6-30-2006

Random Effects Logistic Models for Analyzing Efficacy of a Longitudinal Randomized Treatment With Non-Adherence

Dylan S. Small
University of Pennsylvania

Thomas R. Ten Have
University of Pennsylvania

Marshall M. Joffe
University of Pennsylvania

Jing Cheng
University of Pennsylvania

Recommended Citation

Small, D. S., Ten Have, T. R., Joffe, M. M., & Cheng, J. (2006). Random Effects Logistic Models for Analyzing Efficacy of a Longitudinal Randomized Treatment With Non-Adherence. *Statistics in Medicine, 25* (12), 1981-2007. http://dx.doi.org/10.1002/sim.2313

Random Effects Logistic Models for Analyzing Efficacy of a Longitudinal Randomized Treatment With Non-Adherence

Abstract
We present a random effects logistic approach for estimating the efficacy of treatment for compliers in a randomized trial with treatment non-adherence and longitudinal binary outcomes. We use our approach to analyse a primary care depression intervention trial. The use of a random effects model to estimate efficacy supplements intent-to-treat longitudinal analyses based on random effects logistic models that are commonly used in primary care depression research. Our estimation approach is an extension of Nagelkerke *et al.*'s instrumental variables approximation for cross-sectional binary outcomes. Our approach is easily implementable with standard random effects logistic regression software. We show through a simulation study that our approach provides reasonably accurate inferences for the setting of the depression trial under model assumptions. We also evaluate the sensitivity of our approach to model assumptions for the depression trial.

Keywords
random effects, logistic regression, exclusion restriction, encouragement studies, mental health

Random effects logistic models for analyzing efficacy of a longitudinal randomized treatment with non-adherence

Dylan S. Small
Thomas R. Ten Have
Marshall M. Joffe
Jing Cheng

June 25, 2005

Jing Cheng is a graduate student, Marshall Joffe is Associate Professor, and Thomas Ten Have is Full Professor, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine Blockley Hall, 6th FLR 423 Guardian Dr. Philadelphia, PA 19104-6021 (E-mail:*jcheng@cceb.upenn.edu*, *mjoffe@cceb.upenn.edu*, *ttenhave@cceb.upenn.edu*); Dylan Small is Assistant Professor, Department of Statistics, Wharton School, 464 JMHH/6340, University of Pennsylvania Philadelphia, PA 19104 (E-mail:*dsmall@wharton.upenn.edu*). Correspondence to: Dylan Small, Department of Statistics, Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104 (e-mail:*dsmall@wharton.upenn.edu*).

Abstract

We present a random effects logistic approach for estimating the efficacy of treatment for compliers in a randomized trial with treatment non-adherence and longitudinal binary outcomes. We use our approach to analyze a primary care depression intervention trial. The use of a random effects model to estimate efficacy supplements intent-to-treat longitudinal analyses based on random effects logistic models that are commonly used in primary care depression research. Our estimation approach is an extension of Nagelkerke *et al.* (2000, *Statistics in Medicine*)'s instrumental variables approximation for cross-sectional binary outcomes. Our approach is easily implementable with standard random effects logistic regression software. We show through a simulation study that our approach provides reasonably accurate inferences for the setting of the depression trial under model assumptions. We also evaluate the sensitivity of our approach to model assumptions for the depression trial.

Keywords: random effects, logistic regression, exclusion restriction, encouragement studies, mental health.

1. Introduction

The central goal of a clinical trial is to make inferences about how treatment should be conducted in a general population of patients who will require treatment in the future [1]. Frequent complications for achieving this goal include the sample of patients in the trial being unrepresentative of the general population and the way in which the treatment is administered in the trial being different than the way it would be administered in the general population. Another common challenge for predicting the effect of future treatment programs based on a clinical trial is non-adherence to assigned treatment regime. Non-adherence is a common feature of trials with human subjects because adherence cannot be enforced for ethical reasons. Non-adherence causes difficulties for predicting the effect of future treatment programs when adherence patterns for future treatment programs are expected to differ from adherence patterns in the trial. For the future treatment program of making the treatment generally available to the population after a successful trial, adherence might be higher than in the trial because the treatment is accepted as efficacious as a result of the successful trial or adherence might be lower than in the trial because patients in the general population are given less encouragement to

take the treatment than patients in the trial [2]. When adherence patterns for future treatment programs are expected to differ from adherence patterns in the trial, a key quantity for accurately predicting the effect of a future treatment program based on the trial results is the "efficacy" of the treatment in the trial. The efficacy is a measure of how effective the treatment was relative to the control for those patients (or a subset of those patients) who adhered to the treatment regimen in the trial ([3, 4]; see also Section 10 of this paper). This paper develops a method for estimating efficacy for a study with longitudinal binary outcomes and applies it to a primary care depression treatment study.

Two commonly used methods for analyzing clinical trials are 1) intent-to-treat (ITT) analyses that compare patients assigned to the treatment arm to patients assigned to the control arm and 2) as treated (AT) analyses that compare patients who actually received the treatment to patients who did not receive the treatment. Both methods of analysis have flaws for analyzing efficacy. The ITT analysis does not aim to measure the efficacy of treatment actually received. Instead, the ITT analysis measures the programmatic effectiveness of offering, but not enforcing, treatment in the trial. When there is non-adherence, the programmatic effectiveness will generally differ from the efficacy. Note also that when the pattern of adherence for future treatment programs is expected to differ from that of the trial, the programmatic effectiveness of offering treatment in the trial that is measured by the ITT analysis will not generally be the same as the programmatic effectiveness of future treatment programs. In contrast to the ITT analysis, an AT analysis does aim to measure the efficacy of treatment actually received, but an AT analysis is biased when patients who would adhere to the treatment regimen if randomized to it are not comparable to patients who would not adhere to the treatment regimen if randomized to it. Often, the propensity for a successful outcome among those who would adhere to the treatment if offered it (would be adherers) is greater than among those who would not adhere to the treatment if offered it (would be non-adherers) when both groups do not receive treatment (e.g., [5, 6]). In contrast, in the depression study we consider, we find evidence that the propensity for a successful outcome when not offered treatment is less in would be adherers than would be non-adherers. Unlike AT analyses, instrumental variables (IV) methods for estimating efficacy do not require would be adherers and would-be non-adherers to be comparable to obtain

consistent estimates. Instead, IV methods require an "exclusion restriction" that specifies that the randomization assignment only affects the outcome through its effect on treatment received. IV methods have been developed for several types of studies and data [7, 8, 9]. This paper develops an IV method for estimating efficacy for longitudinal binary outcomes using a random effects logistic regression model.

The study that motivated our work is a randomized trial of an "encouragement" intervention to improve adherence to prescribed depression treatments among depressed elderly patients in primary care practices [10]. Each practice was randomized to either this encouragement intervention, the test treatment (henceforth referred to as the treatment), or to usual care, the control treatment (henceforth referred to as the control). The encouragment intervention is to have a depression specialist (typically a master's level clinician) closely collaborate with the depressed patient and the patient's primary care physician to facilitate patient and clinician adherence to a treatment algorithm and provide education, support and ongoing assessment to the patient. The study measured patients' Hamilton depression scores at baseline and three follow-up visits at 4, 8 and 12 months. A patient was considered to have adhered to the encouragement intervention if the patient had seen a depression specialist in the prior four months of follow-up. Patients in practices randomized to the usual care group did not have access to the depression specialist. One clinical question of interest is what is the effect of a patient's contact with the depression specialist over the past four months on the probability of a 50% or more reduction in a patient's Hamilton score from baseline. The binary outcome of whether or not there is a 50% or more reduction in a patient's Hamilton score from baseline has been advocated by a government panel as a standard for research on primary care treatment of depression [11]. The "transient" effect of the experimental treatment (the effect of contact with the depression specialist over the past four months) is focused on rather than the cumulative effect of treatment from baseline because of the expectation that the effect of contact with the depression specialist does not extend beyond the next visit four months later.

A random effects logistic model was used for the intent-to-treat analysis of the depression study described above in [10]. Random effects logistic models are commonly used in primary care depression treatment research, e.g., [12], because they provide subject-specific inferences. See

Zeger et al. [13, 14] for motivation for and discussion of estimating subject-specific parameters.

Our goal is to provide an analysis of efficacy that supplements the ITT analysis. We use a random effects logistic model to analyze efficacy. Random effects models for analyzing efficacy have several benefits. First, conditioning on random effects in estimating efficacy makes the efficacy estimates comparable to ITT estimates from a random effects model; this was an important motivation for using a random effects logistic model to estimate efficacy for the depression study. Second, random effects models provide a means of accommodating a certain type of informative dropout through a shared parameter model, e.g., [15, 16]. Third, although it is not the focus of this study, random effects models enable information to be borrowed from other subjects for making more accurate treatment decisions for a given patient based on limited longitudinal data for the given patient, e.g., [17, 18].

The methodological contributions of our paper are to formulate a random effects logistic model for analyzing efficacy and provide an easily implementable method for estimating it. Our approach to estimation is an extension of the approximate IV method for cross-sectional logistic models proposed by Nagelkerke et al. [19] and examined by Ten Have et al. [20]. A valuable feature of our approximate IV approach is that it can easily be implemented using standard random effects logistic regression software, e.g., proc NLMIXED in SAS with macros available from the authors. We show that our approximate IV approach produces approximately valid results for the setting of the depression study under model assumptions through a simulation study in Section 8. We also evaluate the sensitivity of our results to various model assumptions.

The depression study we consider is a "clustered encouragement design," meaning that the randomization was done at the cluster level of primary care practices rather than the individual level. Frangakis et al. [21] develop a framework and methods for studying clustered encouragement designs for a cross-sectional setting. In setting up our model, we consider both designs in which the sample is a simple random sample and randomization is done at the individual level and designs in which the sample is a clustered sample and randomization is done at the cluster level. For the depression study, there is only a small correlation between outcomes within practices. In the simulation study of Section 8 that is based on the setup of the depression study, a version of our estimation method which ignores the clustering performs better than a version of

our estimation method which takes the clustering into account.

We focus here on a depression study but the type of data for which our model is designed, longitudinal binary outcome data from a randomized trial with non-adherence, is common. Another example is a randomized trial of treatments for acute myeloid leukemia patients [22]. Literature on estimating efficacy for longitudinal studies with treatment non-adherence includes the following. Robins [8], using g-estimation for linear or log-linear models, focused on estimating cumulative effects of time-varying treatments on final outcome among those who adhere, in contrast to our focus on transient effects on intermediate outcomes. Sato [22] applied g-estimation for linear models without random effects to estimate additive cumulative effects of treatment on repeated measures binary outcomes in a randomized trial. Frangakis et al. [23] developed methodology for estimating the transient effect of a longitudinal treatment using the principal stratification framework for causal inference [24]. Frangakis et al. [23]'s approach differs from ours in that they make population-averaged inferences whereas we make subject-specific inferences using our random effects model. Yau and Little [25] assumed constant compliance status and adherence across time in fitting a random effects linear model in the principal stratification framework. A number of logistic or probit models have been proposed for causal inference based on cross-sectional binary outcomes, e.g., [26, 27, 28, 19, 20, 29].

Our paper is organized as follows. We present descriptive statistics for the depression study in Section 2; causal notation in Section 3; assumptions in Section 4; the model for potential outcomes in Section 5; the estimation approach in Section 6; strategies for assessing assumptions in Section 7; simulation results in Section 8; data analysis results for the depression study in Section 9; discussion of how efficacy is useful in predicting the effect of future treatment programs in Section 10; and general discussion in Section 11.

2. Depression Study

The depression study involved 539 patients in 20 primary care practices. Full details of the study are described in Bruce et al. [10]. The practices were paired in the randomization but in order to focus on main aspects of our methodology, we ignore the pairing in our analysis.

The following are descriptive statistics for the study. The differences in the proportion of successful outcomes between randomized groups diminishes across time, as does the proportion

receiving treatment in the randomized to treatment group. Specifically, the percentages of randomized to treatment patients with a 50% or more reduction in Hamilton score since baseline at 4, 8 and 12 months are 42.7, 46.2 and 52.1% respectively. The corresponding percentages in the randomized to usual care group are 29.1, 35.5 and 42.0 % respectively. The analogous percentages of successful outcomes for those who actually received the treatment are 43.0, 45.5 and 55%, whereas in the group that did not receive the treatment, including those randomized to usual care, the percentages are 29.8, 37.8 and 41.4%. The percentage of the randomized to treatment group that actually receives the treatment declines somewhat across the three follow-up visits: 92.9, 80.9 and 79.7% respectively. The data set is available by request from the authors. We now develop notation for defining the efficacy of receiving the treatment of seeing the depression specialist.

3. Notation

We use the potential outcomes model for causal inference [30, 31] to define the efficacy of an intervention. We shall assume that the clinical trial has a Zelen randomized single consent design [32, 33]. A single consent design has the following features: 1) the control is the best standard method of care (called usual care in the depression study); 2) everyone who does not take the test treatment (including those who are assigned to the test treatment group but do not adhere) receives the control which is the best standard method of care; and 3) the test treatment is not available to patients assigned to the control arm. We shall also assume that adherence with the test treatment is all or none. We consider a longitudinal study with a balanced design and $T > 1$ time points ($T = 3$ in the depression study).

Treatment received and randomization variables. The observed randomization variable is $R_i = 1$ if patient $i(= 1, \ldots, N)$ was randomized to the treatment group and $R_i = 0$ if patient i was randomized to the control group. In the depression study, the treatment entails meeting with the depression specialist and the control arm is usual care. The observed treatment-received variable is defined as follows: $A_{it} = 1$ if the treatment was actually received by patient i during the four months prior to time t ($t = 4, 8$ or 12 months), i.e., $A_{it} = 1$ if patient i actually met with the depression specialist in the four months prior to time t, and $A_{it} = 0$ otherwise. For the single consent design that we consider, $A_{it} = 0$ when $R_i = 0$ because patients assigned to the

control arm do not have access to the treatment.

Compliance status variables. To define the time-varying compliance classes of patients, we first define potential treatment-received variables. Let $A_{it}^{(1)} = a \in \{0, 1\}$ denote whether the ith patient would choose to adhere to the intervention during the four month period prior to time t if she or he were to be randomly assigned to the treatment arm ($r = 1$) and let $A_{it}^{(0)}$ be the corresponding potential treatment-received variable if the ith patient were to be assigned to the control arm ($r = 0$). For a clustered design, $A_{it}^{(r)}$ denotes whether the ith patient would choose to adhere to the intervention if the ith patient's cluster was randomly assigned to arm r. The compliance class of a patient classifies a patient by $(A_{it}^{(0)}, A_{it}^{(1)})$ [7]. For the single consent design, the control group does not have access to the treatment so that $A_{it}^{(0)} = 0$ for all patients and the compliance classes can be defined in terms of $A_{it}^{(1)}$. We denote the compliance class indicator variable as $C_{it} = c$ where $c = 1$ for compliers ($A_{it}^{(1)} = 1$) and $c = 0$ for never-takers ($A_{it}^{(1)} = 0$). Compliers are those patients who would receive the treatment if assigned to it, and never-takers are those patients who would never receive the treatment even if assigned to it. Note that these compliance classes are only partially observed; they are observed if $R_i = 1$ but unobserved if $R_i = 0$. In the terminology of Frangakis and Rubin [24], the vector of compliance classes for patient i, $\mathbf{C}_i = (C_{i1}, \ldots, C_{iT})$, is a principal stratification with respect to adherence to treatment assignment.

Observed and potential outcome variables. The potential outcomes are $Y_{it}^{(1)}$, the binary outcome (50% or more improvement in baseline Hamilton score) that would have been observed had patient i (patient i's cluster for a clustered design) been randomly assigned to the treatment ($r = 1$) arm, and $Y_{it}^{(0)}$, the binary outcome that would have been observed had patient i (patient i's cluster for a clustered design) been randomly assigned to the control ($r = 0$) arm. The corresponding observed outcome is $Y_{it} = y \in \{0, 1\}$, which denotes the outcome that was observed for patient i at time t. Note that we specify a pair of potential outcomes for each of T time points for a patient, but observe only one potential outcome at each time point for a patient.

Observed and potential missed visits and drop-out variables. The potential missed visit and potential drop-out variables are $O_{it}^{(r)} = o \in \{0, 1\}$ and $D_{it}^{(r)} = d \in \{0, 1\}$, which denote whether a

research visit (for O_{it}) or drop-out (for D_{it}) would have occurred at time t had patient i (patient i's cluster for a clustered design) been randomly assigned to arm r assuming that the patient has not dropped out of the trial by time $t-1$. The corresponding observed missed visit and drop out variables are $O_{it} = o \in \{0,1\}$ and $D_{it} = d \in \{0,1\}$, which denote whether a research visit (drop-out) occurred at time t for patient i. We define T_i as the last time a visit occurred for patient i, $T_i \leq T$.

Covariates. The non-treatment covariates include baseline and visit indicator variables. The vector of baseline covariates for patient i is denoted by $\mathbf{X}_i$. For the depression study, the elements of $\mathbf{X}_i$ are baseline Hamilton depression score and baseline suicide ideation score. The vector of time variables is denoted by $\mathbf{T}_t$. For the depression study, $\mathbf{T}_t$ consists of three dummy variables for the three time points (4, 8 and 12 months). We tried specifying $\mathbf{T}_t$ as an intercept and a linear term for time but found that this model did not fit well relative to the saturated model with dummy variables for visits.

Random effects. We define a vector of unobserved random effects, $\boldsymbol{\tau}_i$, for the outcome model for patient i. The elements of this vector include a random intercept, $\tau_{0\,i}$, and if necessary random polynomial terms such as a random slope $\tau_{1\,i}$ for time. The design matrix that links the random effects to the outcomes is $\mathbf{Z}_i$ with t-th row $\mathbf{Z}_{it}$. For a clustered design, we also consider a vector of random effects for the hth cluster, $\boldsymbol{\iota}_h$, with design matrix $\mathbf{V}_h$.

Clusters. For a clustered design, we use the notation that there are n clusters, n_h members of the hth cluster, and the ith member of the hth cluster is indexed by hi, e.g., the baseline covariates for the ith member of the hth cluster are $\mathbf{X}_{hi}$ and the randomization assignment is R_{hi}. The cluster of a given subject j is denoted by P_j, i.e., $P_{hi} = h$. For conciseness, we use the notation for an unclustered design below except when we discuss the clustered design explicitly.

4. Assumptions

Standard assumptions for interpreting or estimating causal effects when there is treatment nonadherence are: 1) Stable Unit Treatment Value Assumption (SUTVA); 2) randomization of treatment assignment; 3) exclusion restriction; and 4) monotonicity. The standards assumptions need to be augmented with additional assumptions that are needed for the logistic link function, longitudinal and missing data, and random effects. We test some of these assumptions and

address sensitivity of our approach to others in Sections 7-9.

4.1 Sampling Assumptions

For an unclustered design, we assume that the vectors

$$\begin{aligned}\mathbf{G}_i \;=\; & (Y_{i1}^{(0)}, Y_{i1}^{(1)}, \ldots, Y_{iT}^{(0)}, Y_{iT}^{(1)}, A_{i1}^{(0)}, A_{i1}^{(1)}, \ldots, A_{iT}^{(0)}, A_{iT}^{(1)}, \\ & O_{i1}^{(0)}, O_{i1}^{(1)}, \ldots, O_{iT}^{(0)}, O_{iT}^{(1)}, D_{i1}^{(0)}, D_{i1}^{(1)}, \ldots, D_{iT}^{(0)}, D_{iT}^{(1)}, \\ & R_i, C_{i1}, \ldots, C_{iT}, Y_{i1}, \ldots, Y_{iT}, A_{i1}, \ldots, A_{iT}, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\tau}_i),\end{aligned}$$

$i = 1, \ldots, n$, are i.i.d., each with the same distribution as the random vector

$$\begin{aligned}& (Y_1^{(0)}, Y_1^{(1)}, \ldots, Y_T^{(0)}, Y_T^{(1)}, A_1^{(0)}, A_1^{(1)}, \ldots, A_T^{(0)}, A_T^{(1)}, O_1^{(0)}, O_1^{(1)}, \ldots, O_T^{(0)}, O_T^{(1)}, \\ & D_1^{(0)}, D_1^{(1)}, \ldots, D_T^{(0)}, D_T^{(1)}, R, C_1, \ldots, C_T, Y_1, \ldots, Y_T, A_1, \ldots, A_T, \mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}), \qquad (1)\end{aligned}$$

where $Y_{it} = Y_{it}^{(R_i)}$, $A_{it} = A_{it}^{(R_i)}$ and $Y_t = Y_t^{(R)}, A_t = A_t^{(R)}$. For a clustered design, we assume that the random vectors $(\mathbf{G}_{h1}, \ldots, \mathbf{G}_{hn_h}, n_h, \boldsymbol{\iota}_h)$, $h = 1, \ldots, n$, are i.i.d. All subsequent probability and expectation statements will be in terms of the random vector (1), where statements like $P(Y_t \mid \mathbf{A}_i, R_i)$ are shorthand for $P(Y_t \mid \mathbf{A} = \mathbf{A}_i, R = R_i)$.

4.2 SUTVA Assumption

The Stable Unit Treatment Value Assumption (SUTVA) assumes that the model's representation of potential variables is adequate to describe the effect of the interventions that are under consideration [34]. Here, our potential outcome variables $Y_{it}^{(r)}$ allow only for differences in treatment assigned r for patient i for an unclustered design and only for differences in treatment assigned r for cluster P_i in a clustered design. For this representation to satisfy SUTVA for the types of randomized trials being considered, we must assume that (1) there is a single value of the potential outcome $Y_{it}^{(r)}$ regardless of the randomization assignment of any other patient $i' \neq i$ in an unclustered design and regardless of the randomization assignment of any other cluster $h' \neq P_i$ in a clustered design; and (2) there is a single value of the potential outcome $Y_{it}^{(r)}$ regardless of the method of treatment assignment or administration.

Assumption (1) allows us to use scalar notation for the treatment-assigned indices of the potential outcomes that refer to patient i rather than vectors of treatment assigned indices when defining potential outcomes for patient i. Assumption (2), often called the SUTVA consistency

assumption, enables us to relate the observed and potential outcomes:

$$Y_{it} = R_i Y_{it}^{(1)} + (1 - R_i) Y_{it}^{(0)}.$$

A similar assumption enables us to relate the observed and potential missed visit and drop-out variables, $O_{it} = R_i O_{it}^{(1)} + (1 - R_i) O_{it}^{(0)}$, and $M_{it} = R_i M_{it}^{(1)} + (1 - R_i) M_{it}^{(0)}$.

4.3 Exclusion Restriction for Never Takers

We assume that for never takers at time t (patients with $A_{it}^{(1)} = 0$), random assignment to the treatment versus the control arm has no effect on potential outcomes, missed visits and drop-out:

$$\text{If } A_{it}^{(1)} = 0, \text{ then } Y_{it}^{(1)} = Y_{it}^{(0)};\ O_{it}^{(1)} = O_{it}^{(0)};\ D_{it}^{(1)} = D_{it}^{(0)} \tag{2}$$

(2) is called an exclusion restriction for never takers because it excludes an effect of randomization assignment on outcomes for never takers. This exclusion restriction is more likely to hold with blinding of treatment assignments to patients and clinicians, which is not the case for the depression study. We assess the robustness of our approach to violations of the exclusion restriction assumption in Section 9.2.2. See Hirano et al. [27] for further discussion of exclusion restriction assumptions.

4.4 Missing Data Assumptions

We assume that the observed missed visit and drop-out processes $(O_1, \ldots, O_T, D_1, \ldots, D_T)$ are independent of the outcomes $(Y_1, \ldots, Y_T) = (Y_1^{(R)}, \ldots, Y_T^{(R)})$ (these represent the observed outcomes and the outcomes that would have been observed if not for missingness) conditional on the observables $(\mathbf{X}, \mathbf{Z}, R)$:

$$\begin{aligned} &\Pr\left(Y_1, \ldots, Y_T, O_1, \ldots, O_T, D_1, \ldots, D_T \mid \mathbf{X}, \mathbf{Z}, R\right) \quad = \\ &\Pr\left(Y_1, \ldots, Y_T \mid \mathbf{X}, \mathbf{Z}, R\right) \Pr\left(O_1, \ldots, O_T, D_1, \ldots, D_T \mid \mathbf{X}, \mathbf{Z}, R\right) \end{aligned} \tag{3}$$

Assumption (3) is a case of covariate-dependent drop-out in the terminology of [35]. Frangakis and Rubin [36] and Mealli et al. [37] provide alternative assumptions about missing data that allow for missingness to be nonignorable conditional on $(\mathbf{X}, \mathbf{Z}, R)$ but ignorable once partially unobserved compliance status is also conditioned on. We consider an alternative assumption that allows for a certain type of informative drop-out in Section 9.2.5.

4.5 Randomization Assumption

For an unclustered design, assignment to the treatment arm is assumed to be random and hence ignorable, i.e., letting $\mathbf{G}_i^{-R_i}$ denote the vector $\mathbf{G}_i$ of Section 3.1 excluding R_i,

$$\Pr(R_1, \ldots, R_n \mid \mathbf{G}_1^{-R_1}, \ldots, \mathbf{G}_n^{-R_n}) = \Pr(R_1, \ldots, R_n), \tag{4}$$

For a clustered design, assignment to the treatment arm is assumed to be random among clusters and hence ignorable when cluster membership is conditioned on,

$\Pr(R_{11}, \ldots, R_{1n_1}, \ldots, R_{n1}, \ldots, R_{nn_n} \mid \mathbf{G}_{11}^{-R_1}, \ldots, \mathbf{G}_{1n_1}^{-R_{1n_1}}, \ldots, \mathbf{G}_{n1}^{-R_{n1}}, \ldots, \mathbf{G}_{nn_n}^{-R_{nn}}, P_{11}, \ldots, P_{nn_n}) =$
$\Pr(R_{11}, \ldots, R_{1n_1}, \ldots, R_{n1}, \ldots, R_{nn_n} \mid P_{11}, \ldots, P_{nn_n})$;

see section 2 under clusters for our notation for clustered designs.

4.6 Monotoncity Assumption

The monotonicity assumption is that $A_{it}^{(1)} \geq A_{it}^{(0)}$ for all i and t, i.e., there are no patients who do the opposite of what they are assigned (i.e., no defiers). For the single consent design, the group assigned to the control arm does not have access to the treatment; thus, monotonicity holds by design.

4.7 Random Effects Assumptions

The random effects vector $\boldsymbol{\tau}$ is assumed to have mean zero conditional on compliance status. For the random effects distribution, we assume $\boldsymbol{\tau}|\mathbf{C}$ is iid $\mathbf{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\mathbf{MVN}$ denotes the multivariate normal distribution. We assume that the random effects design matrix $\mathbf{Z}_i$ contains only functions of the baseline covariates $\mathbf{X}_i$ and the time variables $\mathbf{T}$. For the depression study, we will focus on the case in which $\boldsymbol{\tau}_i$ contains only a random intercept $\tau_{0\,i}$, and consequently, $\boldsymbol{\Sigma} = \sigma_\tau^2$ and $\mathbf{Z}_i$ is a $T \times 1$ vector of ones. In Section 9.2.6, we examine multidimensional random effects vectors for the depression study and find that there is no significant evidence that a multidimensional random effect provides a better model than a random intercept.

We make a version of the usual conditional independence assumption for random effects models. We assume that conditional on the random effects $\boldsymbol{\tau}$, the observed covariates $(\mathbf{X}, \mathbf{Z})$, and the partially unobserved compliance statuses $\mathbf{C} = (C_1, \ldots, C_T)$, the potential outcomes corresponding to an arm r are independent,

$$\Pr(Y_1^{(r)} = y_1, \ldots, Y_T^{(r)} = y_T \mid \boldsymbol{\tau}, \mathbf{X}, \mathbf{Z}, \mathbf{C}) = \prod_{t=1}^{T} \Pr(Y_t^{(r)} = y_t \mid \boldsymbol{\tau}, \mathbf{X}, \mathbf{Z}, \mathbf{C}). \tag{5}$$

For clustered designs, we make an analogous assumption to (5) that involves both the patient level random effects $\boldsymbol{\tau}_{hi}$ and the cluster level random effects $\boldsymbol{\iota}_h$.

5. Model

The assumptions of Section 4 suffice to identify the intention to treat effect for compliers without any further parametric assumptions, following a similar argument to that of Imbens and Angrist [38]. However, to have interpretable parameters, it is useful to consider auxiliary parametric assumptions. The parametric model for potential outcomes we consider is a longitudinal random effects extension of the cross-sectional logistic model that was presented for treatment non-adherence in randomized trials by Nagelkerke et al. [19] and further investigated by Ten Have et al. [20]. The model is as follows:

$$\Pr(Y_t^{(r)} = 1 \mid \boldsymbol{\tau}, \mathbf{X}, R, \mathbf{C}, \mathbf{Z}) = \text{expit}(\boldsymbol{\tau}^T\mathbf{Z} + \boldsymbol{\alpha}^T\mathbf{T}_t + \boldsymbol{\beta}^T\mathbf{X} + \gamma_t C_t + \psi_t r C_t), \tag{6}$$

where $\text{expit}(.) = \exp(.)/[1 + \exp(.)]$. The model makes the assumption that the probability distribution of the potential outcomes $Y_t^{(0)}, Y_t^{(1)}$ at time t is independent of compliance status at times $1, \ldots, t-1, t+1, \ldots, T$ given compliance status at time t, i.e.,

$$Y_t^{(0)}, Y_t^{(1)} \perp\!\!\!\perp C_1, \ldots, C_{t-1}, C_{t+1}, \ldots, C_T | C_t. \tag{7}$$

This assumption is further discussed in Section 10.

We now discuss the interpretation of the parameters in (6) under the assumptions in Section 4. The parameter ψ_t is the log odds ratio comparing the effect of assignment to the treatment arm compared to assignment to the control arm on the outcome at time t among those patients who would adhere to the treatment at time t if assigned to the treatment arm, conditioning on the random effect $\boldsymbol{\tau}$:

$$\begin{aligned} \psi_t \;=\; & \text{logit}\left[\Pr(Y_t^{(1)} = 1 \mid \boldsymbol{\tau}, \mathbf{X}, \mathbf{Z}, C_t = 1, \mathbf{C})\right] \\ & -\text{logit}\left[\Pr(Y_t^{(0)} = 1 \mid \boldsymbol{\tau}, \mathbf{X}, \mathbf{Z}, C_t = 1, \mathbf{C})\right] \end{aligned} \tag{8}$$

In other words, ψ_t is an intention to treat effect for compliers at time t. Because compliers receive the treatment if assigned to the treatment arm and do not receive the treatment if assigned to the control arm, the intention to treat effect for compliers ψ_t can under certain conditions be interpreted as the efficacy of treatment received for compliers at time t [39, 3]; see Section 10.1

for further discussion of the interpretation of ψ_t. Because the group assigned to the control arm does not have access to the treatment in the single consent design, the intention to treat log odds ratio ψ_t for those patients who would adhere to the treatment at time t if assigned to the treatment arm ($C_{it} = 1$) equals the intention to treat log odds ratio for those patients who actually receive the treatment at time t ($A_{it} = 1$) [40]. The parameter γ_t is a log odds ratio parameter for compliance status at time t that reflects how outcomes between compliers and never takers at time t would compare if both groups were assigned to the control arm. The parameter $\boldsymbol{\alpha}$ is a vector of fixed effects for the time variables $\mathbf{T}_t$ and the parameter $\boldsymbol{\beta}$ is a vector of the fixed effects log odds ratio parameters for the baseline covariates. Because of the exclusion restriction for never takers assumption (2), we do not include a parameter for $r(1 - C_t)$ in the model (6).

Under the SUTVA consistency assumption of Section 4.2, $\Pr(Y_t^{(r)} = Y_t \mid R = r) = 1$. Given this SUTVA consistency assumption, the sampling assumptions in Section 4.1 and the missing data assumptions in Section 4.4, the model (6) produces the following model for the observed outcomes:

$$\begin{aligned}&\Pr(Y_t = 1 \mid \boldsymbol{\tau}_i, \boldsymbol{X}_i, \mathbf{A}_i, R_i, \mathbf{C}_i, \mathbf{Z}_i)\\ &\qquad = \quad \text{expit}\left(\boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t C_{it} + \psi_t A_{it}\right). \qquad (9)\end{aligned}$$

6. Estimation

Under the missing data assumption (3), the other assumptions in Section 4 and the model (9), the likelihood function for an unclustered design conditioning on $\mathbf{X}_i, R_i, \mathbf{A}_i, \mathbf{Z}_i$ is the following:

$$\prod_{i=1}^{N} \int \sum_{(c_1,\ldots,c_{T_i})} \omega_i(c_1, \ldots, c_{T_i}) \prod_{t=1}^{T_i} (\pi_{Y_{it},c_t}(\boldsymbol{\tau}))^{Y_{it}} (1 - \pi_{Y_{it},c_t}(\boldsymbol{\tau}))^{1-Y_{it}} f_{\boldsymbol{\tau}}(\boldsymbol{\tau} \mid \Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau}, \qquad (10)$$

where $\omega_i(c_1, \ldots, c_{T_i}) = \Pr(C_1 = c_1, \ldots, C_{T_i} = c_{T_i} \mid \mathbf{X}_i, R_i, \mathbf{A}_i, \mathbf{Z}_i)$; $\pi_{Y_{it},c_t}(\boldsymbol{\tau}) = \Pr(Y_t = 1 \mid \boldsymbol{\tau}, \mathbf{X}_i, R_i, \mathbf{A}_i, \mathbf{Z}_i, C_t = c_t)$; $\sum_{(c_1,\ldots,c_{T_i})}$ is the sum over all 2^{T_i} compliance patterns for patient i; and $f(\boldsymbol{\tau}|\Sigma_{\boldsymbol{\tau}})$ is the normal density with covariance matrix $\Sigma_{\boldsymbol{\tau}}$ [Note: In (10) and all subsequent likelihood expressions, $\prod_{t=1}^{T_i}$ denotes the product over all observations for patient i that are not missing]. Note that for $R_i = 0$, $\omega_i(c_1, \ldots, c_{T_i}) = \Pr(A_1 = c_1, \ldots, A_{T_i} = c_{T_i} \mid \mathbf{X}_i, R = 1, \mathbf{Z}_i)$ and for $R_i = 1$, $\omega_i(A_{i1}, \ldots, A_{iT_i}) = 1$. Thus, letting $\kappa_i(a_1, \ldots, a_{T_i}) = \Pr(A_1 = a_1, \ldots, A_{T_i} = a_{T_i} \mid$

$\mathbf{X}_i, R=1, \mathbf{Z}_i)$, the likelihood function for an unclustered design conditioning on $\mathbf{X}_i, R_i, \mathbf{Z}_i$ is the following:

$$\left[\prod_{i=1|R_i=1}^{n} \int \kappa_i(A_{i1},\ldots,A_{iT_i}) \prod_{t=1}^{T_i} (\pi_{Y_{it},A_{it}}(\boldsymbol{\tau}))^{Y_{it}} (1-\pi_{Y_{it},A_{it}}(\boldsymbol{\tau}))^{1-Y_{it}} f_{\boldsymbol{\tau}}(\boldsymbol{\tau} \mid \Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau}\right] \quad \times$$

$$\left[\prod_{i=1|R_i=0}^{n} \int \sum_{(c_1,\ldots,c_{T_i})} \kappa_i(c_1,\ldots,c_{T_i}) \prod_{t=1}^{T_i} (\pi_{Y_{it},c_t}(\boldsymbol{\tau}))^{Y_{it}} (1-\pi_{Y_{it},c_t}(\boldsymbol{\tau}))^{1-Y_{it}} f_{\boldsymbol{\tau}}(\boldsymbol{\tau} \mid \Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau}\right] \qquad (11)$$

Given a model for $\Pr(A_1 = a_1, \ldots, A_{T_i} = a_{T_i} \mid \mathbf{X}_i, R=1, \mathbf{Z}_i)$, (11) can be maximized using approximate maximum likelihood methods, such as Gaussian quadrature or Monte Carlo EM. However, such methods are not easily implemented using standard software. We focus in this paper on an approximate IV estimation method that is easily implemented using standard random effects logistic regression software. Our approximate IV method is a random effects extension of the approximate IV approach of Nagelkerke et al. [19] for cross-sectional logistic models.

6.1 Approximate Instrumental Variables Estimation

To motivate our approach, first consider a linear version of model (9) for Y_t:

$$E(Y_t \mid \boldsymbol{\tau}_i, \mathbf{X}_i, \mathbf{A}_i, R_i, \mathbf{C}_i, \mathbf{Z}_i) = \boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t C_{it} + \psi_t A_{it}. \qquad (12)$$

Letting $W_{it} = R_i[A_{it} - E(A_t \mid \mathbf{X}_i, R=1)]$ and u_{it} be a mean zero error term, we have

$$\begin{aligned} Y_{it} &= \boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t C_{it} + \psi_t A_{it} + u_{it} \\ &= \boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t W_{it} + \psi_t A_{it} + \gamma_t (C_{it} - W_{it}) + u_{it}. \end{aligned}$$

Note that for patients i with $R_i = 1$, $W_{it} = A_{it} - E(A_t \mid \mathbf{X}_i, R=1)$ and $C_{it} - W_{it} = E(A_t \mid \mathbf{X}_i, R=1)$ and for patients i with $R_i = 0$, $W_{it} = 0$ and $C_{it} - W_{it} = C_{it}$. Consequently, 1) $C_{it} - W_{it}$ is uncorrelated with A_{it} conditional on W_{it}, $\mathbf{X}_i$ and R_i and 2) $C_{it} - W_{it}$ is uncorrelated with W_{it} conditional on A_{it}, $\mathbf{X}_i$ and R_i. A basic property of the linear regression model is that if $E(Y|X_1,\ldots,X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ and $Cov(X_p, X_{p-1} \mid X_1,\ldots,X_{p-2}) = 0$, then $E(Y|X_1,\ldots,X_{p-1}) = \beta_0^* + \beta_1^* X_1 + \ldots + \beta_{p-2}^* X_{p-2} + \beta_{p-1} X_{p-1}$. Therefore, we can obtain consistent estimates of the coefficients $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_T)$ and $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_T)$ in (12) by fitting a uniform correlation linear mixed effects model of Y_{it} on the fixed effects $\mathbf{T}_t$, $\mathbf{X}_i$, W_{it} and A_{it}

with random effects design matrix $\mathbf{Z}_i$. Note that the standard errors from such a linear mixed effects model may not be accurate because

$$\gamma_1(C_{i1} - W_{i1}) + u_{i1}, \ldots, \gamma_{T_i}(C_{iT_i} - W_{iT_i}) + u_{iT_i}, \tag{13}$$

are not necessarily independent as they are assumed to be in the uniform correlation mixed effects model. Note also that the missing data assumption (3) implies that missingness at time t is independent of Y_t conditional on $W_t, \mathbf{A}, \mathbf{X}, \mathbf{Z}$ (conditioning on $W_t, \mathbf{A}, \mathbf{X}, \mathbf{Z}$ is equivalent to conditioning on $R, \mathbf{A}, \mathbf{X}, \mathbf{Z}$). This property of the missing data implies that the above approach provides consistent estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$ in the presence of missing data.

We call the above approach an instrumental variables approach because the randomization indicator is used as an "instrument" to extract variation in A_{it} that is unrelated to omitted confounding variables associated with adherence (the extracted variation is the variation in A_{it} due to R_i) and this variation is used to obtain a consistent estimate of $\boldsymbol{\psi}$ (for an overview of IV methods, see [9]). In Nagelkerke *et al.*'s (2000) graph theory explanation, W_{it} "intercepts" the effect of omitted confounding variables associated with adherence, permitting consistent estimation of $\boldsymbol{\psi}$.

Our approximate IV estimation method extends the above approach to the logistic regression model. Y_{it} can be represented in the following way based on the model (9):

$$\begin{aligned} Y_{it} &= I(\boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t C_{it} + \psi_t A_{it} + u_{it} > 0) \\ &= I(\boldsymbol{\tau}_i^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_i + \gamma_t W_{it} + \psi_t A_{it} + \gamma_t(C_{it} - W_{it}) + u_{it} > 0), \end{aligned} \tag{14}$$

where u_{it} has a logistic distribution. Our estimation method is to fit a logistic mixed effects regression model of Y_{it} on the fixed effects $\mathbf{T}_t$, $\mathbf{X}_i$, W_{it} and A_{it} with random effects design matrix $\mathbf{Z}_i$. We call this estimation method an "approximate" IV approach because, for the logistic regression model (9), this method does not necessarily produce consistent estimates as in the linear regression model (12). The difficulty in the logistic regression model is that even though $C_{it} - W_{it}$ is uncorrelated with A_{it} conditional on W_{it}, $\mathbf{X}_i$, $\mathbf{T}_t$ and R_i, the association between Y_{it} and A_{it} conditional on W_{it} $\mathbf{X}_i$, $\mathbf{T}_t$, R_i and $C_{it} - W_{it}$ that is measured by the logistic regression coefficient ψ_t is not generally equal to the association between Y_{it} and A_{it} conditional on W_{it}, $\mathbf{X}_i$, $\mathbf{T}_t$ and R_i [41, 42]. To gain insight into the difference between these

two associations, we cite some results for the related simpler setting of a logistic regression model $E(Y \mid T, X) = \text{expit}(\theta_0 + \theta_1 T + \theta_2 X)$ for which $Cov(T, X) = 0$. Guo and Geng [42] show that $E(Y \mid T) = \text{expit}(\theta_0' + \theta_1 T)$ if $\theta_2 = 0$ and Gail et al. [41] show that the asymptotic bias in using logistic regression of Y on T to estimate θ_1 is proportional to θ_2^2 multiplied by a function of θ_0 and θ_1 for θ_2 near zero. These results suggest that the magnitude of the bias in estimating the coefficients in (9) by using the approximate IV approach of logistic mixed effects regression of Y_{it} on fixed effects $\mathbf{T}_t, \mathbf{X}_i, W_{it}$, and A_{it} with random effects design matrix $\mathbf{Z}_i$ should be small for γ_t near zero and increase for γ_t of larger magnitude. For the simulation based on the depression study data in Section 8, we find that in fact there is only small bias for $\gamma_t = -0.5$ and $\gamma_t = -1$ but there is somewhat larger bias for $\gamma_t = -2$. Nagelkerke et al. [19] and Ten Have et al. [20] showed through simulations that the cross sectional version of this approximate IV approach exhibits good bias and confidence interval coverage properties for a range of levels of unmeasured confounding due to non-adherence (γ_t). However, under strong confounding (γ_t has large magnitude), the bias and confidence interval coverage deteriorate.

To implement the above approximate IV approach, we need to know $W_{it} = R_i[A_{it} - E(A_t \mid \mathbf{X}_i, R = 1)]$. We follow the usual instrumental variables approach and substitute an estimate $\hat{W}_{it} = R_i[A_{it} - \hat{E}(A_t \mid \mathbf{X}_i, R = 1)]$ for W_{it}. We estimate $E(A_t \mid \mathbf{X}_i, R = 1)$ using a logistic regression model fitted to treatment-received in the randomized to treatment group:

$$\Pr(A_t = 1 \mid \mathbf{X}_i, R_i = 1) \quad = \quad \text{expit}\left(\boldsymbol{\kappa}^T \mathbf{T}_t + \boldsymbol{\xi}^T \mathbf{X}_i\right) . \tag{15}$$

Our approximate IV estimator is then obtained by maximizing the following random effects logistic regression likelihood function over the parameters $\Sigma_\tau^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\psi}^*$:

$$\prod_{i=1}^{n} \int \prod_{t=1}^{T_i} (\pi_{Y_{it}}(\boldsymbol{\tau}^*))^{Y_{it}} (1 - \pi_{Y_{it}}(\boldsymbol{\tau}^*))^{1-Y_{it}} f_{\boldsymbol{\tau}^*}(\boldsymbol{\tau}^* \mid \Sigma_{\boldsymbol{\tau}}^*) d\boldsymbol{\tau}^*, \tag{16}$$

where $\pi_{Y_{it}}(\boldsymbol{\tau}^*) = \text{expit}(\boldsymbol{\tau}^{*T} \mathbf{Z}_{it} + \boldsymbol{\alpha}^{*T} \mathbf{T}_t + \boldsymbol{\beta}^{*T} \mathbf{X}_i + \gamma_t^* \hat{W}_{it} + \psi_t^* A_{it})$ and $f(\boldsymbol{\tau}^* \mid \Sigma_{\boldsymbol{\tau}}^*)$ is the density $N(\mathbf{0}, \Sigma_{\boldsymbol{\tau}^*})$. We estimate $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_T), \boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_T)$ by our estimates $\hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\psi}}^*$ from (16). The integration was performed with the non-adaptive quadrature facility in SAS PROC NLMIXED with 20 quadrature points. SAS macros for our estimation approach are available from the authors. For models in which there are only random intercepts, such as the one we fit to the depression study data, our estimation approach can also be easily implemented in

STATA (using xtlogit) and R (using the glmmML package). We use a quadrature method to approximately maximize (16) rather than a method based on Laplace approximations because there are few observations per subject, a setting for which Laplace approximation methods can work poorly [43]; see Section 8.2 for further discussion. A technical report available from the authors motivates our approximate IV estimator in a different way by showing that (16) is an approximation to the likelihood (10).

Note that $\Sigma_{\boldsymbol{\tau}^*}$ in (16) does not correspond to $\Sigma_{\boldsymbol{\tau}}$ in (10) when $C_{i1} - W_{i1}, \ldots C_{iT} - W_{iT}$ are correlated (see discussion below (13)). To obtain an estimate of $\Sigma_{\boldsymbol{\tau}}$, we consider the conditional likelihood for the subset of randomized to treatment arm patients, conditioning on $(A_{i1}, \ldots, A_{iT_i})$:

$$\prod_{i=1|R_i=1}^{N} \int (\pi_{Y_{it}}(\boldsymbol{\tau}))^{Y_{it}} (1 - \pi_{Y_{it}}(\boldsymbol{\tau}))^{1-Y_{it}} f(\boldsymbol{\tau}|\Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau}, \tag{17}$$

where $\pi_{Y_{it}}(\boldsymbol{\tau}) = \text{expit}(\boldsymbol{\tau}^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T T_t + \boldsymbol{\beta}^T \mathbf{X}_i + \lambda_t A_{it})$; $\lambda_t = \gamma_t + \psi_t$; and $f(\boldsymbol{\tau} \mid \Sigma_{\boldsymbol{\tau}})$ is the normal density with covariance matrix $\Sigma_{\boldsymbol{\tau}}$. Thus, $\Sigma_{\boldsymbol{\tau}}$ can be estimated by approximately maximizing the conditional likelihood (17) using quadrature.

6.2 Model and Estimation for Clustered Encouragement Design

To account for clustering of outcomes, we use the following probability model in place of (6):

$$\begin{aligned} &\Pr(Y_t^{(r)} \mid \boldsymbol{\tau}_{hi}, \boldsymbol{\iota}_h, \mathbf{X}_{hi}, R_{hi}, \mathbf{C}_{hi}, \mathbf{Z}_{hi}, \mathbf{V}_h, P_{hi}) \quad = \\ &\text{expit}(\boldsymbol{\tau}_{hit}^T \mathbf{Z}_{hit} + \boldsymbol{\iota}_h^T \mathbf{V}_{ht} + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X}_{hi} + \gamma_t C_{hit} + \psi_t r C_{hit}) \end{aligned} \tag{18}$$

We can use the approximate IV approach of Section 6.1 to estimate $\boldsymbol{\psi}$ by fitting a logistic mixed effects regression model of Y_{hit} on fixed effects $\mathbf{T}_t$, $\mathbf{X}_{hi}$, $\hat{W}_{hit}$ and A_{hit} with nested random effects $\boldsymbol{\tau}_{hi}$ (with design matrix $\mathbf{Z}_{hi}$) and $\boldsymbol{\iota}_h$ (with design matrix $\mathbf{V}_h$). The corresponding approximate marginal likelihood function to (16) is

$$\prod_{h=1}^{n} \int \prod_{i=1}^{n_h} \int \prod_{t=1}^{T_i} (\hat{\pi}_{Y_{hit}}(\boldsymbol{\tau}, \boldsymbol{\iota}))^{Y_{hit}} (1 - \hat{\pi}_{Y_{hit}}(\boldsymbol{\tau}, \boldsymbol{\iota}))^{1-Y_{hit}} f_{\boldsymbol{\tau}^*}(\boldsymbol{\tau}^* \mid \Sigma_{\boldsymbol{\tau}^*}) f^*_{\boldsymbol{\iota}}(\boldsymbol{\iota}^* \mid \Sigma_{\boldsymbol{\iota}^*}) d\boldsymbol{\tau}^* d\boldsymbol{\iota}^*, \tag{19}$$

where $\hat{\pi}_{Y_{hit}}(\boldsymbol{\tau}) = \text{expit}(\boldsymbol{\tau}_i^{*T} \mathbf{Z}_{hit} + \boldsymbol{\iota}_h^{*T} \mathbf{V}_{ht} + \boldsymbol{\alpha}^{*T} \mathbf{T}_t + \boldsymbol{\beta}^{*T} \mathbf{X}_{hi} + \gamma_t^* \hat{W}_{hit} + \psi_t^* A_{hit})$. (19) cannot be maximized using proc nlmixed in SAS or glmmML in R because it involves nested random effects. (19) can be approximately maximized by Breslow and Clayton [44]'s "penalized quasilikelihood," which is based on Laplace approximations and is implemented in glimmix in SAS and glmmPQL

in R. As we shall see in Section 7, because there is only a small amount of clustering for the depression study and there are at most three observations per subject (which makes the Laplace approximations perform poorly), the estimation method which ignores clustering and uses quadrature to approximately maximize (16) works better than the estimation method which accounts for clustering and uses Laplace approximations to approximately maximize (19).

7. Assessment of the Validity of the IV Analyses

We based conclusions about the validity of the IV analysis for the depression study on six steps: 1) a simulation study to determine whether our approximate IV method produces accurate results for the setting of the depression study; 2) analysis of the sensitivity of the conclusions to the exclusion restriction assumption by including pre-specified direct randomization effects in (6). 3) analyses of the sensitivity to the normal random effects assumption by varying the number of quadrature points; 4) assessment of the assumption (7); 5) assessment of missing data assumptions; and 6) analysis of the sensitivity of conclusions to alternative specifications of the random vector $\boldsymbol{\tau}_i$.

For the simulation study for the step 1 assessment of validity, we generated data from the model for outcomes from Sections 3-4, with probability distribution (9) and $\boldsymbol{\tau}_i = \tau_{0i}$, and the following random effects model for compliance:

$$\Pr(C_{i1} = c_{i1}, \ldots, C_{iT} = c_{iT} \mid \mathbf{X}_i, \eta_{0i}) = \prod_{t=1}^{T_i} \text{expit}(\boldsymbol{\kappa}^T \mathbf{T}_t + \boldsymbol{\xi}^T \mathbf{X}_i + \eta_{0i}), \tag{20}$$

where $\eta_{0i} \sim N(0, \sigma_\eta^2)$. The random effects η_{0i} and τ_{0i} are assumed to be uncorrelated in accordance with our interpretation of τ_{0i} as a mean zero random effect conditional on compliance status (see Section 4.7). To account for clustering, a simulation study was also carried out using the probability model (18) for potential outcomes.

8. Simulations

In order to assess the accuracy of our approximate IV approach for the setting of the depression study, we performed simulations under the model in Section 7 using parameters estimated from the depression study. The parameters in the simulation model were set based on estimates from maximizing (16), using (15) to estimate $\Pr(A_t = 1 \mid \mathbf{X}_i, R = 1)$, for the depression study data. The variance component σ_τ was set based on the estimate from maximizing (17). For all simulations reported, we varied γ_t in the set $\gamma_t \in \{-0.5, -1.0, -2.0\}$. Varying γ_t changes the

strength of the confounding due to treatment non-adherence.

For the outcome model, the following parameters were specified: 1) the variance component of the random intercept $\tau_{0\,i}$, $\sigma_\tau = 2.0$; 2) $\boldsymbol{\alpha}^T = (-0.5, -0.2, -1)$ for the dummy variables corresponding to 4, 8 and 12 month visits respectively; 3) $\psi_1 = 1.0$, $\psi_2 = 0.9$, $\psi_3 = 1.0$; and 4) the model has no baseline covariates. For the compliance model in (20) with a random intercept $\eta_{0\,i}$, the following parameters were specified: 1) the variance component of $\eta_{0\,i}$, $\sigma_\eta = 7.0$; 2) $\boldsymbol{\kappa}^T = (3.78, 2.65, 3)$ for the dummy variables corresponding to 4, 8 and 12 month visits respectively; and 3) the model has no baseline covariates. The number of patients at baseline was 500, approximately the same sample size as the depression study. Additionally, the following design-related probabilities were specified: randomization was 0.5 and drop-out during each period t was 0.10. For each setting considered, the number of simulations done was 1000. Simulation results for the log odds ratio at time 12 months are presented for three estimation approaches: 1) ITT comparison between the randomized to treatment group and the randomized to control groups using a random effects logistic model; 2) AT comparison between the group that actually received the treatment and the group that did not receive the treatment using a random effects logistic model; and 3) the approximate IV approach estimate of ψ_3 described in Section 6.

For each of these estimation approaches, the following simulation statistics averaged over 1000 iterations are presented in Table 1 with respect to true $\psi_3 = 1.0$: 1) mean ITT, AT, and IV estimates and 2) the proportion of times the approximate 95% confidence interval covers $\psi_3 = 1.0$ (labeled coverage). Another set of 1000 iterations with true $\psi_3 = 0.0$ was run and the proportion of times the approximate 95% confidence interval covers $\psi_3 = 0.0$ is reported (labeled size – this is the size of the nominal $\alpha = 0.05$ test of $\psi_3 = 0$).

8.1 Simulations from Model without Clustering by Practice

Table 1 shows the results of simulations with no clustering by practice. For stronger confounding due to non-adherence ($\gamma_t \in \{-1.0, -2.0\}$), the AT estimates of ψ_3 are of smaller magnitude than the corresponding ITT and IV estimates as in the depression study results in Table 3. The approximate IV approach has reasonable coverage (between 92% and 96%) and size (between 0.04 and 0.07) for all all three levels of γ_t. The test size and coverage are worst for the strongest level of confounding ($\gamma_t = -2$). The bias of the IV estimator is reasonably

small (-0.04 and -0.06) for the weaker levels of confounding ($\gamma_t = -0.5$ and -1.0 respectively) but is more substantial (-0.21) for the strongest level of confounding. These results about the approximate IV approach performing best for a small magnitude of confounding due to non-adherence are consistent with the analysis of Section 6 and the results of Nagelkerke et al. [19] and Ten Have et al. [20] for the cross-sectional logistic case.

For all levels of confounding, the bias of the IV estimates for the estimand $\psi_3 = 1.0$ is much less than the ITT and AT estimates and the mean square error of the IV estimator is smaller than the ITT and AT estimates. Furthermore, 95% confidence interval coverage and size of the nominal $\alpha = 0.05$ test are much better for the approximate IV approach than for the ITT and AT approaches. Whereas for the IV estimator, the coverage is between 92 and 96% and the size is between 0.04 and 0.07, for the AT method, the coverage is between 7 and 85% and the size is between 0.14 and 0.89. For the ITT estimator, the size is 0.05 but coverage does not exceed 81%.

The IV estimates of γ_t also perform adequately, although less well than the estimates of ψ_t. There is a positive bias in the estimates of γ_3 of $0.16, 0.32$ and 0.73 for $\gamma_3 = -0.5, -1.0$ and -2.0 respectively. The coverage of 95% confidence intervals for γ_3 are 0.94, 0.92 and 0.82 for $\gamma_3 = -0.5, -1.0$ and -2.0 respectively.

Table 1. Simulation results for unclustered design: mean parameter estimate, mean squared error (MSE) and coverage of 95% confidence interval for true $\psi_3 = 1.0$; and size of test of $\psi_3 = 0$ for true $\psi_3 = 0$.

γ_t	-0.5			-1.0			-2.0		
Statistic	ITT	AT	IV	ITT	AT	IV	ITT	AT	IV
Mean Est.	0.64	0.69	0.96	0.62	0.39	0.94	0.54	0.30	0.79
MSE	0.24	0.21	0.19	0.26	0.49	0.20	0.35	1.84	0.28
Coverage	0.81	0.85	0.96	0.78	0.60	0.95	0.76	0.07	0.92
Size	0.05	0.14	0.04	0.05	0.37	0.05	0.05	0.89	0.07

8.2. Simulations with Clustering by Practice

To estimate the amount of clustering of outcomes by practice for the depression study, we used Breslow and Clayton [44]'s penalized quasi-likelihood (PQL) via the glmmPQL function in R to approximately maximize the analogue of (17) for a clustered design with a random intercept ι_{0q} for practice. We estimate that $\sigma_\iota = 0.46$ with a 95% confidence interval of $(0.23, 0.91)$. To estimate the amount of clustering of treatment received by practice for the depression study, we used glmmPQL to estimate a logistic mixed effects model for treatment received for the randomized to treatment practices with random intercepts for practices and individual patients. We estimate the standard deviation of the random intercept for practice to be 1.12 with a 95% confidence interval of $(0.50, 2.53)$. We consider two estimation methods for a clustered design: (1) approximately maximize (16) by quadrature (described as the quadrature method below) and (2) approximately maximize (19) by penalized quasi-likelihood (described as the PQL method below). To examine the performance of these two estimation methods for the setting of the depression study, we simulated compliance statuses from a mixed effects logistic compliance model with practice and patient random intercepts and outcomes from the probability model (18), with parameters for both models based on estimates from the depression study. For the compliance model, we used the same settings as described at the beginning of Section 8 and a standard deviation of 1.12 for the practice random intercepts. For the outcome model, we used the same settings as described at the beginning of Section 8 and a standard deviation of $\sigma_\iota = 0.46$ for the practice random intercepts.

Table 2 compares the performance of the quadrature and PQL methods for the simulation study. The first feature of note is that the quadrature method (1)'s performance remains reasonable and does not deteriorate in terms of bias and coverage compared to Table 1. The second feature of note is that the quadrature method performs better than the PQL method in terms of bias and confidence interval coverage for all three settings of γ_t, even though the quadrature method does not take into account the practice level clustering. This is partly because the amount of practice level clustering is small, but we found that even for a larger amount of practice clustering, $\sigma_\iota = 2.0$, the quadrature method remained better. The poor performance of PQL for the setting of the depression study is likely related to the presence of small cluster

sizes for which the Laplace approximations underlying the PQL method are inaccurate [43] (the nested clusters of each patient's observations are small, of size at most three, and some of the practice clusters are small). In light of the results in Table 2, we used the quadrature method for the data analysis of the depression study.

Table 2. Simulation results for a clustered design setting similar to the depression study, with practice level random effects: mean parameter estimate for $\psi_3 = 1.0$; mean squared error (MSE); and coverage of 95% confidence interval for estimation method Quad that uses quadrature and ignores clustering and estimation method PQL that takes into account clustering and uses penalized quasi-likelihood.

γ_t	-0.5		-1.0		-2.0	
Estimate	Quad	PQL	Quad	PQL	Quad	PQL
Mean Estimate	1.02	0.82	0.97	0.79	0.84	0.69
MSE	0.23	0.20	0.26	0.23	0.31	0.31
Coverage	0.95	0.90	0.94	0.88	0.93	0.85

9.0 Results for Depression Study

This section presents data analysis results for the depression study described in Section 2. Section 9.1 presents the results of the IV analysis for the study and compares the IV analysis to ITT and AT analyses. Section 9.2 assesses some of the assumptions of the IV approach as described in Section 7.

9.1 IV Analysis and Comparisons to ITT, AT

The IV analysis was carried out using the approximate IV estimation method of Section 6 with $\mathbf{X}_i$ comprised of baseline Hamilton and suicide ideation score and $\mathbf{T}_t$ comprised of dummy variables for the time of the visit (4, 8 or 12 months). Table 3 shows the IV, ITT and AT log odds ratio estimates for random effects logistic models. From the IV analysis, there is strong evidence that contact with a depression specialist is efficacious for compliers – p-values = 0.001, 0.02 and 0.01 for 4, 8 and 12 months respectively. Furthermore, the IV analysis estimates that the efficacy for compliers is substantial – the estimated efficacy odds ratio for compliers (i.e.,

the intent to treat odds ratio for compliers) is 2.97, 2.44 and 2.66 for 4, 8 and 12 months. The point estimates from the IV analysis are quite similar for all three time periods, whereas those from the other methods vary substantially over the time periods. Under the assumed exclusion restriction and model assumptions, the IV analysis provides a better picture of how the efficacy of the treatment varies over time because it does not incorporate changes in compliance over time, as does the ITT analysis, or changes in the confounding due to nonadherence over time, as does the AT analysis.

A comparison of the three sets of estimates shows the advantages of IV over AT as an estimate of efficacy under the assumed exclusion restriction. Under the exclusion restriction, the efficacy of treatment received for compliers should be of at least as large a magnitude as the programmatic effectiveness of offering treatment that is estimated by ITT. The IV estimates are in fact of larger magnitude than the ITT estimates for all three time periods. However, the AT estimates are of smaller magnitude than the ITT estimates for 4 and 8 months. Also, under the exclusion restriction, the p-values for the efficacy and ITT tests of zero treatment effect should be similar [45, 19, 8]. The ITT and IV tests of zero treatment effect in fact give similar p-values at all three time points, but there is a substantial difference between the AT and ITT p-values at 8 months.

The AT estimate is not an accurate estimate of efficacy because it is confounded by omitted variables associated with non-adherence. The relationship between the AT and IV estimates suggests that compliers are less likely than never takers, conditional on baseline covariates, to have their Hamilton score fall by 50% or more from baseline if both groups are given usual care. The estimates of γ_t, which represents the confounding in the AT estimate due to non-adherence at time t, are -0.44, -1.11 and -0.44 for 4, 8 and 12 months respectively.

9.2 Assessment of the Validity of the IV Analyses

We employed the six steps listed in Section 6.0 for performing this assessment

9.2.1 Simulation study to assess accuracy of approximate IV method

A simulation study from the same setting as the depression study that used the parameter estimates from the depression study was described in Section 8. The confidence interval coverage and test size for the approximate IV approach based on approximately maximizing (16) by

Table 3. Visit-specific depression specialist vs. usual care log odds ratios for the depression study with standard errors and p-values in parentheses.

Month	ITT	AT	IV
4	1.01	0.94	1.09
	(0.31; .001)	(0.30; .001)	(0.33; .001)
8	0.74	0.49	0.89
	(0.31; .02)	(0.30; .11)	(0.38; .02)
12	0.73	0.79	0.98
	(0.32; .02)	(0.31; .01)	(0.40; .01)

quadrature were found to be reasonable for the levels of confounding $\gamma_t \in \{-0.5, -1.0, -2.0\}$. The bias was found to be negative for all γ_t in this set with a small magnitude of bias for $\gamma_t \in \{-0.5, -1.0\}$ but a somewhat larger magnitude of bias (-0.20) for $\gamma_t = -2.0$. The estimates of γ_t for the depression study are -0.44, -1.11 and -0.44 with 95% confidence intervals $(-2.04, 1.17)$, $(-2.36, 0.14)$ and $(-1.69, 0.82)$ for the time points 4, 8 and 12 months respectively. A level of confounding due to non-adherence as large in magnitude as $\gamma_t = -2$ (which would mean that the odds ratio comparing compliers to never takers when both are assigned to usual care is 0.14) is considered unlikely by the clinical researchers conducting the study, especially when compared to odds ratios of smaller magnitude for treatment, time and baseline effects, which do not exceed 1.1 on the log scale. Thus, the simulation study provides evidence that the approximate IV approach is reasonably accurate under the model assumptions for the depression study because 1) the simulation study shows that the approximate IV approach is reasonably accurate for $\gamma_t \in \{-0.5, -1.0, -2.0\}$ and 2) the data analysis and a priori beliefs suggest that $|\gamma_t| \leq 2.0$.

9.2.2 Sensitivity to exclusion restriction

An important assumption in model (6) is the exclusion restriction for never takers. A model for the outcomes that departs from the exclusion restriction for never takers is

$$\Pr(Y_t^{(r)} = 1 \mid \boldsymbol{\tau}, \mathbf{X}, R, \mathbf{C}, \mathbf{Z})$$

$$= \quad \text{expit}\left(\boldsymbol{\tau}^T \mathbf{Z}_t + \boldsymbol{\alpha}^T \mathbf{T}_t + \boldsymbol{\beta}^T \mathbf{X} + \gamma_t C_t + \psi_t r C_t + \phi_t r(1 - C_t)\right). \tag{21}$$

The parameter ϕ_t represents the direct effect of randomization at time t for never takers; the exclusion restriction assumes $\phi_t = 0$. For pre-specified ϕ_t, maximizing (16) with $\pi_{Yit}(\boldsymbol{\tau}^*) = \text{expit}(\boldsymbol{\tau}^{*T}\mathbf{Z}_{it} + \boldsymbol{\alpha}^{*T}\mathbf{T}_t + \boldsymbol{\beta}^{*T}\mathbf{X}_i + \gamma_t^*\hat{W}_{it} + \psi_t^* A_{it} + \phi_t R_i(1 - A_{it}))$ provides estimates of ψ_t with the same properties as our approximate IV approach under the assumption that ϕ_t is specified correctly. To examine the sensitivity of our estimates to the exclusion restriction assumption, we considered prespecified values of the direct randomization parameter ϕ_t of 0.10 and 0.50; 0.50 is a substantial direct randomization effect when compared to the original IV estimates of ψ_t in Table 3 that range from 0.89 to 1.09. For $\phi_t = 0.1$, the new estimates of ψ_1, ψ_2 and ψ_3 are 1.08, 0.86 and 0.96 respectively, a drop of between 2 to 3% from the estimates in Table 3. For $\phi_t = 0.5$, the new estimates of ψ_1, ψ_2 and ψ_3 are 1.05, 0.77 and 0.86 respectively, a drop of between 4 to 13% from the estimates in Table 3. Note that the sensitivity of the estimates to violations of the exclusion restriction is higher for the time periods with higher rates of nonadherence (8 and 12 months). By looking at models (6) and (21), we see that, in general, the sensitivity of the estimates based on model (6) to violations of the exclusion restriction will be higher when the rate of non-adherence is higher. For the depression study, the approximate IV results are not highly sensitive to plausible departures from the exclusion restriction.

We have prespecified the parameter ϕ_t in (21) but ϕ_t can actually be estimated by making it a free parameter. However, such estimates have large standard errors, e.g., the estimate of ϕ_1 has a standard error that is more than 34 times as large as that of the estimate of ψ_1 from model (6). In addition, inferences about ϕ_t may be highly sensitive to the assumed logistic link (see [46] for discussion of this type of sensitivity for a nonrandom sampling model).

9.2.3 Assessment of Normal Random Effects Assumption

As a sensitivity analysis of the assumption of a normal random effects distribution, we varied the number of quadrature points from 5 to 30. Doing so altered the shape of the random effects distribution away from normality [47]. The IV estimates based on only 10 quadrature points differed by less than 2% from the IV estimates based on 20 points in Table 3. Increasing the number of quadrature points beyond 20 to 30 did not alter the results beyond a few percentage

points. Reducing the number of quadrature points to below 10 resulted in dramatic changes in estimates.

9.2.4 Assessment of the Form of Dependence of Outcomes on Compliance Status Vector Assumption

The assumption (7) can be tested by nesting it within the following model for $\Pr(Y_t^{(r)})$ that accommodates departures from assumption (7):

$$\begin{aligned}
\Pr(Y_1^{(r)} \mid \boldsymbol{\tau}, \mathbf{X}, R, \mathbf{C}, \mathbf{Z}) &= \text{expit}(\boldsymbol{\tau}^T \mathbf{Z}_t + \boldsymbol{\alpha}^T T_t + \boldsymbol{\beta}^T \mathbf{X} + \gamma_{11} C_1 + \gamma_{12} C_2 + \gamma_{13} C_3 + \psi_1 r C_1), \\
\Pr(Y_2^{(r)} \mid \boldsymbol{\tau}, \mathbf{X}, R, \mathbf{C}, \mathbf{Z}) &= \text{expit}(\boldsymbol{\tau}^T \mathbf{Z}_t + \boldsymbol{\alpha}^T T_t + \boldsymbol{\beta}^T \mathbf{X} + \gamma_{21} C_1 + \gamma_{22} C_2 + \gamma_{23} C_3 + \\
&\qquad \psi_{21} r C_1 + \psi_{22} r C_2), \\
\Pr(Y_3^{(r)} \mid \boldsymbol{\tau}, \mathbf{X}, R, \mathbf{C}, \mathbf{Z}) &= \text{expit}(\boldsymbol{\tau}^T \mathbf{Z}_t + \boldsymbol{\alpha}^T T_t + \boldsymbol{\beta}^T \mathbf{X} + \gamma_{31} C_1 + \gamma_{32} C_2 + \gamma_{33} C_3 + \\
&\qquad \psi_{31} r C_1 + \psi_{32} r C_2 + \psi_{33} r C_3). \qquad (22)
\end{aligned}$$

Under model (22) and the model assumptions in Section 4, the conditional likelihood for the subset of patients with $R_i = 1$ and no missed visits, conditioning on $R_i = 1$ and $\mathbf{A}_i$, is the random effects logistic likelihood

$$\prod_{i=1 | R_i=1, O_{i1}=O_{i2}=O_{i3}=1}^{n} \int (\pi_{Y_{it}}(\boldsymbol{\tau}))^{Y_{it}} (1 - \pi_{Y_{it}}(\boldsymbol{\tau}))^{1-Y_{it}} f(\boldsymbol{\tau} \mid \Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau}, \qquad (23)$$

where

$$\pi_{Y_{it}(\boldsymbol{\tau})} = \text{expit}(\boldsymbol{\tau}^T \mathbf{Z}_{it} + \boldsymbol{\alpha}^T T_t + \boldsymbol{\beta}^T \mathbf{X}_i + \zeta_{t1} A_{i1} + \zeta_{t2} A_{i2} + \zeta_{t3} A_{i3}), \qquad (24)$$

and $\zeta_{11} = \psi_1 + \gamma_{11}$, $\zeta_{12} = \gamma_{12}$, $\zeta_{13} = \gamma_{13}$, $\zeta_{21} = \psi_{21} + \gamma_{21}$, $\zeta_{22} = \psi_{22} + \gamma_{22}$, $\zeta_{23} = \gamma_{23}$, $\zeta_{31} = \psi_{31} + \gamma_{31}$, $\zeta_{32} = \psi_{32} + \gamma_{32}$, $\zeta_{33} = \psi_{33} + \gamma_{33}$. Under the assumption (7), we have $\zeta_{12} = \zeta_{13} = \zeta_{21} = \zeta_{23} = \zeta_{31} = \zeta_{32} = 0$. Thus, we can test assumption (7) by testing $H_0 : \zeta_{12} = \zeta_{13} = \zeta_{21} = \zeta_{23} = \zeta_{31} = \zeta_{32} = 0$. We fit the random effects logistic likelihood (23) with $\pi_{Y_{it}(\boldsymbol{\tau})}$ given by (24) for the patients randomized to the treatment arm with no missed visits (there are 179 such patients) and found that the test of $H_0 : \zeta_{12} = \zeta_{13} = \zeta_{21} = \zeta_{23} = \zeta_{31} = \zeta_{32} = 0$ has a p-value of 0.95, thus providing no evidence against assumption (7). The above test has limitations. First, the test only addresses the validity of assumption (7) for the randomized to treatment potential outcomes and does not address the validity for the randomized to control potential outcomes. Second, the test has no power against certain alternatives, e.g., $\psi_{21} \neq 0, \gamma_{21} \neq 0$ but $\psi_{21} + \gamma_{21} = 0$. Third,

in the context of the depression study, the test has small power because A_1, A_2, A_3 are fairly highly correlated for the randomized to treatment group (correlations range from 0.51 to 0.79); the standard errors for $\zeta_{12}, \zeta_{13}, \zeta_{21}, \zeta_{23}, \zeta_{31}, \zeta_{32}$ range from 1.06 to 1.40. Development of better testing approaches is a valuable topic for future research.

9.2.5 Assessment of Drop-out Assumptions

We have assumed that the drop-out process is noninformative, meaning that drop-outs are independent of the outcomes conditional on the observables $(\mathbf{X}, \mathbf{Z}, R)$. The use of a random effects model enables a certain type of informative drop-out to be modeled through a shared random effects parameter, e.g., [15, 16]. The shared parameter model we consider assumes that drop-out and longitudinal outcomes are independent conditional on the observables $(\mathbf{X}, \mathbf{Z}, R)$ *and* the random effect $\boldsymbol{\tau}$. We model T_i (the last time point at which patient i was observed) by a continuation ratio logit model as in Ten Have et al. [16]:

$$\Pr(T_i = t \mid T_i > t-1, \tau_i, \mathbf{X}_i, R) = \text{expit}(\lambda_t \tau_i + \boldsymbol{\theta}^T \mathbf{T}_t + \boldsymbol{\upsilon}^T \mathbf{X}_i + \varpi R), \tag{25}$$

$t = 2, 3$. To fit the shared parameter model, we maximized the following function over the parameters $\sigma^*_\tau, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*, \boldsymbol{\upsilon}^*, \varpi^*$:

$$\prod_{i=1}^{n} \int \prod_{t=1}^{T_i} (\hat{\pi}_{Y_{it}}(\tau^*))^{Y_{it}} (1 - \hat{\pi}_{Y_{it}}(\tau^*))^{1-Y_{it}} f_{\tau^*}(\tau^* \mid \sigma^*_\tau) f_{T_i}(T_i; \tau^*, \mathbf{X}_i, R) d\tau^*,$$

where $\hat{\pi}_{Y_{it}}(\tau^*) = \text{expit}(\tau^* + \boldsymbol{\alpha}^{*T} \mathbf{T}_{it} + \boldsymbol{\beta}^{*T} \mathbf{X}_i + \gamma^*_t \hat{W}_{it} + \psi^*_t A_{it})$, $f(\tau^* \mid \sigma^*_\tau)$ is the density $N(0, \sigma_{\tau^*})$ and $f(T_i; \tau^*, \mathbf{X}_i, R)$ is the probability that subject i's last time point was T_i conditional on $\tau_i = \tau^*$ given by the continuation logit ratio model (25). The shared parameter model estimates of ψ_1, ψ_2 and ψ_3 are 1.08, 0.88 and 1.00 respectively, negligible changes from Table 3. The coefficients λ_2, λ_3 on τ in the continuation ratio logit model (25) for T_i are not significant (p-values of 0.07 and 0.25 respectively). The shared parameter model represents one type of informative dropout; there are other types of informative dropout, some of which cannot be tested based on the observed data [35]. Development of better methods for testing missing data assumptions and accommodating nonignorable missing data are valuable topics for future research.

9.2.6 Multidimensional Random Effects

We have assumed that the random effects vector $\boldsymbol{\tau}_i$ consists of just a random intercept τ_{0i}. The random effects can be made multidimensional to model variability in the pattern of

patients' outcome probabilities over time. To examine variability in the pattern of patients' outcome probabilities over time, we considered the random effect vector (τ_{0i}, τ_{1i}) where τ_{0i} is a random intercept, τ_{1i} is a random slope for time and $\mathbf{Z}_{it} = (1 \;\; t)$. The estimate variance of τ_{1i} is 0.07 with a standard error of 0.04. Thus, there is evidence of little variability in the slope of patients' outcome probabilities over time. We also considered random slopes for the covariates in $\mathbf{X}_i$, baseline Hamilton and suicide ideation score. For both covariates, the estimated variance of the random effect was less than 0.1 and not significantly different from zero.

10.0 Usefulness of Efficacy for Predicting the Effect of Future Treatment Programs

As noted in the introduction, the main goal of a clinical trial is to predict the comparative effect of future treatment programs. For example, a primary motivation for our analysis of the efficacy of the encouragement intervention for the depression study is to provide guidance for a cost-benefit analysis of implementing the encouragement intervention more widely. The model we study in this paper (6) provides an explanation for the results of the trial in terms of the difference between randomized groups stratified by compliance status [4]. Although the quantities in the model do not directly predict the effect of future treatment programs, these quantities can be important building blocks for making such predictions [48]. This section illustrates, in particular, how the efficacy of treatment received for compliers (i.e., the ITT effect for the strata of compliers) can be an important quantity for extrapolating from the results of the trial to predict the effect of future treatment programs. The results presented in this section are similar in spirit to those of Joffe and Brensinger [49], who provide an illustration of how structural mean model explanatory analyses of randomized trials can be used to predict the effect of future treatment programs.

Consider a situation in which a decision is being made as to whether to make the treatment available to a general population after the trial. Assume that the patients in the trial are representative of the general population. Let $Y_{it}^{*(1)}$ represent the potential outcome for patient i at time t if the treatment is made available to the general population after a trial took place that yielded the same results as the actual trial but did not involve patient i (i.e., in place of patient i, the trial involved a different patient with identical outcomes to patient i; the motivation for excluding patient i from the trial in these potential outcomes is to avoid carryover effects from

the trial). Correspondingly, let $Y_{it}^{*(0)}$ represent the potential outcome at time t for patient i if the treatment is not made available to the general population after a trial took place that yielded the same results as the actual trial but did not involve patient i. Let $A_{it}^{*(1)}$ and $A_{it}^{*(0)}$ represent the corresponding potential treatment receiveds if the treatment is made available/not made available to the general population. Consider the following assumptions:

(a) Similar to the trial, the treatment cannot be received if it is not made available, i.e., $A_{it}^{*(0)} = 0$ for all i.

(b) Other than potentially having a different effect on treatment received, assignment to the treatment/control arm is no different than having the treatment made available/not made available to the general population. Also treatment administration is the same in and out of the trial. Consequently,

$$\text{If } A_{it}^{*(r)} = A_{it}^{(r)}, \text{ then } Y_{it}^{*(r)} = Y_{it}^{(r)}. \tag{26}$$

Note that if (26) fails to hold for compliers in the trial, then the interpretation of the ITT effect for the strata of compliers in the trial as the efficacy of treatment received for compliers in the trial is questionable; see Section 10.1 below.

(c) An exclusion restriction for never takers outside the trial holds that is similar to the exclusion restriction for never takers in the trial (2):

$$\text{If } A_{it}^{*(1)} = 0, \text{ then } Y_{it}^{*(1)} = Y_{it}^{*(0)}. \tag{27}$$

We will now show that, under assumptions (a)-(c), and the assumptions in Section 4, the efficacy for compliers in the trial is a key quantity for extrapolating from the ITT effect in the trial to predict the effect of making the treatment available to the general population versus not making it available. The average ITT effect in the trial under the assumptions in Section 4 equals $P(A_{it}^{(1)} = 1)E[Y_{it}^{(1)} - Y_{it}^{(0)} \mid A_{it}^{(1)} = 1]$. Note that in this section we will focus on marginal average effects for simplicity of presentation but the same general principles apply to the odds ratio conditional on random effects that we have estimated in this paper. The average effect of making the treatment available to the general population versus not making it available is equal

to:

$$\begin{aligned} E[Y_{it}^{*(1)} - Y_{it}^{*(0)}] &= P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1] + \\ & \quad P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0] + \\ & \quad P(A_{it}^{*(1)} = 0, A_{it}^{(1)} = 1)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 0, A_{it}^{(1)} = 1] + \\ & \quad P(A_{it}^{*(1)} = 0, A_{it}^{(1)} = 0)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 0, A_{it}^{(1)} = 0]. \end{aligned}$$

Under the assumptions (a)-(c) above, we have

$$\begin{aligned} E[Y_{it}^{*(1)} - Y_{it}^{*(0)}] &= P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1] + \\ & \quad P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0] \\ &= \text{Average ITT effect in trial} - \\ & \quad [P(A_{it}^{(1)} = 1) - P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1)]E[Y_{it}^{(1)} - Y_{it}^{(0)} \mid A_{it}^{(1)} = 1] + \\ & \quad P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1) \times \\ & \quad \{E[Y_{it}^{(1)} - Y_{it}^{(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 1] - E[Y_{it}^{(1)} - Y_{it}^{(0)} \mid A_{it}^{(1)} = 1]\} + \\ & \quad P(A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0)E[Y_{it}^{*(1)} - Y_{it}^{*(0)} \mid A_{it}^{*(1)} = 1, A_{it}^{(1)} = 0]. \end{aligned} \tag{28}$$

From (28), the difference between 1) the average causal effect of the treatment program of making the treatment available to the general population versus not making it available and 2) the average ITT effect in the trial, depends on the following:

(i) the efficacy for compliers in the trial, $E[Y_{it}^{(1)} - Y_{it}^{(0)} \mid A_{it}^{(1)} = 1]$;

(ii) the proportion of compliers in the trial who would not take the treatment if offered it outside the trial;

(iii) the difference between the efficacy for compliers in the trial who would take the treatment if offered it outside the trial and the efficacy for compliers in the trial who would not take the treatment if offered it outside the trial;

(iv) the proportion of never takers in the trial who would take the treatment if offered it outside the trial;

(v) the average causal effect of taking the treatment outside the trial for never takers in the trial who would take the treatment if offered it outside the trial.

The efficacy for compliers, (i) in the list above, is thus an important quantity for extrapolating from the results of the trial to predict the effect of making the treatment available to the general population versus not making it available. Also, note that for a treatment whose efficacy has small variation across the population, we would expect (v) to be close to the efficacy for compliers and (iii) to have small magnitude. Although the setting considered in this section is simple, the principle it illustates of how efficacy can be a useful quantity for predicting the effects of future treatment programs carries over to many more complicated settings; see [48] for further discussion.

10.1 Interpretation of ψ_t as efficacy

Here we comment further on the interpretation of ψ_t as the efficacy of treatment received for compliers. The parameter ψ_t measures the effect of assignment to the treatment versus assignment to the control on the outcome for compliers at time t, see (9). For $t = 1$, it is reasonable to interpret ψ_1 as the efficacy of treatment received at time 1 on the outcome at time for compliers at time 1 when the stability assumption (26) holds between the trial and future treatment program potential outcomes. Under (26), assignment to the treatment either has no direct effect for the compliers beyond its indirect effect on treatment received or exactly the same direct effect in and out of the trial. In the former case, we can view the effect of assignment to treatment versus control for the compliers as the pure effect of treatment received.

For $t > 1$, it may be misleading to think of ψ_t as the efficacy of treatment received at time t if 1) there is time varying compliance so that some compliers at time t are never takers at time $t-1$ and 2) outcomes at time t are affected by the whole sequence of treatment receiveds up to time t. Under these circumstances, ψ_t is affected by the compliance behavior at time periods before t of compliers at time t and cannot be clearly interpreted. A condition under which it remains reasonable to think of ψ_t as the efficacy of treatment received at time t for $t > 1$ is when the treatment only has a "transient" effect. A formal transience assumption is the following. Let $Y_{it}^{*(a_1,\ldots,a_t)}$ denote the potential outcome for patient i at time t if the treatment is made available ($a_{t'} = 1$) or is not made available ($a_{t'} = 0$) at times $t' = 1, \ldots, t$. Then a formal

transience assumption is

$$Y_{it}^{*(a_1,\ldots,a_t)} = Y_{it}^{*(a_1^{'},\ldots,a_{t-1}^{'},a_t)} \tag{29}$$

Such a transience assumption is plausible for the depression study because of clinical researchers' expectation that the effect of treatment (contact with the depression specialist) does not extend beyond the next visit four months later. The assumption (7) is stronger in some sense than the transience assumption (29) because it can be violated not only if the treatment has cumulative effects but also because the strata of patients with compliance status vector $(C_1,\ldots,C_T)$ might not be comparable to a different strata of patients with compliance status vector $(C_1^{'},\ldots,C_{t-1}^{'},C_t,C_{t+1}^{'},\ldots,C_T^{'})$ in the sense that $E(Y_t^{(0)} \mid C_1,\ldots,C_T) \neq E(Y_t^{(0)} \mid C_1^{'},\ldots,C_{t-1}^{'},C_t,C_{t+1}^{'},\ldots,C_T^{'})$ (analogously $E(Y_t^{*(0,\ldots,0)} \mid C_1,\ldots,C_T) \neq E(Y_t^{*(0,\ldots,0)} \mid C_1^{'},\ldots,C_{t-1}^{'},C_t,C_{t+1}^{'},\ldots,C_T^{'})$).

11.0 Discussion

We have presented a random effects logistic regression approach for estimating the efficacy of treatment for compliers in a randomized study with longitudinal binary outcomes and treatment non-adherence. Our simulation results suggest that while an approximation, our approximate IV approach performs sufficiently well to provide reliable inferences for the setting of the depression study. Our approach is easily implementable using standard software such as SAS with macros available from the authors.

For the depression study considered, our efficacy estimates differ considerably from the as-treated estimates and are more reasonable in their relation to the ITT estimates than the AT estimates under the assumed exclusion restriction. Our efficacy estimates from the IV analysis paint a somewhat different picture of how the efficacy varies over time than the ITT and AT estimates – the IV analysis suggests that there is not much variation over time whereas the ITT and AT estimates suggests some variation over time. This pattern in the IV, ITT and AT estimates would be expected if there is a stable causal mechanism for the effect of treatment received on outcome but, as in our study, the amount of adherence changes over time.

We have formulated our model as a model for the effect of treatment received for the partially unobserved class of patients who would comply with an assignment to treatment. Our formulation is an example of the principal stratification approach to causal inference of Fran-

gakis and Rubin [24] in which causal inferences are made for groups (principal strata) whose membership is not affected by the randomization assignment. Here, the vector of compliance statuses is not affected by the randomization assignment. We could also have formulated our model as a model for the effect of treatment received for the observed class of patients who actually receive treatment. The latter approach to formulating models has been taken by Robins and coworkers in many contributions to causal inference, e.g., [8]. In settings such as ours in which subjects randomized to the control group cannot receive the active treatment, there is a formal equivalence between the estimands generated from conditioning on compliance status and those generated from conditioning on observed treatment received [50].

A principal benefit of our use of a random effects model for the depression study is that it provided an analysis of efficacy that is comparable to the ITT analysis that was done using random effects logistic models. Another valuable feature of the random effects model is that it enabled a certain type of informative drop-out to be accommodated (Section 9.2.5). A useful feature of random effects models that we did not discuss is that they enable information to be borrowed from other subjects for making more accurate treatment decisions for a given subject based on limited longitudinal data [17, 18].

Because the study design considered here involved only baseline randomization and our IV approach requires variability of treatment assignment to estimate causal effects, we cannot estimate the variability of treatment efficacy among patients without strong parametric assumptions. However, for study designs with sequential randomization (discussed by [51]), a random effects model that allows for variability in treatment efficacy can be formulated and such a model can be estimated by methods similar to this paper's.

Several issues concerning random effects models for longitudinal binary outcomes merit further research attention: 1) It would be desirable to have a more accurate estimation method than our approximate IV approach especially if there is strong confounding due to treatment nonadherence. Such an approach could be based on maximizing (11); 2) The study we considered was a clustered encouragement design but the cluster effects were small. Further study and development of appropriate methods for clustered encouragement designs with large cluster effects would be valuable; 3) It would be useful to develop methods to incorporate missing

data assumptions that violate (3) and to develop methods of sensitivity analysis to missing data assumptions; and 4) Methods for accommodating departures from the exclusion restriction would be valuable. Such departures have been accommodated in cross-sectional contexts, e.g., [27, 28, 50].

Acknowledgements

The authors would like to thank two reviewers for very helpful comments and suggestions that greatly improved the paper. The authors are grateful to Michael Elliott for insightful discussion. Funding for this work was provided by NIMH grants R01-MH61892 and P30-MH2129 and a NHLBI grant R29 HL 59184.

References

1. Pocock, S. *Clinical Trials: A Practical Approach*, Wiley, 1983.

2. Robins, J. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, Sechrest, L. (ed). NCHSR, U.S. Public Health Service, 1989: 113-159.

3. Sheiner, L. and Rubin, D. Intention-to-treat analysis and the goal of clinical trials. *Clinical Pharmacology & Therapeutics* 1995; **56**, 6–10.

4. White, I. and Goetghebeur, E. Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference. *Statistics in Medicine* 1998: **17**, 319–339.

5. May, G., Chir, B., Demets, D., Friedman, L., Furberg, C. and Passamani, E. The randomized clinical trial: Bias in analysis. *Circulation* 1981: **64**, 669–673.

6. Tarwotjo, I., Sommer, A., West, K., Djunaedi, E., Loedin, A., Mele, L. and Hawkins, B. Influence of participation on mortality in a randomized trial of vitamin A prophylaxis. *American Journal of Clinical Nutrition* 1987: **45**, 1466–1471.

7. Angrist, J., Imbens, G. and Rubin, D. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996: **91**, 444–455.

8. Robins, J. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics, Theory and Methods* 1994: **23**, 2379–2412.

9. Stock, J. Instrumental variables in statistics and economics. In *International Encyclopedia of the Social & Behavioral Sciences*, Baltes P and Smelser N (eds). Elsevier Science, 2001: 7577–7582.

10. Bruce, M., Ten Have, T., Reynolds, C., Katz, I., Schulberg, H., Mulsant, B., Brown, G., McAvay, G., Pearson, J., and Alexopoulos, G. Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *Journal of the American Medical Association* 2004: **291**, 1081–1091.

11. Depression Guideline Panel. *Depression in Primary Care, 2: Treatment of Major Depression*. U.S. Department of Human Services, Public Health Service, Agency for Health Care Policy and Research, 1993.

12. Unutzer, J., Katon, W., C.M., Callahan, C., Harpole, L., Hunkeler, E., Hoffing, M., Arean, P., Hegel, M., Schoenbaum, M., Oishi, S., and Langston, C. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *Journal of the American Medical Association* 2002: **288**, 2836–2845.

13. Zeger, S., Liang, K-Y., and Albert, P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988: **44**, 1049–1060.

14. Zeger, S., Liang, K-Y., and Albert, P. Response to letter commenting on: "Models for longitudinal data: a generalized estimating equation approach." *Biometrics* 1991: **47**, 1593–1596.

15. Wu, M. and Carroll, R. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988: **45**, 175–188.

16. Ten Have, T., Kunselman, A., Pulkstenis, E. and Landis, J. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* 1998: **54**, 367–383.

17. Sheiner, L., Rosenberg, B. and Melmon, K. Modeling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research* 1972: **6**, 441–459.

18. Berzuini, C. Medical monitoring. In *Markov Chain Monte Carlo in Practice*, Gilks, W., Richardson, S., and Spiegelhalter, D. (eds). London: Chapman-Hall, 1996: 321–337.

19. Nagelkerke, N., Fidler, V., Bernsen, R. and Borgdorff, M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000: **19**, 1849–1864.

20. Ten Have, T., Joffe, M. and Cary, M. Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine* 2003: **22**, 1255–1284.

21. Frangakis, C., Rubin, D. and Zhou, X.-H. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* 2002: **3**, 147–164.

22. Sato, T. A method for the analysis of repeated binary outcomes in randomized clinical trials with non-compliance. *Statistics in Medicine* 2001: **20**, 2761–2774.

23. Frangakis, C., Brookmeyer, R., Varadhan, R., Safaeian, M., Vlahov, D. and Strathdee, S. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association* 2004: **97**, 284–292.

24. Frangakis, C. and Rubin, D. Principal stratification in causal inference. *Biometrics* 2002: **58**, 21–29.

25. Yau, L. and Little, R. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 2001: **96**, 1232–1244.

26. Goldman, D., Bhattacharya, J., McCaffrey, D., Duan, N., Leibowitz, A., Joyce, G., and

Morton, S. Effect of insurance on mortality in an HIV-positive population in care. *Journal of the American Statistical Association* 2001: **96**, 883–894.

27. Hirano, K., Imbens, G., Rubin, D. and Zhou, X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000: **1**, 69–88.

28. Jo, B. and Muthen, B. Longitudinal studies with intervention and noncompliance: estimation of causal effects in growth curve mixture modeling. In *Multilevel Modeling*, Duan N. and Reise S.(eds). New York, Lawrence Erlbaum Associates, 2001: 51–52.

29. Vansteelandt, S. and Goetghebeur, E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B* 2003: **65**, 817–835.

30. Neyman, J. On the application of probability theory to agricultural experiments: Essay on principles. 1923. Translated by D.M. Dabrowska and edited by T.P. Speed. *Statistical Science* 1990: **5**, 465–472.

31. Rubin, D. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974: **66**, 688–701.

32. Zelen, M. A new design for randomized clinical trials. *New England Journal of Medicine* 1979: **300**, 1242–1245.

33. Zelen, M. Randomized consent designs for clinical trials: an update. *Statistics in Medicine* 1990: **9**, 645–656.

34. Rubin, D. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association* 1986: **81**, 961–962.

35. Little, R. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995: **90**, 1112–1121.

36. Frangakis, C. and Rubin, D. Addressing complications of intent-to-treat analysis in the combined presence of all-or-none treatment non-compliance and subsequent missing outcomes. *Biometrika* 1999: **86**, 365–379.

37. Mealli, F., Imbens, G., Ferro, S. and Biggeri, A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* 2004: **5**, 207–222.

38. Imbens, G. and Angrist, J. Identification and estimation of local average treatment effects. *Econometrica* 1994: **62**, 467–476.

39. Sommer, A. and Zeger, S. On estimating efficacy from clinical trials. *Statistics in Medicine* 1991: **10**, 45–52.

40. Robins, J. and Greenland, S. Comment on "Identification of causal effects using instrumental variables" by J.D. Angrist, G.W. Imbens, and D.B. Rubin. *Journal of the American Statistical Association* 1996: **91**, 456–458.

41. Gail, M., Wieand, S. and Piantados, S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984: **71**, 431–444.

42. Guo, J. and Geng, Z. Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society, Series B* 1995: **57**, 263–267.

43. Lai, T. and Shih, M.-C. A hybrid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* 2003: **90**, 859–879.

44. Breslow, N. and Clayton, D. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993: **88**, 9–25.

45. Branson, M. and Whitehead, J. A score test for binary data with patient non-compliance. *Statistics in Medicine* 2003: **22**, 3115–3132.

46. Copas, J. and Li, H. Inference for non-random samples. *Journal of the Royal Statistical Society, Series B* 1997: **59**, 55–95.

47. Babiker, A. and Cuzick, J. A simple frailty model for family studies with covariates. *Statistics in Medicine* 1994: **13**, 1679–1692.

48. Sheiner, L. Is intent-to-treat analysis always (ever) enough? *British Journal of Clinical Pharmacology* 2002: **54**, 203–211.

49. Joffe, M. and Brensinger, C. Weighting in instrumental variables and g-estimation. *Statistics in Medicine* 2003: **22**, 1285–1303.

50. Ten Have, T., Elliott, M., M., J., Zanutto, E. and Datto, C. Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association* 2004: **99**, 8–16.

51. Murphy, S. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B* 2003: **65**, 331–355.