Published in final edited form as:

Stat Med. 2012 July 20; 31(16): . doi:10.1002/sim.4510.

Longitudinal Structural Mixed Models for the Analysis of Surgical Trials with Noncompliance

Colleen M. Sitlani^{a,b,*}, Patrick J. Heagerty^a, Emily A. Blood^c, and Tor D. Tosteson^d

^aDepartment of Biostatistics, University of Washington, F-600 Health Sciences Building, Box 357232, Seattle, WA 98195

^bCardiovascular Health Research Unit, University of Washington, 1730 Minor Ave, Suite 1360, Box 358085, Seattle, WA 98101

^cClinical Research Program, Children's Hospital Boston and Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115

^dSection of Biostatistics and Epidemiology, Dartmouth Medical School, One Medical Center Drive, 7927 Rubin Building, Lebanon, NH 03756

Abstract

Patient noncompliance complicates the analysis of many randomized trials seeking to evaluate the effect of surgical intervention as compared to a non-surgical treatment. If selection for treatment depends on intermediate patient characteristics or outcomes, then “as-treated” analyses may be biased for the estimation of causal effects. Therefore, the selection mechanism for treatment and/or compliance should be carefully considered when conducting analysis of surgical trials. We compare the performance of alternative methods when endogenous processes lead to patient crossover. We adopt an underlying longitudinal structural mixed model that is a natural example of a structural nested model. Likelihood-based methods are not typically used in this context; however, we show that standard linear mixed models will be valid under selection mechanisms that depend only on past covariate and outcome history. If there are underlying patient characteristics that influence selection, then likelihood methods can be extended via maximization of the joint likelihood of exposure and outcomes. Semi-parametric causal estimation methods such as marginal structural models, g-estimation and instrumental variable approaches can also be valid, and we both review and evaluate their implementation in this setting. The assumptions required for valid estimation vary across approaches; thus the choice of methods for analysis should be driven by which outcome and selection assumptions are plausible.

Keywords

longitudinal data; noncompliance; causal inference; endogeneity; structural nested models

1. Introduction

Randomized trials with the goal of evaluating the long-term benefit of a surgical intervention as compared to non-surgical treatment are often faced with serious patient noncompliance. Frequently subjects assigned to surgery delay or subsequently refuse surgery, while non-surgical subjects may ultimately seek and receive surgery. For example,

in the Spine Patient Outcomes Research Trial (SPORT) evaluation of surgery for lumbar intervertebral disk herniation, 60% of those randomized to surgery actually received it, while 45% of those randomized to nonoperative treatment subsequently received surgery [1]. Standard intent-to-treat (ITT) analyses do not capture the full efficacy of surgery in the presence of such noncompliance [2, 3]. In an effort to estimate the efficacy of treatment, investigators often supplement ITT analyses with “as-treated” analyses that incorporate subjects’ actual treatment received instead of their assigned treatment. There are several statistical challenges associated with longitudinal ‘as-treated’ analyses which seek to estimate average causal effects (ACEs) attributable to surgery [2, 3].

The goal of this manuscript is to overview both statistical models and estimation methods that may be appropriate for the analysis of surgical non-compliance. We first anchor the statistical objective by defining the causal effect of interest based on a structural model for the longitudinal outcomes. Second, we describe possible selection models which characterize the baseline and time-dependent factors that influence the timing of treatment. Together, the structural longitudinal model and the selection model determine the joint distribution of the observed longitudinal treatment and outcome data. Finally, given a causal target parameter and a framework for characterizing the full longitudinal data process we consider both parametric and semi-parametric methods of estimation and delineate the assumptions required for their validity.

The biostatistics literature contains substantial development of semi-parametric methods, but to our knowledge no comparison has been made with valid parametric options. Specifically, in section 2.1 we describe a distribution for potential outcomes using an underlying longitudinal structural mixed model which then allows identification of the surgical ACE of interest. We show that the proposed mean model is a natural example of a structural nested model (SNM), as introduced by Robins [4]. In Section 2.2 we discuss possible selection mechanisms which can characterize the likelihood that patients deviate from their treatment assignment. In Section 3 we discuss options for estimating the ACE when endogenous processes lead to patient crossover. Analysis approaches include standard methods such as linear mixed models [5] and generalized estimating equations [6, 7], as well as tailored causal estimation methods such as marginal structural models [8], instrumental variable methods [9, 10], and g-estimation [4]. We compare the performance of alternative methods both in terms of bias and efficiency using a variety of simulation scenarios in Section 4. We use these methods in Section 5 to illustrate analysis of data collected in the SPORT trial, in which crossover occurred for a large fraction of subjects. In Section 6 we discuss the implications of our findings.

1.1. Motivating Example: SPORT Trial

The Spine Patient Outcomes Research Trial (SPORT) is a randomized clinical trial comparing surgical and nonoperative treatment for several different back conditions [1, 11, 12, 13]. One focus of this study is the treatment for lumbar intervertebral disk herniation. The primary outcome measures are changes from baseline for the Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) bodily pain and physical function scales and the modified Oswestry Disability Index. The initial reports for SPORT focused on outcomes that were measured at 6 weeks, 3 months, 6 months, 1 year, and 2 years from enrollment.

Between March 2000 and November 2004, a total of 501 patients were randomized to either standard open lumbar discectomy or nonoperative treatment. Noncompliance with assigned treatment was high: at three months post-enrollment, only 50% of patients assigned to surgery had received surgery, and 30% of patients assigned to nonoperative treatment had received surgery. By two years, these numbers had increased to 60% and 45%, respectively.

The pre-specified intent-to-treat analyses did not yield strong evidence for a treatment benefit, at least in part due to the large amount of crossover that occurred. To explore the impact of this crossover, and to account for the time of surgery, additional comparisons were made between those who actually received surgery and those who didn't, despite their randomized treatment assignment. Because the observed treatment groups were not defined entirely by randomization, methods of controlling for confounding in observational studies were considered, including adjustment for longitudinally evolving individual health status outcomes. The scientific goal was to evaluate the impact of treatment on key outcomes, but because many patients did not comply with their assigned treatment, investigation of the factors that lead to the decision to have surgery is essential in order to obtain valid inference. It is likely that recent past measures of pain or disability influenced individual treatment timing decisions. The "as-treated" analyses used in the SPORT papers adjusted for past outcomes in order to compensate for this possible selection bias. The objectives of this paper are to discuss additional methods that might be used to account for such treatment endogeneity and to evaluate their relative merits.

2. Statistical Models for Longitudinal Outcomes and Time-dependent Exposures

To consider estimation approaches for longitudinal data subject to non-compliance, we first outline a class of causal models for the longitudinal outcomes, and then present a hierarchy of assumptions for factors that influence changes in exposure over time. By combining structural mean models with standard random effects assumptions, the longitudinal outcome model allows identification of the causal treatment comparisons of interest and provides a complete likelihood for the observed data.

2.1. Outcomes: Structural Models for Longitudinal Data

Often the primary goal of medical research is to determine whether one or more treatments cause improvement in a medical condition. One way to define a causal effect of treatment is via potential outcomes [14]. That is, if the outcome of interest is Y , then one can define a potential outcome $Y_i(x_k)$ for each subject i under each possible treatment x_k . Then, for subject i , a causal effect comparing treatment x_k to the referent treatment x_0 is $Y_i(x_k) - Y_i(x_0)$. Unfortunately this causal effect cannot in general be directly observed because each subject receives only one treatment. However, it is well known that the ACE can be estimated by randomizing subjects to the treatment(s) of interest or to a referent group, and then comparing the average outcome in each treatment group to the average outcome in the referent group. That is, one can calculate $E[Y_i(x_k)] - E[Y_j(x_0)]$ where subjects $i=1, \dots, I$ receive treatment x_k and subjects $j=1, \dots, J$ receive the referent treatment x_0 .

With longitudinal data, definition of causal effects needs to consider the repeated exposure and/or outcome measurements that can be observed. If response measurements are made only at one pre-defined timepoint, then the outcome of interest will be defined as before, but treatment may become a vector of treatments over defined intervals prior to the outcome measurement, and one must consider the entire treatment path $\mathbf{x}_k$ that represents the vector of treatment indicators at each time t where $t = 1, 2, \dots, T$: $\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kT})$. The ACE can then be estimated via $E[Y_i(\mathbf{x}_k)] - E[Y_j(\mathbf{x}_0)]$ where subjects $i=1, \dots, I$ follow treatment path $\mathbf{x}_k$ and subjects $j=1, \dots, J$ follow the referent treatment path $\mathbf{x}_0$.

When outcome data is also collected at several different points in time, each subject has a vector of potential outcomes, one at each observation time t , under every possible treatment path: $\mathbf{Y}_i(\mathbf{x}_k) = [Y_{i1}(\mathbf{x}_k), Y_{i2}(\mathbf{x}_k), \dots, Y_{iT}(\mathbf{x}_k)]$. In this manuscript we restrict discussion to "point treatments" that occur at a single time s and where we assume subjects remain in the

treated group at all times after s since this reflects the structure of a surgical intervention. Note that s can be any time during the observation period. In this situation the vector of potential outcomes can be denoted $\mathbf{Y}_i(s)$ and the individual causal effect at time t is $Y_{it}(s=0) - Y_{it}(s=\infty)$ which contrasts the outcome at time t when a subject has surgery immediately after baseline ($s=0$) to the outcome when surgery is withheld ($s=\infty$, or alternatively $s>T$). The corresponding ACE is $E[Y_{it}(s=0) - Y_{it}(s=\infty)] = \Delta_s$.

Characterization of individual and average causal effects requires an individual-level, time-specific causal model that compares a person's potential outcomes under the two treatment paths of interest at any given time. We propose a longitudinal structural mixed model (LSMM) that contains three components: a group average, a subject average, and individual observations. This model is "structural" because its coefficients represent causal effects based on potential outcomes [14], and it is "mixed" because it incorporates both fixed and random effects to account for the correlation between repeated (potential) measurements on the same subject. In the context of a monotone treatment such as surgery, the LSMM can be written as follows:

$$Y_{it}(s) = \mathbf{X}_i(t, s)' \beta \quad \text{group average} \\ + \mathbf{Z}_i(t, s)' \mathbf{b}_i \quad \text{subjects within a group} \\ + e_{it}(s) \quad \text{observations within a subject}$$

In this context, t is the time of measurement and s is the time of surgery. The group average component is typically the one of interest, and the vector $\mathbf{X}_i(t, s)$ is often separated into a baseline component $\mathbf{X}_i^0(t)$ and a time-dependent exposure $\mathbf{X}_i^{TX}(t, s)$. The group average can then be written as follows:

$$\mathbf{X}_i(t, s)' \beta = \mathbf{X}_i^0(t)' \lambda + \mathbf{X}_i^{TX}(t, s)' \Delta(t, s).$$

The first term represents the average outcome over time when no treatment occurs and can be written in terms of a time function basis $\mathbf{B}(t)$ so that

$\mathbf{X}_i^0(t)' \lambda = \lambda_0 + \lambda_1 B_1(t) + \dots + \lambda_p B_p(t)$. The second term represents the change in outcome attributable to surgery, which may depend on the time at which surgery occurs, e.g.

$\mathbf{X}_i^{TX}(t, s) = \mathbb{1}_{[t \geq s]}$ and $\Delta(s, t) = \gamma_0 + \gamma_1 \cdot (t - s)$.

The subject-specific latent effects, represented by $\mathbf{b}_i$, account for the correlation between measurements on the same individual by assuming that each subject follows their own trajectory. The design matrix for the random effects is $\mathbf{Z}_i$. Random effects are assumed to be independent of the measurement errors $e_{it}(s)$. A rank-preserving model is one in which the rank of individuals' outcomes are preserved across all possible treatment paths [15]. For example, if subject 1 will have a higher outcome than subject 2 in the absence of treatment, then by a rank-preserving assumption he will also have a higher outcome in the presence of treatment. If $e_{it}(s)$ does not depend on the surgery time and treatment effects are assumed to be homogeneous, then the LSMM is rank-preserving. A non-rank-preserving model, which is often considered to be more realistic, can be constructed either by allowing $e_{it}(s)$ to differ before and after surgery occurs or by allowing treatment effects to be heterogeneous via a random effect.

For each subject, at a specified time, the LSMM permits characterization of the potential outcomes both for his observed treatment time and for the hypothetical treatment times that

we want to compare, e.g. $s = 0$ and $s = \infty$. One example of a non-rank-preserving LSMM is the following:

$$\begin{aligned} \mathbf{X}_i^0(t)' \lambda &= \lambda_0 + \lambda_1 \cdot t \\ \mathbf{X}_i^{TX}(t, s) &= \mathbb{1}_{[t \geq s]} \\ \Delta(t, s) &= \gamma_0 + \gamma_1 \cdot (t - s) \\ \mathbf{Z}_i(t, s)' \mathbf{b}_i &= b_{i0} + b_{i1} \cdot t + b_{i2} \cdot \mathbf{X}_i^{TX}(t, s) \\ e_{it}(s) &= e_{it}^0 \cdot \mathbb{1}_{[t \leq s]} + e_{it}^1 \cdot \mathbb{1}_{[t > s]} \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}) \\ (\mathbf{e}_i^0, \mathbf{e}_i^1) &\text{ iid } N(0, \sigma^2) \end{aligned}$$

Figure 1 displays the potential outcomes defined by this model, both at the group average and subject-specific levels. On average, a subject follows the trajectory of the untreated group until he receives treatment, at which point there is an immediate effect of treatment as well as increased improvement over time. The outcome trajectory clearly depends on the treatment path, i.e. whether or not treatment occurs, and if so, then when it occurs. The average causal effect at any given time is the difference, at that time, between the trajectory corresponding to treatment just after enrollment and the one corresponding to no treatment, i.e. $E[Y_i(s=0)] - E[Y_i(s=\infty)] = \gamma_0 + \gamma_1 \cdot t$. Therefore this model provides a framework for identifying treatment effects of interest.

An alternate way to formulate the model is as a structural nested model (SNM) [4]. A SNM models the effect on the outcome caused by one increment of treatment, as a function of treatment history, covariate history (time-invariant and/or time-varying), outcome history, and time between treatment increment and outcome measurement. That is, it compares the conditional outcome at a time k for subjects on the treatment path $(x_1, \dots, x_{m-1}, \underline{x}_m, 0, \dots, 0)$ to that for subjects on the treatment path $(x_1, \dots, x_{m-1}, \underline{0}, 0, \dots, 0)$ [4]. The effect of the treatment at time m that one subject receives and the other doesn't is often referred to as a "blip" in the literature. When the SNM is specified for conditional *mean* outcomes, it is referred to as a structural nested mean model (SNMM); when it is specified for conditional outcome *distributions*, it is referred to as a structural nested distribution model (SNDM) [16].

When treatment is monotone and binary, as for surgery, this model provides a convenient context for characterizing the causal effect. Up until the time of surgery, all subjects will have the same treatment history (no surgery), and outside of the one time period in which surgery occurs, treated subjects will have no additional surgery. Therefore, the average causal effect of interest, i.e. the average difference in outcome between people who never undergo surgery $[\mathbf{x}_k = (0, \dots, 0, 0, 0, \dots, 0)]$ and people who undergo surgery at a specified time $[\mathbf{x}_k = (0, \dots, 0, 1, 0, \dots, 0)]$ is precisely the "blip" of a SNM. Therefore, the structural model introduced here is a special case of a SNDM due to the simple nature of the exposure under consideration. If we had specified the structural model only at the group average level, then this would be a SNMM. To see the connection between LSMs and SNMMs, consider Robin's Example 1 [4], in which the "blip" is modeled as a function of exposure at time m (a_m), assigned treatment arm (r), a time-varying covariate (w_m), and the times of outcome measurement and incremental exposure (t_k and t_m). If this model is simplified to be only a function of a_m , t_k , and t_m so that the model is $\psi_{0,1}a_m + \psi_{0,6}a_m(t_k - t_m)$, then this is equivalent to the structural group-average model proposed here.

2.2. Exposure: Statistical Models for Endogenous Treatment Processes

For longitudinal data with noncompliance, the structural model alone is not sufficient for estimation of average causal effects. The exposure process, i.e. characterization of who receives treatment and when they receive it, must also be considered. If treatment is determined solely by factors that are separate from the longitudinal outcomes, then the treatment process is said to be exogenous. Specifically, given all previous exposures and time-invariant covariates, the exposure at time t is exogenous if it is independent of all preceding outcome measurements [17]. One example of such an exogenous exposure is treatment in a randomized trial with perfect compliance. For example, in a surgery trial, surgery is exogenous if the subjects assigned to control never undergo surgery and the subjects assigned to surgery undergo surgery immediately after enrollment. In this case, the timing of surgery would only occur at enrollment so that outcomes after enrollment would have no bearing on treatment received.

However, in the case of an unblinded trial with crossover between treatment arms, the treatment received is often influenced by patient factors prior to randomization or, in the case of longitudinal trials, by interim patient factors including interim outcomes. That is, if a patient is assigned to the control group, but has poor outcomes as measured prior to treatment, then he may undergo surgery after these measurements despite having been assigned to non-surgical therapy. Any such treatment process that does not meet the condition for exogeneity is said to be endogenous.

We will consider two different types of selection that cause endogeneity: *direct* and *indirect*. To define these terms, we will assume the following model for p_{it} = the probability of subject i being treated in the interval $(t-1, t]$: $\text{logit}(p_{it}) = \alpha(t, \mathbf{L}_i^0) + \eta(\bar{\mathbf{L}}_{i,t-1}) + \delta(\mathbf{U}_i)$ where $\mathbf{L}_i^0$ is a vector of baseline covariates, $\bar{\mathbf{L}}_{i,t-1}$ is a vector of time-varying covariates observed through time $t-1$, and $\mathbf{U}_i$ is a vector of unobserved covariates. The treatment process is exogenous when both $\eta(\bar{\mathbf{L}}_{i,t-1}) = 0$ and $\delta(\mathbf{U}_i) = 0$; it is endogenous when one or both of these functions is nonzero, as in the following definitions of direct and indirect selection.

Direct selection occurs when covariates observed prior to a time of interest directly determine whether or not a subject receives treatment at that time, i.e. $\eta(\bar{\mathbf{L}}_{i,t-1}) \neq 0$ and $\delta(\mathbf{U}_i) = 0$. An example of direct selection is when a subject undergoes surgery as a result of having had a poor outcome at earlier times and then actively seeking out a change in treatment.

Indirect selection occurs when unobserved factors drive treatment choice, i.e. $\delta(\mathbf{U}_i) \neq 0$. One indirect selection mechanism is when an underlying, unmeasured patient characteristic impacts both treatment status and outcome. For example, rather than a subject undergoing surgery because his last outcome was poor, he may undergo surgery because his physician thinks that he is likely to benefit from it, based on unmeasured aspects of available clinical information such as imaging data. Choosing patients selectively based on their propensity to benefit, which can be thought of as an unmeasured confounder, is a form of selection bias since not all of the target subjects are treated. If we could measure this latent propensity for benefit and condition on it, then treatment and outcome would be conditionally independent of each other.

In our development we allow selection to include both direct and indirect selection by using the following model: $\text{logit}(p_{it}) = \alpha(t, R_i) + \eta(\bar{\mathbf{Y}}_{i,t-1}^O) + \delta(\mathbf{b}_i)$, where R_i is 0 if subject i is assigned to the control group and 1 if subject i is assigned to the surgery group. The vector $\bar{\mathbf{Y}}_{i,t-1}^O$ includes all outcomes for subject i that are observed through time $t-1$. The random effects $\mathbf{b}_i$ are assumed to capture the unmeasured covariates $\mathbf{U}_i$ that contribute to selection to

treatment. Simpler selection models can be constructed by setting one or more of the terms in this model equal to zero.

The observed data likelihood for longitudinal data with an endogenous exposure process must incorporate both this selection model and the LSMM specified in the previous section. In order to validly estimate the average causal effect when the exposure is endogenous, assumptions must be made about the underlying selection model. Analysis options and their corresponding assumptions will be described in the next section.

3. Estimation: Methods of Inference with Time-dependent Covariates

Standard estimation methods used to analyze longitudinal data will not necessarily be valid in the presence of endogenous treatment exposure [17]. In this section we discuss both the assumptions needed for validity of standard estimators, as well as alternate estimators that relax key assumptions. Alternate estimators include maximization of more complex joint likelihoods, as well as methods tailored to estimate causal effects in the presence of endogeneity, such as marginal structural models, instrumental variable estimators, and g-estimators. All methods rely on the stable unit treatment value assumption (SUTVA), which implies that the potential outcomes for one individual are unrelated to the treatment status of other individuals and that the observed outcome equals a potential outcome [18]. In the subsections that follow we overview the general approach of each candidate method and emphasize key assumptions that are required for valid application.

3.1. Likelihood-Based Methods

Our goal is to estimate the average causal effects of treatment over time, represented for example by $\gamma_0 + \gamma_1 \cdot t$ in the simple LSMM presented in section 2.1. If the treatment process were exogenous, then estimation of the average causal effect could proceed through standard evaluation of the difference between the average outcome in the subjects treated at time 0 and the average outcome in the subjects never treated, with the time-dependent treatment process being considered ancillary. Given the absence of selection bias, the major statistical consideration is to properly account for within-subject correlation of outcomes using standard tools such as the likelihood-based linear mixed effect (LME) model or generalized estimating equations (GEE).

LME and GEE estimators are not necessarily valid when exposure is endogenous [17], so they are not typically considered for use in as-treated analyses. We will show that there are reasonable conditions under which LME estimators are valid, and that even when they are not valid, a modified likelihood-based approach is. Both conditions for the use of LME estimators and the modified likelihood approach will be illustrated through consideration of maximization of the joint likelihood of observed outcomes and treatment under explicit assumptions about 1) selection for treatment and 2) the underlying structure of treatment effects. Similar methods have been used to model outcomes and informative dropout [19, 20, 21], but to our knowledge such methods have not been implemented in the context of noncompliance.

General Approach—To express the joint model mathematically, we need notation for the observed outcomes, as opposed to the counterfactual outcomes that we have considered thus far. Let each observed outcome be Y_{it}^O , let X_{it}^{TX} be an indicator of whether subject i has received treatment by time t , and let random effects remain denoted as $\mathbf{b}_i$. The vectors of observations for each subject will be denoted $\mathbf{Y}_i^O$ and $\mathbf{X}_i^{TX}$, and the subsets of these vectors that represent histories through time t will be denoted $\bar{\mathbf{Y}}_{it}^O$ and $\bar{\mathbf{X}}_{it}^{TX}$. In general we want to

maximize the joint likelihood $\prod_i [\mathbf{Y}_i^O, \mathbf{X}_i^{TX}]$. Using a telescoping factorization over time, the likelihood can be rewritten as:

$$\prod_i \int [\mathbf{Y}_i^O, \mathbf{X}_i^{TX} | \mathbf{b}_i] dF(\mathbf{b}_i) = \prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}, \mathbf{b}_i] [X_{it}^{TX} | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}, \mathbf{b}_i] dF(\mathbf{b}_i). \quad (1)$$

If we assume conditional independence of the outcomes, given treatment and random effects, then the joint likelihood becomes:

$$= \prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i] [X_{it}^{TX} | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}, \mathbf{b}_i] dF(\mathbf{b}_i). \quad (2)$$

This likelihood is the most general formulation of the joint model that we will consider in this manuscript. The underlying causal assumptions appear in the first part of the likelihood $[Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i]$, while the selection model appears in the second part $[X_{it}^{TX} | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}, \mathbf{b}_i]$. Given assumptions about each of these models, the likelihood can be maximized to estimate the target parameters in the first part, which will represent average causal treatment effects.

The general joint likelihood differs from the likelihood maximized by standard LME methods because it allows the selection model to include shared random effects. If additionally we assume that treatment depends only on previous outcomes and not on random effects, then we can simplify the selection model and obtain:

$$= \prod_i \int \prod_t [Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i] [X_{it}^{TX} | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}] dF(\mathbf{b}_i). \quad (3)$$

Now the simplified joint likelihood can be factored to allow separate consideration of the structural model and the selection model:

$$= \prod_i \left\{ \left[\prod_t [X_{it}^{TX} | \bar{\mathbf{Y}}_{it-1}^O, \bar{\mathbf{X}}_{it-1}^{TX}] \right] \left[\prod_t [Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i] dF(\mathbf{b}_i) \right] \right\}. \quad (4)$$

Factorization permits estimation of structural parameters by separately maximizing

$\prod_i [\prod_t \int [Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i] dF(\mathbf{b}_i)]$. LME models are fitted via maximization of $\prod_i \int [\mathbf{Y}_i^O | \mathbf{X}_i^{TX}, \mathbf{b}_i] dF(\mathbf{b}_i)$. Therefore LME models will give consistent estimates of treatment effect when the joint likelihood can be rewritten as in (4) because this likelihood takes the same form as the LME likelihood.

Key Assumptions—LME estimators are obtained by maximizing the likelihood of the structural model, and therefore rely on the specification of probability distributions for both the random effects and the measurement errors for validity [5, 22]. If these distributions are not correctly specified, then LME estimators can be biased. In the presence of an endogenous treatment process, LME estimators, which do not explicitly incorporate a selection model, can be biased unless selection is based only on previous outcomes and not on shared latent effects. If selection does not depend on an indirect latent effect, then the

LME estimator provides a consistent estimate of treatment effects provided that the random effect structure is correctly specified and that there is no serial correlation among outcomes.

GEE estimates are quasi-likelihood-based in that they only require specification of the mean and variance, not the entire probability distribution [6, 23]. In fact, consistency can be obtained even with incorrect specification of the correlation structure in the variance [6, 23]. In most cases, GEE estimates do not benefit from the factorization of the joint likelihood that enables consistency of LME estimates when selection depends only on measured history. With an endogenous treatment process, GEE estimates will only be consistent in the special case where random intercepts accurately capture the within-subject dependence structure and GEE is implemented via an exchangeable working covariance matrix. In this case GEE estimation is asymptotically equivalent to an LME estimator that assumes random intercepts.

When there is an indirect latent effect and/or there is serial correlation between outcomes, LME and GEE estimates will both be biased. Such indirect selection is comparable to the nonignorable random-coefficient-based dropout described by Little in the context of missing data [24].

When indirect selection exists, the joint likelihood that explicitly incorporates such selection should be the focus of likelihood-based estimation. Direct maximization of this joint likelihood will give consistent estimates, provided dependence on $\mathbf{b}_i$ for selection is correctly specified. An analytic solution to this maximization problem may be possible, but would require specialized numerical methods. A simple alternative to pure likelihood analysis would be to adopt Bayesian analysis that explicitly incorporates the selection model. For example, the posterior distribution corresponding to the likelihood in equation (2) is proportional to

$$\prod_i \prod_t [Y_{it}^O | \bar{\mathbf{X}}_{it}^{TX}, \mathbf{b}_i; \beta, \delta, \sigma^2] [X_{it}^{TX} | \bar{\mathbf{Y}}_{i,t-1}^O, \bar{\mathbf{X}}_{i,t-1}^{TX}, \mathbf{b}_i; \alpha] [\mathbf{b}_i; D] \cdot \pi(\beta, \delta, \sigma^2, \alpha, D).$$

Assuming independence between components of the prior distribution, it can be decomposed into $\pi(\boldsymbol{\beta}, \boldsymbol{\delta}) \cdot \pi(\sigma^2) \cdot \pi(\boldsymbol{\alpha}) \cdot \pi(D)$. Bayesian estimation can then be implemented in freely-available software, such as WinBUGS [25]. Sample code illustrating prior and likelihood specification is provided in web-based supporting materials.

As shown in equation (4), LME estimates are consistent when the joint likelihood can be factored into a structural component and a selection component. However, this factorization relies on the assumption that the selection model includes only variables that are also included in the structural model, such as past outcomes and shared latent effects. However, there are scenarios where additional time-dependent patient characteristics contribute to treatment decisions, yet the additional variables may be intermediate outcomes and therefore should not be included as covariates in the structural model. For example, in the SPORT example, when the outcome of interest is patient function as measured by the SF-36 physical function subscale, a subjects' pain score may contribute to treatment decisions and may be correlated with function measures. Including lagged pain scores as covariates in $\mathbf{X}_i^0(t)$ would likely attenuate the estimated marginal effect of treatment on physical function because controlling for previous changes in pain would effectively control for intermediate outcomes. Therefore the requirement of using the same covariates in both models is a disadvantage to LME estimates. On the other hand, maximization of the joint likelihood

allows separate specification of the selection and structural models, providing increased flexibility.

3.2. Marginal Structural Models

As discussed in Section 3.1, methods that can more flexibly account for the various patient and/or provider factors that influence treatment timing are an important generalization for non-compliance analysis. LME methods do not provide this flexibility; joint likelihood methods do, but they require stronger assumptions than marginal structural models (MSMs), which provide this flexibility under marginal modeling assumptions instead of joint ones.

General Approach—MSMs require the specification of a separate selection model that can include any covariates that are predictive of treatment. For example, with a time-varying dichotomous treatment, the selection model could be [8]

$$\text{logit } P[X_t^{TX}=1 | \bar{\mathbf{X}}_{t-1}^{TX}=\bar{\mathbf{x}}_{t-1}, \bar{\mathbf{L}}_t=\bar{\mathbf{l}}_t] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x_{t-1} + \alpha_3 \cdot l_t + \alpha_4 \cdot x_{t-1} \cdot l_t$$

where $\mathbf{L}$ is a vector of all covariates that predict longitudinal treatment variables, which may or may not include previous outcome measurements $\bar{\mathbf{Y}}_{t-1}^O$. The selection model could also be a more complicated function of covariate history, but whatever model is deemed to be appropriate can be used to calculate the probability that each subject received his/her own treatment, conditional on past treatment and observed covariates, i.e.

$$p_{it} = P[X_t^{TX}=1 | \bar{\mathbf{X}}_{t-1}^{TX}=\mathbf{0}, \bar{\mathbf{L}}_t=\bar{\mathbf{l}}_t] \text{ or } (1-p_{it}) = P[X_t^{TX}=0 | \bar{\mathbf{X}}_{t-1}^{TX}=\mathbf{0}, \bar{\mathbf{L}}_t=\bar{\mathbf{l}}_t].$$

Weighting the data by the inverse of these estimated probabilities creates a pseudopopulation in which the average causal effect is no longer confounded. However, the weights can be highly variable, making the estimate of the causal effect highly imprecise, so a stabilized version of the weights is recommended [8]. In addition to the probabilities p_{it} , stabilized weights (sw_{it}) also require a model for the probability that each subject received his/her own treatment, conditional only on past treatment and baseline covariates, i.e.

$$p_{it}^* = P[X_t^{TX}=1 | \bar{\mathbf{X}}_{t-1}^{TX}=\mathbf{0}] \text{ or } (1-p_{it}^*) = P[X_t^{TX}=0 | \bar{\mathbf{X}}_{t-1}^{TX}=\mathbf{0}].$$

The probabilities p_{it} and p_{it}^* are used to calculate the weights:

$$sw_{it} = \frac{\prod_{k=0}^t (p_{ik}^*)^{x_{ik}} (1-p_{ik}^*)^{(1-x_{ik})}}{\prod_{k=0}^t (p_{ik})^{x_{ik}} (1-p_{ik})^{(1-x_{ik})}}.$$

Once the weights have been specified, then a weighted GEE analysis of the structural model discussed in Section 2.1, using independent working covariance, will provide consistent estimates of the causal parameters of interest [8].

Key Assumptions—The selection probabilities are generally not known, and are therefore estimated via a logistic regression model that incorporates prior treatment history

($\bar{\mathbf{X}}_{t-1}^{TX}$) and factors influencing exposure ($\mathbf{L}$), such as one or more previous outcomes, baseline covariates, and time. Fitted estimates from the logistic selection model conditional on $\bar{\mathbf{X}}_{t-1}^{TX}$ and $\mathbf{L}$, as well as estimates from one conditional only on past exposure $\bar{\mathbf{X}}_{t-1}^{TX}$, contribute to estimation of the stabilized weights $\widehat{sw}_{it}$. In order for the MSM analysis to be consistent, two conditions must be met by these selection models: 1) there can be no unmeasured confounders, i.e. $\mathbf{L}$ must contain all characteristics that influence both future outcomes and future treatment, and 2) the form of the selection models must be correctly specified [26]. Additionally, the form of the structural model must be correctly specified [26]. Similar to LME estimators, MSM estimators will be invalid when latent effects impact selection so that current and/or future outcomes are associated with current treatment.

An experimental treatment assignment (ETA) assumption is also required. That is, for all possible covariate combinations, all realizations of the treatment must be possible at all times [27]. This would be violated if a subset of subjects with a specific covariate profile always received treatment or never received it. Violation of ETA can lead to biased estimation [27]; however, this assumption is testable and can be accommodated via study design, e.g. by restricting study populations. Assumptions implicit in the use of MSMs, and particularly in the specification of the weight model, are discussed further by Cole et al [28]. Brumback et al explore the consequences of unmeasured confounding [29]. Sensitivity analyses can be implemented to explore the impact of modeling assumptions [28, 30, 31].

3.3. G-estimation and Instrumental Variables

We consider a simple form of g-estimation that relies on initial randomization [4], instead of the doubly-robust g-estimation that relies on sequential randomization in observational data [32, 33]. Both initial-randomization-based-g-estimation and instrumental variables (IV) estimators rely on estimating equations of the form:

$$\sum_i h_1(R_i, L_i) \cdot h_2(X_i^{TX}, L_i) = 0. \quad (5)$$

Although both methods use equations of this form, the motivation for the equations and the functions h_1 and h_2 differ. G-estimation exploits the idea that treatment-free potential outcomes for those subjects randomized to treatment should be equal, on average, to treatment-free potential outcomes for those subjects randomized to no treatment. Treatment-free outcomes are the outcomes that each subject would have had if they had received no treatment instead of the treatment that they actually received, and therefore are incorporated into h_2 . Use of randomization status in h_1 ensures that average treatment-free outcomes are equal in the two randomized groups. IV estimators, on the other hand, are typically motivated as two-stage least squares (2SLS) estimates. 2SLS methods solve a system of equations in which the first equation is the causal model relating outcome and exposure, while the second equation uses an IV, such as randomization, to predict exposure. The joint solution to these two equations falls within the class of g-estimators [34, 35], with the second equation contributing to h_1 and the first contributing to h_2 . 2SLS methods can be applied to longitudinal data after the data are transformed to remove correlation between measures obtained from a single subject; this transformation can introduce differences in the performance of IV relative to g-estimation.

General Approach - G-estimation—In the original formulation of g-estimation for continuous outcomes in randomized data with noncompliance, Robins denotes the vector of treatment-free outcomes for subject i at times $t=1, \dots, K$ as $\mathbf{H}_i(\boldsymbol{\psi})$ [4]. He uses $\boldsymbol{\psi}$ to represent the average causal effects of treatment, whereas we use $\boldsymbol{\gamma}$. For subjects who were never

treated, $\bar{\mathbf{H}}_i(\psi)$ is equivalent to the vector of observed outcomes. For those who were treated, it is the vector of observed outcomes until the treatment occurs, after which it is the vector of observed outcomes minus the causal effect of treatment. In terms of the LSMM described in section 2.1, the subject-specific treatment-free outcome at time t , which is the element of the vector $\mathbf{H}_i$ at time t , is:

$$\begin{aligned} H_{it}(\psi) &= Y_{it}(s) - X_i^{TX}(t, s) \cdot \Delta(t, s) \\ &= \mathbf{X}_i^0(t)' \lambda + \mathbf{Z}_i(t, s)' \mathbf{b}_i + e_{it}(s), \end{aligned}$$

i.e. each treatment-free outcome is the potential outcome minus the component of the LSMM that includes the treatment effects γ . As a function of exposure, the treatment-free outcomes comprise part of h_2 in equation (5).

Due to randomization and the assumption of no-current-treatment-interaction, the average treatment-free outcomes in the group assigned to control should be the same as those in the group assigned to treatment, i.e.

$$\frac{1}{\sum_i \mathbb{1}_{[R_i=0]}} \sum_{i \text{ with } R_i=0} \bar{\mathbf{H}}_i(\psi) = \frac{1}{\sum_i \mathbb{1}_{[R_i=1]}} \sum_{i \text{ with } R_i=1} \bar{\mathbf{H}}_i(\psi).$$

This can be rewritten in the form of equation (5) by incorporating weights that depend on randomization group as specified by $h_1 = \mathbf{d}$: $\sum_i \mathbf{d}'(R_i) \bar{\mathbf{H}}_i(\psi)$. Incorporating baseline covariates X_{i0} , potentially including (or even limited to) the baseline outcome Y_{i0} , Robins has shown [4] that the more general estimating equations with arbitrary functions $\mathbf{d}$ and $\mathbf{q}$:

$$n^{-1/2} \sum_i \mathbf{d}'(R_i, X_{i0}) [\bar{\mathbf{H}}_i(\psi) - \bar{\mathbf{q}}(X_{i0})]$$

provide consistent estimates of the causal effects as long as the function $\mathbf{d}$ has the form $\mathbf{g}(R_i, X_{i0}) - E[\mathbf{g}(R_i, X_{i0})/X_{i0}]$ for some function $\mathbf{g}$. This is a singly-robust version of g-estimation that relies on correct specification of the structural model and exploits the existence of randomization, instead of relying solely on the assumption of sequential randomization that is typically used in g-estimation for observational data.

Robins notes that the semi-parametric-efficient “optimal” choices of $\mathbf{d}$ and $\mathbf{q}$, based on Proposition 1 in section 3 of Chamberlain’s paper [36], are as shown in Table 1. The optimal $\mathbf{g}$ is a function of an expected derivative $\mathbf{D}$ (Robins’ μ) and weights $\mathbf{w}$ that are an inverse variance. This choice results in $E[\mathbf{g}(R_i, X_{i0})/X_{i0}] = 0$, so that $\mathbf{g}_{opt}$ essentially replaces $\mathbf{d}$ in the estimating equations. The optimal $\mathbf{q}$ centers the treatment-free outcomes at their expected value, conditional on baseline covariates.

Assuming that the LSMM in section 2.1 represents the truth, the optimal functions $\mathbf{q}_{opt}$ and $\mathbf{g}_{opt}$ used to construct g-estimates can be written in terms of the LSMM parameters, as shown in Table 1. The function $\mathbf{q}_{opt}$ is $\mathbf{X}_i^0(t)' \lambda$, which makes the core of the estimating equations $\bar{\mathbf{H}}_i(\psi) - \mathbf{q}_{opt}$ equal to the residual $\varepsilon_{it} = \mathbf{Z}_i(t, s)' \mathbf{b}_i + e_{it}(s)$. The function $\mathbf{g}_{opt}$ depends on $\mathbf{D}$ and $\mathbf{w}$. The vector $\mathbf{D}$ is the partial derivative, with respect to the causal effects γ , of $\bar{\mathbf{H}}_i(\psi) = \mathbf{Y}_i(s) - \mathbf{X}_i^{TX}(s) \cdot \Delta(t, s)$. Thus in our example it is a vector over K two-dimensional conditional expectations of observed treatment and time since treatment, i.e.

$$\mathbf{D}_{it} = E \left(\begin{pmatrix} -X_i^{TX}(t, s) \\ -X_i^{TX}(t, s) \cdot (t-s) \end{pmatrix} \right) = \begin{pmatrix} -P[t \geq s | R_i, X_{i0}] \\ -P[t \geq s | R_i, X_{i0}] \cdot (t - E[s | t \geq s]) \end{pmatrix}.$$

The vector $\mathbf{w}$ is the inverse variance of $\mathbf{H}_i(\psi) - \mathbf{q}_{opt}$, conditional on randomization group and baseline covariates. For our LSMM, this is the inverse variance of $\mathbf{Z}_i(t, s)\mathbf{b}_i + e_{it}(s)$, which is a function of the time-varying treatment through the random effect on treatment.

Regardless of the choices of $\mathbf{q}_i$ and $\mathbf{g}_i$, the estimate of ψ (or γ in our notation) and its variance would then be [37]:

$$\hat{\psi} = (\mathbf{d}'\mathbf{M})^{-1} \mathbf{d}'(\mathbf{Y} - \bar{\mathbf{q}})$$

$$\text{Var}[\hat{\psi}] = (\mathbf{d}'\mathbf{M})^{-1} \frac{1}{n} \sum_i \{ \mathbf{d}_i'(\bar{\mathbf{H}}_i - \bar{\mathbf{q}}_i) (\mathbf{d}_i'(\bar{\mathbf{H}}_i - \bar{\mathbf{q}}_i))' \} (\mathbf{d}'\mathbf{M})'^{-1}$$

where $\mathbf{Y}$ and $\mathbf{q}$ are column vectors consisting of stacked vectors of post-baseline observations and q_i for each subject, $\mathbf{d}$ is the matrix consisting of stacked subject-specific matrices $\mathbf{d}_i$ each with the number of rows equal to the number of post-baseline observations and the number of columns equal to the number of treatment effects to be estimated, and $\mathbf{M}$ is the design matrix associated with the causal effects, also consisting of stacked subject-specific design matrices each with the same dimensions as $\mathbf{d}_i$.

Past attempts to implement efficient g-estimation with continuous outcomes [37, 38] have relied on several simplifying assumptions: 1) the control group had no access to treatment, 2) outcome measurements were available at baseline and one follow-up time only, and 3) $\mathbf{w}$ was assumed to be constant (and thus independent of R_i and X_{i0}). When the first assumption is satisfied, $\mathbf{H}_i(\psi)$ among control subjects does not depend on ψ , so $\mathbf{q}_{opt}$ and $\mathbf{w}$ can be estimated among the control group, assuming that there is no dependence of these functions on treatment group. The vector $\mathbf{D}$ will be zero among controls, so its dependence on R_i and X_{i0} can be estimated in the treated group. When the third assumption is satisfied, the expression for $\mathbf{g}_{opt}$ simplifies to $\mathbf{w}\{\mathbf{D} - E_{R_i}[\mathbf{D}/X_{i0}]\}$, eliminating the need to estimate $E_{R_i, X_{i0}}(\mathbf{w})^{-1}$.

We do not want to make any of these assumptions. The first two are easy to relax; however, if we do not make the third simplifying assumption, i.e. if we allow $\mathbf{w}$ to vary by treatment group, then it cannot simply be estimated in the control group and extended to the treatment group. This is the scenario that we face when the random effects $\mathbf{b}_i$ depend on treatment group, e.g. via b_{i2} in our example model. To implement g-estimation in this case, the functions $\mathbf{g}_{opt}$ and $\mathbf{q}_{opt}$ must be estimated among all subjects. Assuming no available baseline covariates, one algorithm for accomplishing this is the following:

1. choose starting values for both the λ and γ parameters;
2. estimate $\mathbf{H}_i(\psi) - \mathbf{q}_{opt}$ via $Y_{it} - X_i^{TX}(t, s) \cdot \Delta(t, s) - \mathbf{X}_i^0(t)' \lambda$;
3. estimate $\mathbf{g}_{opt}$ for each randomization group by using the design matrix for γ and the estimates of $\mathbf{H}_i(\psi) - \mathbf{q}_{opt}$ from 2);
4. update λ via linear regression of estimated treatment-free outcomes on the design matrix for λ ;
5. update γ using the Newton-Raphson method; and

6. repeat steps 2) through 5) until the estimating equations equal zero.

Key Assumptions - G-estimation—Consistency of g-estimation for uncensored quantitative outcomes relies on three key assumptions [4]: 1) the counterfactual outcomes must be independent of baseline randomized treatment assignment, 2) the structural model must be correctly specified, and 3) there must be no current-treatment-interaction, i.e. the effect of treatment at a specified time must be the same for subjects who receive it and for those who don't [39]. This is also referred to as sequential randomization, which is achieved if there are no unmeasured confounders. Semi-parametric efficiency requires correct specification of the functions in Table 1, but consistency does not. Correct specification of the structural model (assumption 2) should include consideration of interactions between causal effects and baseline covariates [37].

General Approach - Instrumental Variables—An instrumental variable (IV) is a variable whose effects on the outcome are exerted only via the exposure of interest [40]. Therefore any association between the IV and the outcome can be fully explained by exposure. One common example of an IV is randomization status - it determines exposure, but in most scenarios will have no other impact on outcomes. The IV estimator solves the same estimating equations as the g-estimator, but it does not use the optimal q and g functions, so it is not semi-parametric efficient. It is, however, still a valid estimator of the parameters of the SNM for which g-estimation was designed. Its implementation in standard software motivates us to include it as a comparator in our analyses.

Causal inference using IVs is typically understood in the context of a system of equations. The first equation is the structural model of interest, and the second equation uses one or more IVs to predict exposure [41], e.g.

$$Y_i = \lambda + \gamma X_i + \xi_i \quad (6)$$

$$X_i = \zeta_0 + \zeta_1 R_i + \omega_i \quad (7)$$

where ω_i has mean zero and can depend on ξ_i . The predicted exposure from equation (7) represents the expected exposure that is due to the IV(s). By substituting this expected exposure into the structural equation (6), one can recover the effect of exposure on outcome in the absence of selection because only the exposure that is due to the IV(s) is included. This approach is called two-stage least squares (2SLS).

In the context of longitudinal data, one or both sets of model parameters can be estimated via GEE or LME estimators to account for the correlation between outcomes collected at different timepoints [10]. Alternatively, 2SLS can be cast as a method of moments (MM) or generalized method of moments (GMM) estimator, depending on the number of IVs [9]. Econometricians have developed tools to implement these GMM methods in the presence of random intercepts, by first transforming the data to eliminate correlation between measurements on a single subject [42]. Specifically, the outcomes, exposure, and instruments are all passed through a feasible GLS transform that centers values within a subject in a way that incorporates estimated variances of $\mathbf{b}_0$ and e and diagonalizes the variance-covariance matrix. Conventional 2SLS methods are then applied to the transformed, uncorrelated data [42].

Key Assumptions - Instrumental Variables—The validity of IV estimators relies on the assumptions regarding the definition of the IV(s). For cross-sectional data, assumptions

include the following [40]: 1) random treatment assignment, 2) exclusion restriction, which implies that any effect of the instrument on outcome must be via its effect on treatment, 3) nonzero ACE of instrument on treatment, and 4) monotonicity, which implies that there is no individual treated with the opposite of his assignment regardless of the assignment. The last assumption of monotonicity is required for identification of the causal effect when the causal effect is heterogeneous, i.e. when it varies across subjects. There are other, weaker assumptions that could be used instead [43]. Or one can forgo a point estimate and report only confidence bounds [43, 44]; unfortunately these bounds are often uselessly wide. The sensitivity of IV estimates to IV assumptions increases with the amount of noncompliance to treatment [44, 45, 46].

For longitudinal data, the instrumental variable must satisfy the first three assumptions in every time period where a causal effect is to be estimated [41]. Further assumptions would then need to be made for identifiability. Careful consideration should be given to the inclusion of interactions with baseline covariates in both equations (6) and (7) [10].

3.4. Summary of Estimation Options

The properties of the methods presented in this section are summarized in Table 2. All methods except joint likelihood maximization and g-estimation have been implemented in standard software. Joint likelihood maximization can be approximately implemented in WinBUGS as illustrated in web-based supporting materials. To our knowledge, a general implementation of semi-parametric-efficient g-estimation has not been distributed, except with simplifying assumptions that are not satisfied in our problem of interest.

The choice of a method should include careful consideration of the most plausible selection model for treatment. If there is reason to believe that selection may depend on a latent variable, e.g. because all covariates that influence selection have not been recorded or have not been used in analysis, then joint likelihood estimators should be used because they are valid in the presence of indirect selection, while other methods are not.

If selection does not depend on a latent variable, then there are more options for valid analysis. LME estimates, joint estimates, and MSM estimates are all valid options. Initial-randomization-based-g-estimates and IV estimates can also be valid, given specific assumptions about random effects. LME estimates are generally preferable to joint estimates in this scenario because they require less strong assumptions. MSM or g-estimates are preferable to IV because they are designed for use with time-varying covariates. The choice from among LME estimates, MSM estimates, and g-estimates may depend on whether all factors that influence selection also belong in the structural model. If the two models include the same factors, then LME methods, which do not require explicit modeling of selection, can be used; however, if the two models include different factors, then methods that explicitly model selection separately from causal effects, such as MSM and g-estimates, are preferable. If there are heterogeneous treatment effects that are likely to be correlated with other subject-specific random effects, then initial-randomization-based-g-estimates are not guaranteed to be consistent, so MSM estimates should be used.

4. Simulations

We generated data that approximated the experimental conditions of the SPORT trial. We then estimated the treatment effect using a number of different methods and compared the results with the known treatment effect to evaluate the bias and variance of the analysis approaches. The goals of these simulations are twofold: 1) to anchor the treatment effects at known values using a LSMM and 2) to compare estimators for bias and efficiency under different selection scenarios.

4.1. Data Generation

We simulated data for 500 subjects, half of whom were assigned to surgery and half of whom were assigned to a control group. Outcome data ($\mathbf{Y}^O$) was generated at baseline and at four follow-up times: 6, 13, 26, and 52 weeks.

We assumed that no subjects had had surgery at baseline, and evaluated two different models for selection to surgery after baseline. In both selection models, subsequent surgery status at each follow-up time was determined randomly by a Bernoulli trial with probability p_{it} that depended on assigned treatment group, time, and the single most recent outcome measure. The first selection model – direct selection – included only these factors, while the second selection model – indirect selection – also included the unmeasured propensity to benefit from surgery in the model for p_{it} . Specifically, p_{it} was determined by the following expression, with $\delta(b_{i2})$ set to zero in the direct selection model:

$$\begin{aligned}\text{logit}(p_{it}) &= \alpha(t, R_i) + \eta(Y_{i,t-1}^O) + \delta(b_{i2}) \\ \alpha(t, R_i) &= (-1.5 - 0.05 \cdot t) \cdot (1 - R_i) + (-0.75 - 0.04 \cdot t) \cdot R_i \\ \eta(Y_{i,t-1}^O) &= 0.5 \cdot (\mu_{t-1} - Y_{i,t-1}^O) \\ \delta(b_{i2}) &= 4 \cdot b_{i2}\end{aligned}$$

with R_i equal to zero for subjects randomized to the control group and equal to one for subjects randomized to the treatment group, $(t-1)$ representing the observation time one prior to time t , and μ representing the population mean among those subjects not yet treated with surgery. The coefficients in these expressions were chosen so that the treatment proportions would be similar to those observed in the SPORT trial. The amount of correlation between past outcome and current treatment could be controlled by the coefficient in the expression for η . A Bernoulli trial was conducted at each follow-up time t ; when the trial result was that surgery had occurred, the time of surgery s was determined randomly with a Uniform distribution on the time period $(t-1, t]$.

Because treatment status depended on previous outcomes, we generated the outcome data sequentially, using the following model for subject i at time t (in weeks since enrollment):

$$Y_{it}^O = [28 + 0.05 \cdot t + 0.75 \cdot \min(t, 26)] + X_i^{TX}(t, s_i) [3 + 0.4 \cdot \min(t - s_i, 26)] + [b_{i0} + b_{i1} \cdot t + b_{i2} \cdot X_i^{TX}(t, s_i)] + e_{it}$$

where $\mathbf{b}_i$ were random effects specific to each subject, and e_{it} represented measurement error. The vector $\mathbf{b}_i$ was assumed to have a multivariate normal distribution centered at $(0,0,0)$, with variances $(2.67, 0.05, 1)$ and pairwise correlations of 0.8. No dependence on time of surgery was incorporated into measurement errors, making this a rank-preserving model when b_{i2} is set to zero. Measurement error was assumed to be normally distributed with mean 0 and variance 1.33.

4.2. Specification of Estimators

Our goal was to estimate the treatment effects of surgery over time, represented by the parameters γ_0 and γ_1 in our simulations. We first estimated these coefficients in the intent-to-treat (ITT) framework via LME estimators that assumed $s = 0$ for those randomized to surgery and $s = \infty$ for those randomized to nonoperative treatment. Several random effects structures were implemented, but all gave similar results, so only the most complex one is reported in Section 4.3. We then implemented LME and GEE estimators using actual surgery times, including covariates for time, $\min(t, 26)$, treatment status $\mathbf{1}_{\{t \geq s\}}$, and time since treatment $\mathbf{1}_{\{t \geq s\}} \cdot \min(t - s, 26)$.

We then implemented methods designed to account for the time-dependent confounding induced by selective treatment. For the marginal structural model approach, we created a stabilized weight for each subject and then obtained estimates using weighted GEE with independent covariance structure. To calculate the stabilized weights, we created two logistic regression models that estimated the probability of treatment for those subjects who had not yet received treatment, first with $(\widehat{p}_{it})$ and then without $(\widehat{p}_{it}^*)$ use of previous outcomes. Assigned treatment group and time, which were used to determine treatment status, were included in both logistic regression models, with most recent outcome also included in the first model. Interactions with time were allowed for both assigned treatment group and most recent outcome because the probability of treatment varied over time. One stabilized weight value was calculated for each subject at each time. The value was a ratio, with numerator = the product of $(1 - \widehat{p}_{it}^*)$ for all times up to and including time t where no surgery occurred and $\widehat{p}_{it}^*$ at the time when surgery occurred (if it did), and denominator = the product of $(1 - \widehat{p}_{it})$ for all times up to and including time t where no surgery occurred and $\widehat{p}_{it}$ at the time when surgery occurred (if it did).

For the IV approach, we evaluated 2SLS using the `xtivreg` command with the `re` option in Stata 10 [42]. The `re` option assumes that the true random effect structure includes only random intercepts. Randomization status and its interaction with time were used as IVs for the immediate and subsequent treatment effects. Covariates included t and $\min(t, 26)$.

We implemented an initial-randomization-based-g-estimator using a routine in R that takes starting values for the parameters λ and γ , then updates the parameter values iteratively until the estimating functions equal zero. The residuals $\mathbf{H} - \mathbf{q}$ are estimated by substituting the current estimates of λ and γ into the expression $Y_{it} - X_i^{TX}(t, s) \cdot \Delta(t, s) - \mathbf{X}_i^0(t)' \lambda$. The optimal functions $\mathbf{g}_{opt}$ are estimated within each randomized group using the residuals and

the observed design matrix $X_i^{TX}(t, s) \cdot \begin{pmatrix} 1 \\ (t-s) \end{pmatrix}$ for γ . The λ parameters are updated using a linear regression model of the estimated $t - s$ treatment-free outcomes

$\mathbf{H}_{it} = Y_{it} - X_i^{TX}(t, s) \cdot \Delta(t, s)$ on the design matrix $\mathbf{X}_i^0(t)$ for λ ; the δ parameters are updated using a Newton-Raphson method. We used the known true parameter values as starting values.

In summary, we used the following estimators:

1. ITT, implemented using LME assuming random intercepts and random slopes for both time and X_t
2. LME, assuming random intercepts
3. LME, assuming random intercepts and random slopes for time
4. LME, assuming random intercepts and random slopes for both time and X_t
5. GEE, assuming independent covariance structure
6. GEE, assuming exchangeable covariance structure [same as #1]
7. MSM, implemented using GEE with independent covariance structure and estimated stabilized weights
8. 2SLS, assuming random intercepts
9. G-EST, with randomized group as the only baseline covariate

We evaluated each of these estimators in 1000 simulations. In order to compare the results of these analyses to the actual values used to generate the data, we averaged the estimated

coefficients and their corresponding standard errors across the simulations. We also calculated the standard deviation of the estimated coefficients to evaluate precision.

The joint likelihood estimator that we have proposed (Section 3.1) is not conducive to evaluation in simulations due to the amount of time that it would take to repeatedly implement it in a Bayesian framework. However, we did implement this estimator using a representative data set from the set of simulations that incorporated indirect selection. For purposes of comparison, we implemented two estimators in the Bayesian framework - one that incorporated the selection model when computing the joint likelihood and one that did not. The latter is an alternate way to implement mixed model methods, while the former should give a valid answer even when indirect selection exists, so long as the selection model is correctly specified. By comparing the estimates obtained using these two models, we can evaluate the sensitivity of the mixed model estimates to the assumption of no indirect selection. We can also evaluate the indirect selection assumption itself by looking at the estimate of the coefficient of the random effect in the selection model. Proper but weakly informative priors were used to implement the joint likelihood maximization. The corresponding WinBUGS code can be found in web-based supporting materials.

4.3. Results

Estimates for γ_0 and γ_1 are displayed in Table 3. In comparing traditional methods for longitudinal data - GEE and LME - estimation in the presence of selection bias is less biased in LME estimators than in GEE estimators. LME estimators incorporate changes over time within individuals in the estimation process, rather than just incorporating cross-sectional averages, so they are more robust to selection bias. However, when indirect selection exists, maximization of the joint likelihood of treatment and outcomes must be used to obtain consistency. Properties of tailored causal methods - MSMs, IV estimators, and g-estimators - vary. MSMs provide consistent estimates when selection is history-driven, but are biased when selection is latent-variable-driven. IV estimates are consistent under HDS with random intercepts only, and demonstrate no bias under LVDS. G-estimates are consistent under HDS when there is not a random effect on treatment, but are not consistent under LVDS. All three tailored methods are less efficient than likelihood-based methods when likelihoods are correctly specified.

4.3.1. Likelihood-based Estimators—As discussed previously, there are many cases when LME estimators are consistent provided that the random effects are correctly specified. This can be seen in the first three scenarios displayed in Table 3. The LME estimates of γ_0 and γ_1 from these first three sets of simulations illustrate the importance of using a random effects structure that is at least as complex as the true one. The downside to using a more complex structure is an increased chance that the estimation algorithm will not converge. For example, if the LSMM includes only random intercepts, but we try to obtain an LME estimator with random intercepts and random slopes on time, convergence is not obtained 15% of the time. If we add random slopes on treatment as well, then convergence is not obtained 73% of the time. That said, if the truth (as reflected in the LSMM) has random slopes on both time and treatment, then they must be included in the LME estimator to avoid bias. In data generated with both random slopes, the LME estimator including them did not converge 23% of the time in our simulations. The fact that the standard deviation for this “correct” estimator is only calculated from 770 model fits, as opposed to 1000 for the other estimators, explains why it does not appear to be the most efficient estimator.

Although LME estimators can be valid, there are scenarios when they are not consistent even with correct specification of the random effects structure. The fourth situation in Table

3 illustrates the bias that can occur if indirect selection exists in addition to direct selection. This bias increases with the strength of the indirect selection, and can be greater than 10%.

To obtain valid estimates in the presence of indirect selection, we implemented an estimator that relied on maximization of the joint likelihood of a representative data set from the set of simulations that incorporated indirect selection. For comparison, the Bayesian implementation of a mixed model with random intercepts and slopes on both time and treatment gave the following estimated effects and standard deviations: $\hat{\gamma}_0=4.05(0.108)$ and $\hat{\gamma}_1=0.42(0.0097)$. On the other hand, the maximization of the joint likelihood that incorporated the selection model gave the following: $\hat{\gamma}_0=2.95(0.199)$ and $\hat{\gamma}_1=0.40(0.0100)$. In the latter implementation we were able to detect the indirect selection based on propensity to benefit from treatment via its estimated coefficient in the selection model: 4.45(1.052). This joint likelihood estimator provides an alternative to the methods explored in simulations that will remain valid even in the presence of indirect selection. The need for this alternative should be explored when there are plausible pathways for indirect selection.

4.3.2. Tailored Causal Methods—Methods that are specifically tailored to account for the endogenous exposure, such as MSM, IV, and g-estimators, are also unbiased in the first two scenarios in Table 3. The downside to these methods is substantial loss of efficiency, though MSM is more efficient than IV or g-estimation in our simulations. The IV estimates did not suffer from the nonconvergence of LME estimates, but some values were extreme. Therefore the lowest 2.5% of the estimates were replaced with the 2.5%ile and the highest 2.5% were replaced with the 97.5%ile, i.e. the estimates were Winsorized [47]. Then medians were reported in place of means. In the third scenario, all three estimates were biased downward, but only the IV estimate was more than 10% biased. Inspection of the estimating equations for g-estimation reveals that the bias results from correlation between the random treatment effect and the random intercept and slope on time. Details are in Appendix A. In the fourth scenario, with indirect selection, the IV method appears to be unbiased, but MSM and g-estimation do not; justification for the bias in g-estimation is again in Appendix A. The rationale for the performance of the IV estimator is similar to that for g-estimation; however, the centering transformation that is applied prior to implementation of 2SLS sometimes alleviates the potential for bias that is apparent in g-estimation.

In data not shown, we investigated the bias for all of the methods when different underlying models were assumed. As expected, stronger selection bias generated more bias in estimation, as did more variance in random effects. Weaker correlation between random effects, at least those not involved in indirect selection processes, decreased bias.

5. Illustration

The estimation methods used in simulations were applied to data from the SPORT randomized trial for treatment of lumbar intervertebral disk herniation [1, 11]. All 501 subjects with known surgery time (or known not to have had surgery) were included, even if some outcome observations were missing. The outcome of interest was the SF-36 physical function score, and all available observations through one year were used, providing a total of 2155 observations.

The results of selected analyses are displayed in Figure 2. The first plot includes the ITT and as-treated (AT) analyses that have been published previously [1, 11]. The ITT analysis showed little difference between the groups randomized to surgical versus nonoperative treatment, as would be expected in a trial with substantial crossover between treatment groups. The AT analysis, which used a longitudinal mixed effects regression model based on

the time of surgery and adjusted for baseline factors and interim outcomes affecting compliance, showed a substantial difference between those who did and did not receive surgery.

The previous analyses for SPORT had identified an underlying time trend in the SPORT data that was more complicated than the spline with one knot at 26 weeks that was used for our simulations. Therefore, in obtaining the estimators discussed in this section, we added knots at 6 and 13 weeks, using the following model:

$$E[Y_{it}] = \lambda_0 + \sum_{k=1}^4 \lambda_k \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} + \mathbb{I}_{[t \geq s_i]} [\gamma_0 + \gamma_1 \cdot \min(t - s_i, 26)]$$

where $(t_1, t_2, t_3, t_4) = (0, 6, 13, 26)$ are the knots in the time spline and s_i is an individual's surgery time. We maintained the two-component treatment effect comprised of an immediate effect of treatment (γ_0) and an effect of treatment over time up to 26 weeks after surgery (γ_1). Estimates of these treatment effects, along with the cumulative estimated effect at 26 weeks, are presented in Table 4.

Explicit selection models were necessary for implementing the MSM estimators. There was not evidence of difference in the impact of previous outcomes on treatment in the two randomized groups, but there was evidence that the impacts of both previous outcomes and randomization group varied with time. Therefore, the first selection model used in MSM estimation, which includes interactions with the time spline, is:

$$\begin{aligned} \text{logit } P[X_t^{TX} = 1 | X_{t-1}^{TX} = 0, \bar{Y}_{t-1}, \mathbf{b}_i] = & \alpha_0 + \sum_{k=1}^4 \alpha_k \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} + \alpha_5 \cdot R + \alpha_6 \cdot Y_{t-1} \\ & + \sum_{k=1}^4 \alpha_{k+6} \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} \cdot R + \sum_{k=1}^4 \alpha_{k+10} \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} \cdot Y_{t-1} \end{aligned} \quad (8)$$

where $(t_1, t_2, t_3, t_4) = (0, 6, 13, 26)$. In equation (8), both randomization and previous outcome were strong predictors of treatment, with those randomized to the treatment group being more likely to receive treatment and those with higher previous outcomes being less likely to receive treatment. Controlling for randomization assignment, at 6 weeks after enrollment, subjects with a five-unit higher previous outcome had, on average, odds of receiving surgery that were 0.91 (95% CI = 0.86–0.98) times the odds among those with previous outcome five units lower. This odds ratio decreased over time: 0.83 (0.78–0.89) at 13 weeks and 0.78 (0.71–0.85) at 26 weeks, implying that among those not yet treated with surgery, lower previous outcomes were more influential in the decision to perform surgery as time progressed.

It is plausible that selection depended on more than just the most recent previous outcome. Therefore a second selection model that added terms for the difference between the two most recent outcomes was also used in MSM estimation:

$$\begin{aligned} \text{logit } P[X_t^{TX} = 1 | X_{t-1}^{TX} = 0, \bar{Y}_{t-1}, \mathbf{b}_i] = & \alpha_0 + \sum_{k=1}^4 \alpha_k \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} + \alpha_5 \cdot R + \alpha_6 \cdot Y_{t-1} \\ & + \sum_{k=1}^4 \alpha_{k+6} \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} \cdot R + \sum_{k=1}^4 \alpha_{k+10} \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} \cdot Y_{t-1} \\ & + \alpha_{15} \cdot [Y_{t-1} - Y_{t-2}] + \sum_{k=1}^4 \alpha_{k+15} \cdot (t - t_k) \cdot \mathbb{I}_{[t > t_k]} \cdot [Y_{t-1} - Y_{t-2}]. \end{aligned} \quad (9)$$

We assumed that the difference in two previous outcomes was zero at the second visit because two previous visits were not available at this time. After controlling for randomization assignment, time, and the one most recent outcome, the difference in the two previous outcomes was not significantly associated with surgery treatment.

To implement g-estimation, the algorithm used in simulations was modified to account for missing data, as discussed in section 6 of Robins' work [4]. Missingness was assumed to be completely at random (MCAR), so that the inverse probabilities used for weighting depended only on time and randomization group. Standard errors were computed as in the work of Robins et al [48], but without a contribution from the estimation of inverse probability weights. Robins et al note that such standard errors are conservative.

As observed in simulations, the ITT analysis estimated treatment effects to be close to zero, which is consistent with the published results. The GEE analysis with working independence, which can be thought of as a naive as-treated estimation approach, gave a larger cumulative effect estimate of approximately 10 points. The mixed models that employed different random effects structures provided variable estimates of the immediate effect of surgery, ranging from 1.66 to 3.61, and similar estimates of about 0.7 for the effect over time following surgery. All three estimates, which assume no indirect selection, translated into a cumulative 26-week treatment effect of approximately 20 points. The increased treatment effect estimates given by LME, when compared to GEE, were consistent with our simulation results, and can be rationalized by recognizing that subjects with lower previous outcomes were more likely to cross over into the surgery group, as described in the weight model for the MSM. Therefore the trajectory of the untreated group is overestimated by GEE, leading the effect estimates to be smaller than the true effects. This can be seen by comparing the two middle panels in Figure 2.

Compared to the LME estimates, the MSM analysis that incorporated only the most recent past outcome into the selection model provided a larger effect estimate just after surgery and a comparable effect estimate over time following surgery. At 23 points, the MSM cumulative treatment effect estimate was just slightly higher than that from LME estimators. The increased standard errors were not as pronounced in this example as they were in our simulations, indicating that this analysis approach may be a useful alternative that relaxes the assumption of exogenous exposure. Its validity, however, depends on the assumption that the assumed selection model, equation (8), is correctly specified. Sensitivity analysis using the selection model in equation (9) led to a cumulative effect estimate of 21, which was similar to our first MSM estimate and to the LME estimates.

The IV and g-estimation analyses had substantially higher standard errors than all of the other analyses. The estimate of the effect at the time of surgery was much higher than estimates from other methods. The estimate of the effect over time after surgery was lower, and even negative with IV which would indicate that patients' physical function worsened over time. Combining these effect estimates led to cumulative effect estimates of less than 10 points using IV and 17 points using g-estimation. The IV estimate is not consistent with the rest of the analyses, but the g-estimate is not too different from likelihood-based estimates.

To illustrate the joint likelihood methods with data from the SPORT trial, we altered the WinBUGS code so that it incorporated variable visit times and a more flexible underlying time trend. Four estimators using this modified code are included in Table 4. All estimators assumed random intercepts (b_{0i}) and random slopes on both time (b_{1i}) and treatment (b_{2i}) in the structural model. What varied across the estimators were the assumptions regarding the selection model. First, no explicit selection was incorporated, providing the Bayesian

implementation of a corresponding LME estimator. Then three different selection models were illustrated:

$$\text{logit } P[X_t^{TX}=1|X_{t-1}^{TX}=0, \bar{Y}_{t-1}, \mathbf{b}_i] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot R + \alpha_3 \cdot t \cdot R + \alpha_4 \cdot b_{i0} \quad (10)$$

$$\text{logit } P[X_t^{TX}=1|X_{t-1}^{TX}=0, \bar{Y}_{t-1}, \mathbf{b}_i] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot R + \alpha_3 \cdot t \cdot R + \alpha_4 \cdot b_{i0} + \alpha_5 \cdot b_{i1} \quad (11)$$

$$\text{logit } P[X_t^{TX}=1|X_{t-1}^{TX}=0, \bar{Y}_{t-1}, \mathbf{b}_i] = \alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot R + \alpha_3 \cdot t \cdot R + \alpha_4 \cdot b_{i0} + \alpha_5 \cdot b_{i1} + \alpha_6 \cdot b_{i2} \quad (12)$$

All include time, randomization group, and an interaction between the two, but each includes a different combination of random effects as predictors: 1) only random intercepts, 2) random intercepts plus random slopes on time, and 3) random intercepts plus random slopes on both time and treatment.

Coefficient estimates from the selection part of the joint likelihood are displayed in Table 5. Probability of surgery was negatively associated with both outcome level (b_{0i}) and trend in outcome (b_{1i}). When only outcome level was included in the selection model, the cumulative effect estimate of 23 points was similar to the MSM estimate, which was a bit higher than the LME estimates. When trend in outcome was added, the effect estimate was substantially increased to 31 points. In the third selection model that also included propensity for treatment benefit, the significance of the associations between probability of surgery and both outcome level and trend in outcome were attenuated. The resulting estimated model included a positive association between probability of treatment and propensity to benefit from treatment, confirming that there is indirect selection based on latent treatment benefit. The effect estimate of 26 points from this model was less than that from the second model, but still higher than the MSM estimate. Accounting for this indirect selection primarily influenced the estimate of the immediate benefit of surgery, which makes sense given that the latent treatment benefit was tied to the immediate effect, not the effect over time. The decrease in $\hat{\gamma}_0$ can be attributed to the idea that it was overestimated using the other estimators due to failure to account for the fact that people with higher immediate treatment effects were more likely to be treated.

6. Discussion

Estimation of the average causal effect of treatment is a primary goal of randomized clinical trials. We have considered a scenario in which longitudinal data is available and interest lies in a one-time treatment, such as surgery, that is administered with noncompliance to randomized treatment assignment. The causal effect to be estimated can be specified by constructing a LSMM for the potential outcomes. The formulation of a LSMM is useful because it clearly identifies the effect of interest and is compatible with a number of different estimation methods. In choosing an estimation method, both standard longitudinal methods and methods designed specifically to estimate causal effects should be considered.

Because noncompliance with treatment assignment can create confounding, standard longitudinal methods are not necessarily expected to perform well. Our simulations support the existence of bias when using GEE estimators in the presence of such endogeneity. However, LME estimators perform well, at least in the case where indirect selection, i.e. selection based on both past and future outcome values, is not present. For maximum efficiency and minimum bias, LME models should incorporate the true random effects structure. However, models using complex random effects structures did not converge in a

large percentage of simulations, implying that simpler structures may need to be employed for practical applications.

Methods specifically designed to estimate causal effects in the presence of direct selection can provide unbiased effect estimates, but can also be inefficient. Increased variability in MSM is often the result of the presence of extreme weights, which can exist even when the weights are stabilized. Cole and Hernan [28] suggest weight truncation as one way to increase the efficiency of MSM estimates. The increase in standard errors was greater in our simulations than in the illustration using data from the SPORT trial. In both analyses, though, IV and g-estimation methods had much higher standard errors relative to the other methods of estimation.

When there is indirect selection, methods that explicitly maximize the joint likelihood of treatment and outcomes are necessary to obtain consistent estimation. We have shown how this can be accomplished in a Bayesian framework. Maximizing the joint likelihood in this way also permits exploration of the extent to which indirect selection exists in a given set of data.

One potential limitation of this research is that the indirect selection that we have studied is specific to random effects. The random effects capture information that would be contained in unmeasured covariates. Therefore similar simulation results would be obtained if unmeasured covariates were generated and input into the indirect selection model instead of random effects. However, for practical application, unmeasured covariates cannot be incorporated into an analysis, whereas their proxy – random effects – can.

In randomized clinical trials, estimation of causal effects via as-treated analyses is typically secondary, with primary analysis based on the intent-to-treat principle. However, in trials with such high noncompliance that the ITT analysis is rendered uninformative, the as-treated analysis becomes of primary importance. We recommend performing multiple analyses to estimate the causal effect so that sensitivity to the different assumptions required by the different methods can be assessed. However, if a single analysis must be chosen, our recommendation depends on whether the analyst believes that all likely confounders have been measured. We have shown that LME estimates can be valid for as-treated analyses with noncompliance, but that they are limited by their lack of flexibility in inclusion of different covariates in the selection model versus the structural model. Therefore, when analysts are fairly confident in the assumption of no unmeasured confounders, we suggest either MSM estimates or g-estimates, instead of LME estimates, due to their flexibility. In practice, MSM estimates are easier to implement than g-estimates, so we expect that they will continue to be used more frequently. However, when there is concern that important covariates may have not been collected, i.e. that unmeasured confounders exist, we recommend a likelihood-based joint model that can at least partially account for the unmeasured confounders via random effects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract/grant sponsor: US National Institutes of Health [T32 HL07183, R01 HL072966, and UL1 RR025014]

References

1. Weinstein J, Tosteson T, Lurie J, Tosteson A, Hanscom B, Skinner J, Abdu W, Hilibrand A, Boden S, Deyo R. Surgical vs Nonoperative Treatment for Lumbar Disk Herniation: The spine patient outcomes research trial (sport) randomized trial. *JAMA*. 2006; 296(20):2441–2450. [PubMed: 17119140]
2. Flum D. Interpreting Surgical Trials With Subjective Outcomes: Avoiding UnSPORTsmanlike Conduct. *JAMA*. 2006; 296(20):2483–2485. [PubMed: 17119146]
3. Ellenberg J. Intent-to-Treat Analysis Versus As-Treated Analysis. *Drug Inf J*. 1996; 30:535–544.
4. Robins J. Correcting for Noncompliance in Randomized Trials using Structural Nested Mean Models. *Commun Statist - Theory Meth*. 1994; 23(8):2379–2412.
5. Laird N, Ware J. Random-Effect Models for Longitudinal Data. *Biometrics*. 1982; 38(4):963–974. [PubMed: 7168798]
6. Liang KY, Zeger S. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*. 1986; 73(1):13–22.
7. Zeger S, Liang KY, Albert P. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 1988; 44(4):1049–1060. [PubMed: 3233245]
8. Robins J, Hernan M, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epi*. 2000; 11(5):550–560.
9. Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*. MIT Press; Cambridge, MA: 2002.
10. Bond S, White I, Walker A. Instrumental Variables and interactions in the causal analysis of a complex clinical trial. *Statist Med*. 2007; 26(7):1473–1496.
11. Weinstein J, Lurie J, Tosteson T, Skinner J, Hanscom B, Tosteson A, Herkowitz H, Fischgrund J, Cammisa F, Albert T, et al. Surgical vs Nonoperative Treatment for Lumbar Disk Herniation: The spine patient outcomes research trial (sport) observational cohort. *JAMA*. 2006; 296(20):2451–2459. [PubMed: 17119141]
12. Weinstein J, Lurie J, Tosteson T, Hanscom B, Tosteson A, Blood E, Birkmeyer N, Hilibrand A, Herkowitz H, Cammisa F, et al. Surgical vs Nonsurgical Treatment for Lumbar Degenerative Spondylolisthesis. *N Engl J Med*. 2007; 356(22):2257–2270. [PubMed: 17538085]
13. Weinstein J, Tosteson T, Lurie J, Tosteson A, Blood E, Hanscom B, Herkowitz H, Cammisa F, Albert T, Boden S, et al. Surgical vs Nonsurgical Treatment for Lumbar Spinal Stenosis. *N Engl J Med*. 2008; 358(8):794–810. [PubMed: 18287602]
14. Greenland S, Robins J, Pearl J. Confounding and Collapsibility in Causal Inference. *Stat Sci*. 1999; 14(1):29–46.
15. Hernan M, Cole S, Margolick J, Cohen M, Robins J. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidem Dr S*. 2005; 14(7):477–491.
16. Robins J. Correction for Non-compliance in Equivalence Trials. *Statist Med*. 1998; 17:269–302.
17. Diggle, P.; Heagerty, P.; Liang, K.; Zeger, S. *Analysis of Longitudinal Data*. 2. Oxford University Press; Oxford, UK: 2002.
18. Bellamy S, Lin J, Have TT. An introduction to causal modeling in clinical trials. *Clin Trials*. 2007; 4:58–73. [PubMed: 17327246]
19. Hogan J, Laird N. Mixture Models for the Joint Distribution of Repeated Measures and Event Times. *Statist Med*. 1997; 16(3):239–257.
20. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*. 2000; 1(4):465–480. [PubMed: 12933568]
21. Schluchter M, Konstan M, Davis P. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. *Statist Med*. 2002; 21:1271–1287.
22. Lindstrom M, Bates D. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *JASA*. 1988; 83 (404):1014–1022.
23. Zeger S, Liang KY. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*. 1986; 42(1):121–130. [PubMed: 3719049]

24. Little R. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *JASA*. 1995; 90(431): 1112–1121.
25. Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput*. 2000; 10 (4):325–337.
26. Hernan M, Brumback B, Robins J. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statist Med*. 2002; 21:1689–1709.
27. Mortimer K, Neugebauer R, van der Laan M, Tager I. An Application of Model-Fitting Procedures for Marginal Structural Models. *Am J Epidemiol*. 2005; 162(4):382–388. [PubMed: 16014771]
28. Cole S, Hernan M. Constructing Inverse Probability Weights for Marginal Structural Models. *Am J Epidemiol*. 2008; 168(6):656–664. [PubMed: 18682488]
29. Brumback B, Hernan M, Haneuse S, Robins J. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statist Med*. 2004; 23:749–767.
30. Ko H, Hogan J, Mayer K. Estimating Causal Treatment Effects from Longitudinal HIV Natural History Studies Using Marginal Structural Models. *Biometrics*. 2003; 59(1):152–162. [PubMed: 12762452]
31. Robins, J.; Rotnitzky, A.; Scharfstein, D. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: Halloran, M.; Berry, D., editors. *Statistical Methods in Epidemiology: The Environment and Clinical Trials*. Springer Verlag; New York: 1999. p. 1-94.
32. Bang H, Robins J. Doubly-robust estimation in missing data and causal inference models. *Biometrics*. 2005; 61:962–973. [PubMed: 16401269]
33. Moodie E, Richardson T, Stephens D. Demystifying Optimal Dynamic Treatment Regimes. *Biometrics*. 2007; 63:447–455. [PubMed: 17688497]
34. Joffe M, Brensinger C. Weighting in instrumental variables and G-estimation. *Statist Med*. 2003; 22(8):1285–1303.
35. Dunn G, Bental R. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions. *Statist Med*. 2007; 26(26):4719–4745.
36. Chamberlain G. Efficiency Bounds for Semiparametric Regression. *Econometrica*. 1992; 60(3): 567–596.
37. Goetghebeur E, Lapp K. The Effect of Treatment Compliance in a Placebo-controlled Trial: Regression with Unpaired Data. *Appl Statist*. 1997; 46 (3):351–364.
38. Fischer-Lapp K, Goetghebeur E. Practical Properties of Some Structural Mean Analyses of the Effect of Compliance in Randomized Trials. *Control Clin Trials*. 1999; 20(6):531–546. [PubMed: 10588294]
39. Joffe M, Small D, Have TT, Brunelli S, Feldman H. Extended Instrumental Variables Estimation for Overall Effects. *Int J Biostat*. 2008; 4(1):Article 4. [PubMed: 20231915]
40. Angrist J, Imbens G, Rubin D. Identification of Causal Effects Using Instrumental Variables. *JASA*. 1996; 91(434):444–455.
41. Hogan J, Lancaster T. Instrumental Variables and Inverse Probability Weighting for Causal Inference from Longitudinal Observational Studies. *Stat Meth Med Res*. 2004; 13(1):17–48.
42. StataCorp. Longitudinal/Panel Data Reference Manual, Release 10. Stata Press; College Station, TX: 2007.
43. Hernan M, Robins J. Instruments for Causal Inference: An Epidemiologist's Dream? *Epi*. 2006; 17(4):360–372.
44. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epi*. 2000; 29:722–729.
45. Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *JASA*. 1995; 90:443–450.
46. Martens E, Pestman W, de Boer A, Belitser S, Klungel O. Instrumental Variables: Applications and Limitations. *Epi*. 2006; 17:260–267.
47. Ruppert, D. Trimming and Winsorization. *Encyclopedia of Statistical Sciences* 2006.

48. Robins J, Rotnitzky A, Zhao L. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *JASA*. 1995; 90 (429):106–121.

Appendix A. Structural Mixed Models and g-estimation

G-estimation will provide unbiased estimates for the treatment effects when the expectation of the estimating equations equals zero. We rewrite the expectation of the estimating equations, conditioning separately on randomization group and on random effects and measurement error, in order to emphasize the conditions under which bias will occur. We assume that the correct LSMM is the example given in section 2.1 and that the selection model is as specified for simulations in section 4.1. Note that $X_{it} = \mathbb{1}_{\{t \geq s\}}$ and R_i is indicator of randomization group.

$$\begin{aligned}
 E_{Y_i, R_i, X_i} [d'_i(\mathbf{H}_i - \bar{\mathbf{q}}_i)] &= E_{\mathbf{b}_i, \mathbf{e}_i, R_i, X_i} [d'_i(\mathbf{H}_i - \bar{\mathbf{q}}_i)] \\
 &= E_{\mathbf{b}_i, \mathbf{e}_i, R_i} \{E_{X_i | \mathbf{b}_i, \mathbf{e}_i, R_i} [d'_i(\mathbf{H}_i - \bar{\mathbf{q}}_i)]\} \\
 &= E_{R_i} (E_{\mathbf{b}_i, \mathbf{e}_i | R_i} \{E_{X_i | \mathbf{b}_i, \mathbf{e}_i, R_i} [d'_i(\mathbf{H}_i - \bar{\mathbf{q}}_i)]\}) \\
 &= E_{R_i} (d'_i E_{\mathbf{b}_i, \mathbf{e}_i | R_i} \{E_{X_i | \mathbf{b}_i, \mathbf{e}_i, R_i} [b_{i0} + b_{i1} \cdot \mathbf{t} + b_{i2} \cdot \mathbb{1}_{\{t \geq s\}} + \mathbf{e}_i]\}) \\
 &= E_{R_i} (d'_i E_{\mathbf{b}_i, \mathbf{e}_i | R_i} \{b_{i0} + b_{i1} \cdot \mathbf{t} + b_{i2} \cdot E_{X_i | \mathbf{b}_i, \mathbf{e}_i, R_i} [\mathbb{1}_{\{t \geq s\}}] + \mathbf{e}_i\}) \\
 &= E_{R_i} (d'_i E_{\mathbf{b}_i, \mathbf{e}_i | R_i} \{b_{i2} \cdot P[t \geq s | \mathbf{b}_i, \mathbf{e}_i, R_i]\})
 \end{aligned}$$

Direct Selection: $P\{\mathbf{t} \geq s | \mathbf{b}_i, \mathbf{e}_i, R_i\}$ depends on R_i and $\mu_{t-1} - Y_{it-1} = b_{i0} + b_{i1} \cdot \mathbf{t}$

Indirect Selection: $P\{\mathbf{t} \geq s | \mathbf{b}_i, \mathbf{e}_i, R_i\}$ depends on R_i , $b_{i0} + b_{i1} \cdot \mathbf{t}$, and b_{i2}

If b_{i2} are correlated with $P\{\mathbf{t} \geq s | \mathbf{b}_i, \mathbf{e}_i, R_i\}$, then the estimating equations will be biased because the expectation will not be zero. This occurs either if b_{i2} is correlated with the other random effects (b_{i0} , b_{i1}) [scenario 3 in Table 3] or if selection depends on b_{i2} [scenario 4 in Table 3].

The relationships between X_i , Y_i , and $\mathbf{b}_i$ are displayed graphically in Figure 3. The random effects b_{i0} and b_{i1} always contribute to Y_i (and thus to X_i via Y_i) and sometimes contribute to X_i separately (dotted lines). The random effect b_{i2} contributes to Y_i only if the corresponding X_i equals 1, so doesn't contribute to X_i via Y_i , but sometimes contributes to X_i separately (dotted lines), as in scenario 4 in Table 3. If b_{i2} is correlated with b_{i0} and/or b_{i1} (dot-dash line), then it will be correlated with X_i even if it does not contribute separately to it, as in scenario 3 in Table 3.

(a) Population-average potential outcomes.

(b) Subject-average potential outcomes and observations.

Figure 1.

Model for potential outcomes over time, based on time from enrollment and treatment time. The model for population-average potential outcomes is $\lambda_0 + \lambda_1 \cdot t + \mathbb{1}_{\{t \geq s\}}[\gamma_0 + \gamma_1 \cdot (t - s)]$, where t = time since enrollment (in weeks) and s = treatment time (in weeks since enrollment). The subject average outcomes add random effects $\mathbf{b}_i$ to the population-average model. The additional incorporation of measurement error yields the subject-specific observations. Note that surgery can occur at any time, not just at the observation times displayed here, and that $s = \infty$ if surgery never occurs.

Figure 2.

Fitted mean models using SPORT data. The curves in the left frame are from the published SPORT analyses. The other three frames contain our intent-to-treat (ITT) estimator using dotted lines, and solid lines for (left to right) the generalized estimating equation (GEE) estimator, the linear mixed effect (LME) estimator with random intercepts and random slopes on both time and treatment, and the marginal structural model (MSM) estimator. Circles represent subjects with surgery just after enrollment, while triangles represent non-operative treatment.

Figure 3. Directed acyclic graphs (DAG) illustrating possible relationships between treatment, outcomes, and random effects. Subfigure (a) includes direct (solid) and indirect (dotted) selection, while subfigure (b) includes only direct selection.

Table 1

Functions needed to implement semi-parametric efficient g-estimation.

Function	General Notation	Scenario in Section 2.1
$\mathbf{q}_{opt;i}^-$	$E_{R_i}[H_{\lambda}(\psi) X_{i0}]$	$\mathbf{X}_i^0(\mathbf{t})' \lambda$
$\mathbf{g}_{opt;i}$ $\mathbf{D}_i$	$\mathbf{w}_i\{\mathbf{D}_i - E_{R_i, X_{i0}}[\mathbf{w}_i]^{-1} E_{R_i}[\mathbf{w}_i \mathbf{D}_i X_{i0}]\}$ $E\left[\frac{\partial \bar{\mathbf{H}}_1(\psi)}{\partial \psi'} R_i, X_{i0}\right]$	$\mathbf{w}_i\{\mathbf{D}_i - E_{R_i}[\mathbf{w}_i]^{-1} E_{R_i}[\mathbf{w}_i \mathbf{D}_i]\}$ $E\left(\begin{array}{c} -\mathbf{X}_i^{TX}(s) \\ -\mathbf{X}_i^{TX}(s) \cdot (\mathbf{t} - s) \end{array} \middle R_i\right)$
$\mathbf{w}_i$	$\{E[(\mathbf{H}_i(\psi) - \mathbf{q}_{opt}^-)(\mathbf{H}_i(\psi) - \mathbf{q}_{opt}^-)' R_i, X_{i0}]\}^{-1}$	$E[(\mathbf{Z}_i(\mathbf{t}, s)' \mathbf{b}_{i+} e_{it}(s))(\mathbf{Z}_i(\mathbf{t}, s)' \mathbf{b}_{i+} e_{it}(s))' / R_i]^{-1}$

Table 2

Properties of methods for estimating causal effects in longitudinal data.

	<u>Estimator</u>				
	GEE	LME	JOINT	MSM	G-EST
Implemented in standard software	yes	yes	no	yes	no
Requires correctly specified structural model	yes	yes	yes	yes	yes
Requires correctly specified selection model	no	no	yes	yes	no
Permits selection model and structural model to include different covariates	no	no	yes	yes	yes

Table 3

Mean and standard deviations of coefficient estimates across simulations. The true values were $\gamma_0 = 3$ and $\gamma_1 = 0.4$. Estimates with more than 10% bias are in bold. The first three scenarios assume a selection model with no indirect selection, i.e., $\mathcal{A}(b_2) = 0$. The third scenario includes all three random effects, as well as both direct and indirect selection. All scenarios assume two-way correlations of 0.8 between pairs of b_i .

Selection	LSMM	Estimation	$E[\widehat{\gamma_0}] (sd[\widehat{\gamma_0}])$	$E[\widehat{\gamma_1}] (sd[\widehat{\gamma_1}])$	% Non convergent
Direct	Random Intercepts Only	ITT	-0.02 (0.17)	0.08 (0.02)	0
		LME-intercept only (GEE-exchangeable)	3.00 (0.11)	0.40 (0.01)	0
		LME-intercept + slope on time	3.00 (0.10)	0.40 (0.01)	15
		LME-intercept + slopes on time and X_t^{TX}	3.00 (0.11)	0.40 (0.01)	73
		GEE-independence	2.02 (0.14)	0.39 (0.01)	0
		MSM, with estimated stabilized weights	2.97 (0.19)	0.40 (0.01)	0
		IV, assuming random intercepts*	3.00 (2.02)	0.40 (0.07)	0
Direct	Random Intercepts Plus Slopes on Time	G-EST	3.06 (0.94)	0.40 (0.04)	0
		ITT	-0.03 (0.15)	0.07 (0.019)	0
		LME-intercept only (GEE-exchangeable)	1.28 (0.24)	0.08 (0.03)	0
		LME-intercept + slope on time	3.00 (0.11)	0.40 (0.01)	0
		LME-intercept + slopes on time and X_t^{TX}	2.98 (0.11)	0.40 (0.01)	51
		GEE-independence	0.47 (0.25)	0.08 (0.03)	0
		MSM, with estimated stabilized weights	2.88 (0.61)	0.36 (0.07)	0
Direct	Random Intercepts Plus Slopes on Time and Treatment	IV, assuming random intercepts*	2.87 (1.94)	0.40 (0.21)	0
		G-EST	3.06 (0.93)	0.40 (0.09)	0
		ITT	-0.06 (0.15)	0.07 (0.02)	0
		LME-intercept only (GEE-exchangeable)	0.81 (0.25)	0.08 (0.03)	0
		LME-intercept + slope on time	2.56 (0.13)	0.40 (0.01)	0
		LME-intercept + slopes on time and X_t^{TX}	2.99 (0.14)	0.40 (0.01)	23
		GEE-independence	0.01 (0.28)	0.08 (0.03)	0
		MSM, with estimated stabilized weights	2.88 (0.70)	0.36 (0.07)	0

Selection	LSMM	Estimation	$E[\hat{\gamma}_0] \text{ (sd}[\hat{\gamma}_0])$	$E[\hat{\gamma}_1] \text{ (sd}[\hat{\gamma}_1])$	% Non convergent
Direct and Indirect	Random Intercepts Plus Slopes on Time and Treatment	IV, assuming random intercepts*	2.61 (2.10)	0.41 (0.22)	0
		G-EST	2.76 (0.98)	0.41 (0.09)	0
		ITT	-0.05 (0.16)	0.03 (0.02)	0
		LME-intercept only (GEE-exchangeable)	3.38 (0.27)	0.75 (0.03)	0
		LME-intercept + slope on time	3.95 (0.11)	0.43 (0.01)	0
		LME-intercept + slopes on time and X_t^{TX}	3.73 (0.13)	0.43 (0.01)	46
		GEE-independence	4.77 (0.27)	0.77 (0.03)	0
		MSM, with estimated stabilized weights	4.06 (0.28)	0.63 (0.027)	0
		IV, assuming random intercepts*	3.01 (10.48)	0.43 (0.46)	0
		G-EST	3.23 (1.59)	0.38 (0.13)	0

* In some cases the IV estimates had extreme values, so in all cases we replaced the outer 5% with the 2.5%ile and the 97.5%ile, then reported medians and standard deviations.

Table 4

Coefficient estimates and their standard errors, plus estimated treatment effect at 26 weeks, using SPORT data with spline knots at each time point. The previously-published AT 26-week effect estimate was 21.0 (2.283). All joint likelihood methods include random intercepts and random slopes on time and X_t^{TX} in the structural model. Assumptions about the selection model vary across joint likelihood estimators.

Estimation	$\widehat{\gamma}_0$ (se)	$\widehat{\gamma}_1$ (se)	26-week Effect $\widehat{\gamma}_0 + \widehat{\gamma}_1 \cdot 26$ (se)
ITT	0.61 (2.14)	0.06 (0.08)	2.1 (2.1)
LME-intercept only (GEE-exchangeable)	3.06 (1.92)	0.67 (0.10)	20.4 (1.7)
LME-intercept + slope on time	3.61 (1.83)	0.73 (0.10)	22.5 (1.9)
LME-intercept + slopes on time and X_t^{TX}	1.66 (2.08)	0.68 (0.10)	19.4 (2.0)
GEE-independence	-3.50 (2.52)	0.54 (0.11)	10.6 (2.3)
MSM, weights: past Y level	4.48 (2.74)	0.71 (0.14)	22.9 (3.3)
MSM, weights: past Y level and Y trend	4.05 (2.70)	0.66 (0.14)	21.1 (3.2)
IV, assuming random intercepts	11.25 (31.23)	-0.14 (1.12)	7.5 (13.8)
G-EST	13.17 (11.40)	0.14 (0.71)	16.9 (14.0)
JOINT, no explicit selection	2.10 (2.17)	0.69 (0.10)	20.0 (2.2)
JOINT, selection on b_{i0}	4.42 (2.24)	0.70 (0.10)	22.7 (2.3)
JOINT, selection on b_{i0} and b_{i1}	-3.36 (3.07)	1.33 (0.13)	31.3 (3.0)
JOINT, selection on b_{i0} , b_{i1} , and b_{i2}	-6.01 (3.31)	1.25 (0.13)	26.4 (4.0)

Table 5

Coefficient estimates and their standard errors for the random-effects terms in the selection model part of the joint likelihood.

Estimation	$\widehat{\alpha}_4$ (se)	$\widehat{\alpha}_5$ (se)	$\widehat{\alpha}_6$ (se)
JOINT, no explicit selection	-	-	-
JOINT, selection on b_{i0}	-0.03 (0.005)	-	-
JOINT, selection on b_{i0} and b_{i1}	-0.04 (0.008)	-5.41 (1.08)	-
JOINT, selection on b_{i0} , b_{i1} , and b_{i2}	-0.02 (0.011)	-3.87 (1.24)	0.05 (0.02)