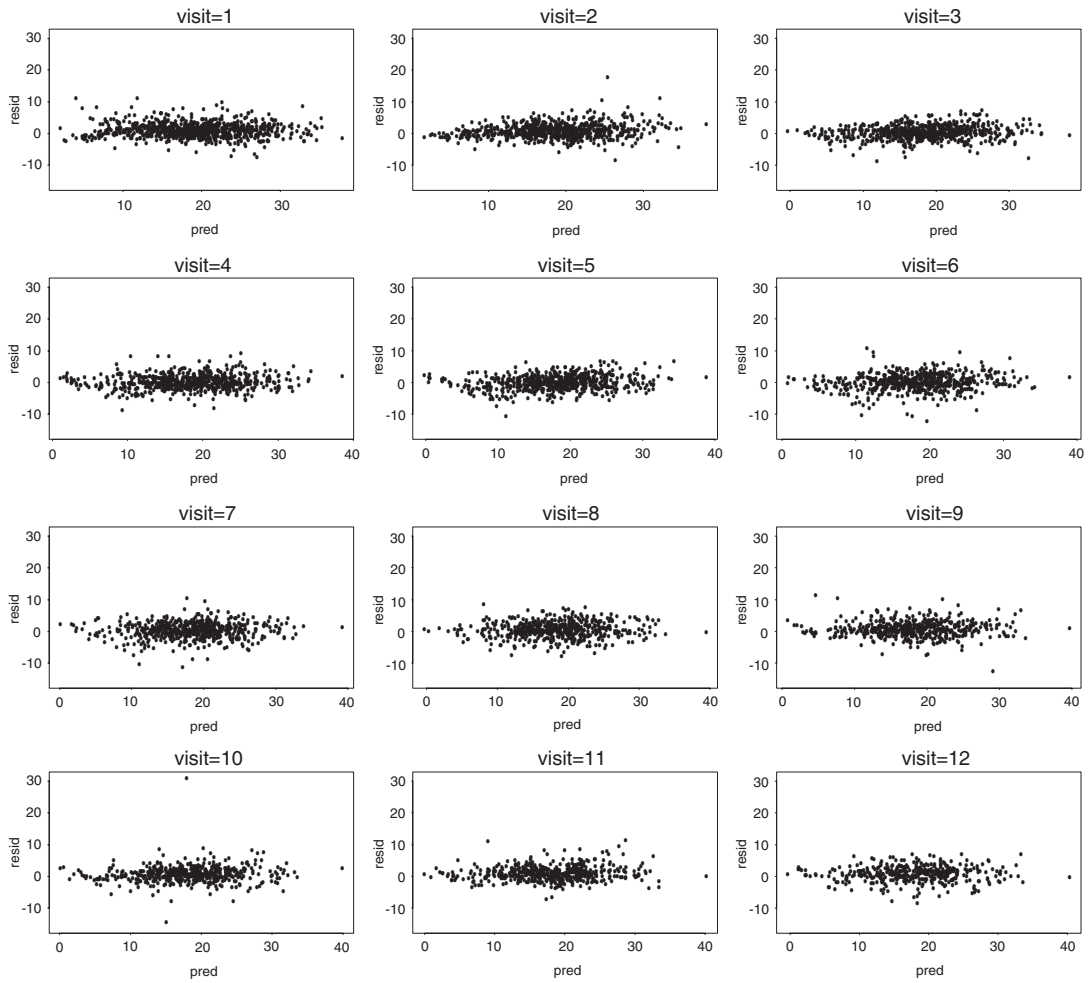

Figure 7. For 3-category pattern mixture model, plot of the residuals against the predicted values at each visit.

the plot we did identify a potential outlier at visit 10 (the residual has value of about 30). We removed this subject from the data and fitted the models again, but this subject's data had little impact on the estimates and standard errors.

6. DISCUSSION

Drop-out and missing data are important issues in longitudinal data analysis, and despite a flurry of recent research activity on methods for handling drop-out, it is clear that no single methods is flexible enough to handle all—or even most—instances of drop-out. In our tutorial, we have focused on three key issues: terminology and assumptions, a review of methods and

models and detailed illustrations of two methodologies. The first four sections of our tutorial are designed to give readers a context within which to engage the still evolving literature on methods for drop-out and missing data; the fifth section (with the applications) can be used as a template for implementing various popular and flexible methods; here we illustrated use of inverse probability weighting, joint likelihood analysis for repeated binary responses and pattern mixture models. In short, it is our hope to have put some modern and powerful tools into the hands of practitioners.

Since this tutorial is by nature introductory, there are a number of important issues that either have not been included here or have been discussed only superficially. Perhaps foremost among these is sensitivity analysis, for the simple reason that models of outcome-related drop-out rest heavily on assumptions cannot be empirically critiqued. In fact the design, execution and interpretation of sensible sensitivity analyses is gaining momentum as a research area in its own right. Readers are referred to References [77, 82, 83, 108] for methodology related to pattern-mixture models. Other key references include References [69, 109]. Information on sensitivity analysis in semi-parametric models can be found in References [13, 94].

Properly handling death as a cause of drop-out is among the more vexing issues in an analysis of longitudinal data. Most methods for drop-out assume that subjects who drop-out could have been measured after their drop-out time, which seems implausible when drop-out results from death. Two possible approaches have been outlined by Rubin and Frangakis [110] and Robins *et al.* [111], who essentially envision inferences about the subpopulation of individuals who would survive, or who have non-zero probability of surviving, to a certain time t. Again because of the introductory nature of our tutorial, we have sidestepped the issue in our analysis of the HIV data in Section 5, but it remains an important one to resolve.

A third topic that appears to be drawing attention of methodologists is multiple-cause drop-out. In our examples, we have treated all drop-outs the same, but in practice participants may have different reasons for dropping out; these types of drop-out may be related to outcomes in different ways (or not at all). A simple example is when some subjects have outcome-related drop-out and others do not [85, 89]; in other cases, drop-out may be related to treatment for some and to outcome for others [13].

Finally, there does not appear to exist a unified terminology for describing drop-out mechanisms in longitudinal studies. Clearly the framework of Rubin [50] forms the basis for classifying missing data mechanisms, but its application to longitudinal data is not straightforward. We have attempted to draw what we view as a key distinction between sequential MAR, which seems naturally suited to stochastic process formulations, and MAR, which is a multivariate version of definitions given in References [25, 50]. To get a sense of the various ways MAR is defined for longitudinal data, readers are directed to papers by Diggle and Kenward [8], Little [3], Robins *et al.* [37], Fleming and Harrington [52, p. 100] and Gill and Robins [59].

A substantial portion of information for decision making in public health and clinical medicine derives from longitudinal studies of various types, including clinical trials and cohort studies, and drop-out tends to be more the rule than the exception. Failure to handle it correctly leaves the results susceptible to selection biases, which in turn may lead to erroneous conclusions and poorly informed decisions. Applied with care, the methods reviewed and illustrated here can help analysts and decision makers appreciate the nature of possible selection biases, to correct them to some degree, but perhaps most importantly to understand the limitations of the information contained in their data.

APPENDIX A: SMOKING CESSATION EXAMPLE

This appendix includes data excerpts and sample code for the models fit in our case studies. For simplicity, the sample code for the CTQ study ignores data from the first 4 weeks (although our data analysis includes those observations).

A.1. Data excerpt

```
ID:              identification number for each subject
WEEK:            variable indicating the week number
R:               missing data indicator (1 = observed, 0 = missing)
TX:              treatment indicator (1 = exercise, 0 = control)
QUIT:            smoking status indicator (1 = quit, 0 = not quit)
WEIGHT:          baseline weight in pounds
FAGER:           baseline nicotine dependence score (Fagerstrom Index)
RATE:            baseline average number of cigarettes smoked per day
CESD:            baseline depression score (CESD scale)
MAXQUIT:         the longest interval in days of previous quit attempt

SUBJECT:=001
 id    tx    week    quit    r    weight    fager    brate    bcesd    maxquit
001     0       1       0    1       238        8       60       32         57
001     0       2       0    1       238        8       60       32         57
001     0       3       0    1       238        8       60       32         57
001     0       4       0    1       238        8       60       32         57
001     0       5       1    1       238        8       60       32         57
001     0       6       1    1       238        8       60       32         57
001     0       7       1    1       238        8       60       32         57
001     0       8       1    1       238        8       60       32         57
001     0       9       1    1       238        8       60       32         57
001     0      10       1    1       238        8       60       32         57
001     0      11       1    1       238        8       60       32         57
001     0      12       1    1       238        8       60       32         57

SUBJECT:= 003
 id    tx    week    quit    r    weight    fager    brate    bcesd    maxquit
003     1       1       0    1       199        7       20        6         15
003     1       2       0    1       199        7       20        6         15
003     1       3       0    1       199        7       20        6         15
003     1       4       0    1       199        7       20        6         15
003     1       5       1    1       199        7       20        6         15
003     1       6       1    1       199        7       20        6         15
003     1       7       .    0       199        7       20        6         15
```

A.2. Sample SAS code for semi-parametric regression analysis

```
/* DATA STORED LOCALLY IN FILE DATA.SAS7BDAT */
libname sss '';

/* RESTRICT ANALYSIS TO DATA FROM WEEKS 5 TO 12 */
data model; set sss.data; if WEEK ge 5; N=1;

/* LOGISTIC REGRESSION MODEL FOR ESTIMATING WEIGHTS */
```



```
proc genmod data=model;
class ID WEEK;
model R/N = WEEK TX TX*WEEK QUITPR BRATE TX*BRATE WEIGHT FAGER TX*FAGER
            BCESD TX*BCESD MAXQUIT TX*MAXQUIT / link=logit dist=bin obstats;
make 'obstats' out=stats;

/* DATA STEP FOR COMPUTING WEIGHTS
   PRED:  MARGINAL PROBABILITY OF BEING OBSERVED AT T
   OBSER: CUMULATIVE PROBABILITY OF BEING OBSERVED AT T
   W:     INVERSE OF THE CUMULATIVE PROBABILITY (WEIGHT) */
proc sort data=stats; by ID WEEK;
data weight; set stats; by ID WEEK;
if first.ID then do; OBSER = 1; end;
retain ID OBSER;
OBSER = OBSER * PRED;
W     = 1 / OBSER;

/* UNWEIGHTED GEE (ASSUMES MCAR) */
proc genmod data=model;
class ID WEEK;
model QUIT/N = TX / link = logit dist = bin;
repeated subject = ID / withinsubject = WEEK type = ind;

/* WEIGHTED GEE (ASSUMES MAR)  */
proc genmod data=weight;
class ID WEEK;
scwgt W;
model QUIT/N = TX / link = logit dist = bin;
repeated subject = ID / withinsubject = WEEK type = ind;
```

APPENDIX B: HIV COHORT EXAMPLE

B.1. Data excerpt

```
ID:       Participant ID number
TIME:     Visit number (6 month intervals)
CD4SQRT:  Square root of CD4 cell count
D:        Drop-out pattern (1, 2, 3)
VLCATB:   Viral load at baseline
SYMPTOMB: Number of HIV-related symptoms at baseline (0 to 5)
ART6B:    Indicator of receiving antiviral treatment in
          6 months prior to enrolment (0 = no, 1 = yes)
AWAREYRB: Number of years aware of HIV infection at baseline
```



```
SUBJECT ID:=1
time    cd4sqrt    d    vlcatb    symptomb    art6b    awareyrb
  0     18.9499    1     >30k         2         0          3
  1     15.3362    1     >30k         2         0          3
  2     14.9097    1     >30k         2         0          3

SUBJECT ID:=2
time    cd4sqrt    d    vlcatb    symptomb    art6b    awareyrb
  0     29.4444    3    500-5k        0         0          1
  1     22.9456    3    500-5k        0         0          1
  2     29.4856    3    500-5k        0         0          1
  3     22.2998    3    500-5k        0         0          1
  4     21.1660    3    500-5k        0         0          1
  5     17.8606    3    500-5k        0         0          1
  6     17.6593    3    500-5k        0         0          1
  7     18.9989    3    500-5k        0         0          1
  8     26.0707    3    500-5k        0         0          1
  9     25.9230    3    500-5k        0         0          1
 10     26.8384    3    500-5k        0         0          1
```

B.2. SAS code

```
/* DATA STORED LOCALLY IN DROP-OUT.SAS7BDAT */
libname sss '';
data dropout; set sss.dropout;

/* REGRESSION MODEL UNDER MCAR ASSUMPTION
   ESTIMATION METHOD EQUIVALENT TO GEE UNDER WORKING INDEPENDENCE
   ROBUST S.E. IMPLEMENTED USING 'EMPIRICAL' OPTION */
proc mixed data=dropout empirical noclprint;
class id  vlcatb;
model cd4sqrt = time symptomb art6b awareyrb vlcatb time*symptomb
                time*art6b time*awareyrb time*vlcatb / s;
repeated / subject=id;

/* REGRESSION MODEL UNDER MAR ASSUMPTION
   ESTIMATION METHOD IS MAXIMUM LIKELIHOOD UNDER M.V. NORMAL MODEL
   VARIANCE MATRIX STRUCTURED USING RANDOM INTERCEPT AND SLOPE */
proc mixed data=dropout empirical noclprint;
class id vlcatb;
model cd4sqrt = time symptomb art6b awareyrb vlcatb time*symptomb
                time*art6b time*awareyrb time*vlcatb / s;
random intercept time / subject=id type=un g gcorr;

/* REGRESSION MODEL USING PATTERN MIXTURE APPROACH (3 CATEGORIES) */
/* STEP 1:  USE PROC MIXED TO ESTIMATE WITHIN-PATTERN REGRESSION
            COEFFICIENTS FROM MODEL OF [Y | D, X] */
```



```
proc mixed data=dropout empirical noclprint;
ods output SolutionF=esti1; ods output CovB=COVB1;
class id vlcatb d;
model cd4sqrt = d time*d   d*vlcatb d*symptomb d*art6b d*awareyrb
                d*time*vlcatb d*time*symptomb
                d*time*art6b d*time*awareyrb / noint s covb;
random intercept time / subject=id type=un;

/* RETAIN REQUIRED COLUMNS OF COV(BETAHAT)
   FROM MODEL OF (Y | D, X) FOR READING INTO PROC IML */
data COVB1; set COVB1; keep Col1-Col48;

/* STEP 2:  USE PROC IML TO CALCULATE MARGINALIZED REGRESSION PARAMETERS
        AND STANDARD ERRORS (USE DELTA METHOD FOR S.E.) */
proc iml;

/* BETAHAT = REGRESSION PARAMS FROM PROC MIXED ABOVE */
use esti1; read all var{Estimate} into betahat;

/* V_BETA = COV(BETAHAT) */
use COVB1; read all var _num_ into V_beta;

/* PIHAT = OBSERVED PROPORTIONS FOR EACH DROP-OUT CATEGORY (USER-SUPPLIED)
   V_PI  = COV(PIHAT) */
pihat = { 199 109 537 }‘ / 845;
V_pi  = ( diag(pihat) - pihat * pihat‘ ) / 845;

/* CONSTRUCT VARIANCE/COVARIANCE MATRIX FOR BETAHAT AND PIHAT */
p = nrow(betahat); q = nrow(pihat); z = shape(0, q, p);
Vhat1 = V_beta || z‘; Vhat2 = z || V_pi; Vhat  = Vhat1 // Vhat2;

/* COMPUTE MARGINAL COVARIATE EFFECTS BY AVERAGING OVER PATTERN.
   FUNCTION I(K) GENERATES IDENTITY MATRIX OF DIMENSION K
   OPERATOR @ IS KRONECKER PRODUCT
   OPERATOR || (//) IS HORIZONTAL (VERTICAL) CONCATENATION
   OPERATOR ## IS ELEMENT-WISE EXPONENTIATION */
Imat    = I( int(p/q) );
e1  = {1 0 0};      e1mat   = Imat @ e1;
e2  = {0 1 0};      e2mat   = Imat @ e2;
e3  = {0 0 1};      e3mat   = Imat @ e3;
beta_m  = ( (pihat[1]*e1mat) + (pihat[2]*e2mat) + (pihat[3]*e3mat) )*betahat;

/* CONSTRUCT JACOBIAN MATRIX FOR DELTA METHOD CALCULATION
   LET THETA = (BETA, PI) WHERE
        BETA = COEFFICIENT VECTOR FROM [Y | D, X] MODEL
    PI   = VECTOR OF DROP-OUT PROBABILITIES */
```



```
dth_db  =  (pihat[1] * e1mat) + (pihat[2] * e2mat) + (pihat[3] * e3mat);
dth_dp1 = e1mat * betahat; dth_dp2 = e2mat*betahat; dth_dp3 = e3mat * betahat;
Jac     = dth_db || dth_dp1 || dth_dp2 || dth_dp3;

/* COMPUTE STANDARD ERRORS FOR MARGINAL COVARIATE EFFECTS */
V_beta_m    = Jac * Vhat * Jac';
se_beta_m   = ( vecdiag( V_beta_m ) )##(0.5);
results         = beta_m || se_beta_m; print results;
```

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to Ralph D'Agostino and the editors of *Statistics in Medicine* for providing the opportunity to write this paper, and for their patience during its development. We are also grateful to several colleagues for their generous assistance and feedback: two anonymous referees provided careful reading and detailed critique, which helped clarify several important points and led to significant improvements of the original manuscript; Mike Daniels, Peter Diggle, Garrett Fitzmaurice, Robin Henderson and Nan Laird read early drafts and provided valuable and detailed comments; Patrick Heagerty provided software to fit the marginalized transition models in Section 6.1, and Don Alderson assisted with implementation; Bess Marcus provided data from the Commit to Quit Study; and Lytt Gardner granted permission to use data from the HER Study. Project funded through NIH grants R01-AI50505 and P30-AI42853 (Lifespan/Tufts/Brown Center for AIDS Research). Data for Commit to Quit was collected under grant R01-CA77249 from the NIH, and data for HERS was collected under CDC cooperative agreements U64-CCU106795, U64-CCU206798, U64-CCU306802 and U64-CCU506831.

REFERENCES

1. Meinert CL. Beyond CONSORT: need for improved reporting standards for clinical trials. *Journal of the American Medical Association* 1998; **279**:1487–1489.
2. Wu MC, Hunsberger S, Zucker D. Testing for differences in changes in the presence of censoring: parametric and non-parametric methods. *Statistics in Medicine* 1994; **13**:635–646.
3. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112–1121.
4. Hogan JW, Laird NM. Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997; **16**:159–272.
5. Kenward MG, Molenberghs G. Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research* 1999; **8**:51–83.
6. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data.* Springer: New York, 2000.
7. Diggle, PJ. Dealing with missing values in longitudinal studies. In *Recent Advances in the Statistical Analysis of Medical Data*, Everitt BS, Dunn G (eds). Arnold: London, 1998; 203–228.
8. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Clarendon Press: Oxford, 1994.
9. Fitzmaurice, GM. Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica* 2003; **57**: 75–99.
10. Hogan JW, Laird NM. Intention to treat analysis for incomplete repeated measures data. *Biometrics* 1996; **52**: 1002–1017.
11. Little RJ, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 1996; **52**: 1324–1333.
12. Fragakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika* 1999; **86**:365–379.
13. Rotnitzky A, Scharfstein DO, Su TL, Robins JM. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* 2001; **57**:103–113.
14. Yau LHY, Little RJ. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 2001; **96**:1232–1244.
15. Hogan JW, Daniels MJ. A hierarchical modelling approach to analysing longitudinal data with dropout and noncompliance, with application to an equivalence trial in paediatric AIDS. *Applied Statistics* 2002; **51**:1–21.
16. Smith DK, Warren DL, Vlahov D, Schuman P, Stein MD, Greenberg BL, Holmberg SD. Design and baseline participant characteristics of the human immunodeficiency virus epidemiology research study (HERS). *American Journal of Epidemiology* 1997; **146**:459–469.

17. Marcus BH, King TK, Albrecht AE, Parisi AF, Abrams DB. Rationale, design, and baseline data for commit to quit: an exercise efficacy trial for smoking cessation in women. *Preventive Medicine* 1997; **26**:586–597.
18. Marcus BH, Albrecht AE, King TK, Parisi AF, Pinto B, Roberts M, Niaura RS, Abrams, DB. The efficacy of exercise as an aid to smoking cessation in women: a randomized controlled trial. *Archives of Internal Medicine* 1999; **159**:1229–1234.
19. Yao Q, Wei LJ, Hogan JW. Analysis of incomplete repeated measurements with dependent censoring times. *Biometrika* 1998; **85**:139–149.
20. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 1996; **91**:444–472.
21. Holland PW. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 1986; **81**:945–970.
22. Robins JM. Association, causation and marginal structural models. *Synthese* 1999; **121**:151–179.
23. Rubin DB. Estimating causal effects in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
24. Goetghebeur EJT, Shapiro SH. Analysing non-compliance in clinical trials: ethical imperative or mission impossible? *Statistics in Medicine* 1996; **15**:2813–2826.
25. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. Wiley: New York, 1987.
26. Lichtenstein E, Glasgow RE. Smoking cessation: what have we learned over the past decade?. *Journal of Consulting and Clinical Psychology* 1992; **60**:518–527.
27. Borrelli B, Hogan JW, Bock B, Pinto B, Roberts M, Marcus B. Predictors of quitting and dropout among women in a clinic-based smoking cessation program. *Psychology of Addictive Behaviors* 2002; **16**:22–27.
28. Mayer KH, Hogan JW, Smith D *et al.* Clinical and immunologic progression in hiv-infected us women before and after the introduction of highly active antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes* 2003; **33**:614–624.
29. Daniels M, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1965–1982.
30. Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 2001; **96**:103–113.
31. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**:465–480.
32. Ibrahim JG. Incomplete data in generalized linear models. *Journal of the American Statistical Association* 1990; **85**:765–769.
33. Little RJA. Regression with missing x's: a review. *Journal of the American Statistical Association* 1992; **87**:1227–1237.
34. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
35. Horton NJ, Laird NM. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* 1999; **8**:37–50.
36. Roy J, Lin X. The analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* 2002; **97**:40–52.
37. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
38. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**: 29–46.
39. Hernán M, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 2001; **96**:440–448.
40. Robins JM, Greenland S, Hu FC. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 1999; **94**:687–700.
41. Hogan JW, Lancaster T. Instrumental variables and marginal structural models for estimating causal effects from longitudinal observational data. *Statistical Methods in Medical Research* 2004; **13**:17–48.
42. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
43. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
44. Fitzmaurice GM, Laird NM, Rotnitzky AG. Regression models for discrete longitudinal responses. *Statistical Science* 1993; **8**:284–299.
45. Fitzmaurice GM, Laird NM. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 1993; **80**:141–151.
46. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* 2002; **58**:342–351.
47. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.

48. Diggle P. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**:959–971.
49. Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* 1992; **11**:1825–1839.
50. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
51. Tsiatis AA, Davidian M. An overview of joint modeling of longitudinal and time-to-event data. *Statistica Sinica* 2003; to appear.
52. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in Medicine* 1994; **13**:955–968.
53. Hogan JW, Laird NM. Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research* 1998; **7**:28–48.
54. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
55. Glynn RJ, Laird NM, Rubin DB. Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association* 1993; **88**:984–993.
56. Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* 1996; **91**:473–520.
57. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
58. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* 1998; **93**:1321–1339.
59. Gill RD, Robins JM. In *Sequential Models for Coarsening and Missingness*, *Proceedings of the First Seattle Symposium on Survival Analysis*, Lin DY, Fleming T (eds). Springer: New York, 1997; 295–305.
60. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
61. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; **39**:1–37.
62. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. Springer: New York, 2003.
63. Wulfsohn MS, Tsiatis A. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**:330–339.
64. DeGruttola V, Tu XM. Modeling progression of cd4 lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003–1014.
65. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine* 1996; **15**:1663–1685.
66. Huang WZ, Zeger SL, Anthony JC, Garrett E. Latent variable model for joint analysis of multiple repeated measurements and bivariate event times. *Journal of American Statistical Association* 2001; **96**:906–914.
67. Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979; **47**:153–161.
68. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* 1998; **17**:2723–2732.
69. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics* 2001; **57**:7–14.
70. Troxel AB, Lipsitz SR, Harrington DP. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* 1994; **94**:1096–1120.
71. Fitzmaurice GM, Laird NM, Zahner GEP. Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association* 1996; **91**:99–108.
72. Molenberghs G, Kenward MG, Lesaffre E. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* 1997; **84**:33–44.
73. Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal data with informative dropouts. *Journal of the Royal Statistical Society, Series B* 1995; **57**:691–704.
74. Conaway MR. The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association* 1992; **87**:817–824.
75. Oswald software, David M. Smith and Lancaster University Department of Mathematics and Statistics.
76. Bugs software, Imperial College School of Medicine at St. Mary‘s, London.
77. Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* 1977; **72**:538–543.
78. Wu MC, Bailey KR. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* 1988; **7**:337–346.
79. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989; **45**:939–955.
80. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; **81**:471–483.
81. Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**:125–134.
82. Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 1996; **52**:98–111.

83. Daniels MJ, Hogan JW. Reparameterization of the pattern mixture model for sensitivity analysis under informative dropout. *Biometrics* 2000; **66**:1241–1248.
84. Molenberghs G, Michiels B, Kenward MG, Diggle PJ. Monotone missing data and pattern-mixture models. *Statistica Neerlandica* 1998; **52**:153–161.
85. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **16**:239–257.
86. Fitzmaurice GM, Laird NM. Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* 2000; **1**:141–156.
87. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; **51**:151–168.
88. Albert PS, Follmann DA. Modeling repeated count data subject to informative dropout. *Biometrics* 2000; **56**: 667–677.
89. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1987; **44**:175–188.
90. Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV-positive patients. *Journal of the American Statistical Association* 1998; **93**:46–53.
91. Schluchter MD. Methods for the analysis of informatively censored longitudinal data (Disc: P1881–1885). *Statistics in Medicine* 1992; **11**:1861–1870.
92. Mori M, Woodworth G, Woolson RF. Application of empirical bayes inference to estimation of rate of change in the presence of informative right censoring. *Statistics in Medicine* 1992; **11**:621–631.
93. Pulkstenis EP, Ten Have TR, Landis JR. Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication. *Journal of the American Statistical Association* 1998; **93**:438–450.
94. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 1999; **94**:1096–1120.
95. Glidden DV, Wei LJ. Rank estimation of treatment differences based on repeated measurements subject to dependent censoring. *Journal of the American Statistical Association* 1999; **94**:888–895.
96. Lipsitz SR, Fitzmaurice GM, Molenberghs G, Zhao LP. Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Applied Statistics* 1997; **46**:463–476.
97. Technical Report, Center for Statistical Sciences, Brown University. *Controlled Clinical Trials*, in revision.
98. Lichtenstein E, Glasgow RE. Smoking cessation: what have we learned over the past decade? *Journal of Consulting and Clinical Psychology* 1992; **60**:518–527.
99. Chen HY, Little R. A test of missing completely at random for generalised estimating equations with missing data. *Biometrika* 1999; **86**:1–13.
100. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics, Part A—Theory and Methods* 1980; **9**:1043–1069.
101. Cook RD, Weisberg S. *An Introduction to Regression Graphics*. Wiley, New York, 1994.
102. Ko H, Hogan JW, Mayer KH. Estimating causal treatment effects from longitudinal HIV natural history studies using marginal structural models. *Biometrics* 2003; **59**:152–162.
103. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1991; **1**:54–75.
104. Stiratelli R, Laird NM, Ware JH. Random-effects models for serial obervations with binary response. *Biometrics* 1984; **40**:961–971.
105. Hedeker D, Gibbons RD. A random effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
106. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
107. Fitzmaurice GM, Laird NM, Lipsitz SR. Analysing incomplete longitudinal binary responses: a likelihood-based approach. *Biometrics* 1994; **50**:601–612.
108. Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. *Biostatistics* 2001; **3**:245–265.
109. Copas JB, Li HG. Inference for non-random samples. *Journal of the Royal Statistical Society, Series B* 1997; **59**:55–77.
110. Rubin DB, Frangakis CE. Comment on estimation of the causal effect of a time varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 1999; **94**: 702–704.
111. Robins JM, Greenland S, Hu FC. Reply to comments on estimation of the causal effect of a time varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 1999; **94**:708–712.

