The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 22

Causal Inference from Longitudinal Studies with Baseline Randomization

Sengwee Toh, *Department of Epidemiology, Harvard School of Public Health*
Miguel A. Hernán, *Department of Epidemiology, Harvard School of Public Health, and Harvard-MIT Division of Health Sciences and Technology*

Recommended Citation:
Toh, Sengwee and Hernán, Miguel A. (2008) "Causal Inference from Longitudinal Studies with Baseline Randomization," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 22.
DOI: 10.2202/1557-4679.1117

Causal Inference from Longitudinal Studies with Baseline Randomization

Sengwee Toh and Miguel A. Hernán

Abstract

We describe analytic approaches for study designs that, like large simple trials, can be better characterized as longitudinal studies with baseline randomization than as either a pure randomized experiment or a purely observational study. We (i) discuss the intention-to-treat effect as an effect measure for randomized studies, (ii) provide a formal definition of causal effect for longitudinal studies, (iii) describe several methods -- based on inverse probability weighting and g-estimation -- to estimate such effect, (iv) present an application of these methods to a naturalistic trial of antipsychotics on symptom severity of schizophrenia, and (v) discuss the relative advantages and disadvantages of each method.

KEYWORDS: causal inference, inverse probability weighting, marginal structural model, g-estimation, large simple trial

Author Notes: This research was funded by NIH grant R01 HL080644-01. The authors thank Douglas Faries, PhD, and Sonia Hernández-Díaz, M.D., Dr.P.H., for her comments to an earlier version of this manuscript.

1 Longitudinal studies with baseline randomization

A study is said to be a longitudinal, or a follow-up, study when subjects are followed from study entry until the determination of certain outcome of interest, loss to follow-up, or the administrative end of follow-up, whichever comes first. Longitudinal studies are often referred to as cohort studies by epidemiologists and as panel studies by social scientists. When the goal is estimating the causal effect of certain treatment on the outcome, longitudinal studies are preferred over non longitudinal (i.e., cross-sectional) ones in which the temporal order of treatment and outcome may be unclear. Longitudinal studies are usually classified as either experiments (the treatment is assigned by the investigators) or observational studies (the investigators play no role in treatment assignment). Experiments are said to be randomized when the investigators assign the treatment at random. Randomized experiments are considered the mainstay design for causal inference. Data from randomized experiments are usually analyzed in a very straightforward manner: the distribution of the outcome is compared between those assigned to each treatment group. If a difference is found, then treatment is declared to have a causal effect on the outcome. Below we discuss some advantages and disadvantages of this "intention-to-treat" analysis.

Despite the apparently clear distinction between randomized experiments and observational studies, in practice it is common to find longitudinal studies that combine characteristics from both designs. For example, consider a conventional two-arm randomized clinical trial in which the investigators select a group of subjects based on stringent eligibility criteria, randomly assign them to one of two treatments (or placebo) at baseline, and monitor them closely until the end of follow-up. Some of the subjects participating in this study may, at any time, deviate from the trial's protocol by switching to a treatment other than that assigned to them at baseline or by dropping out of the study completely. In the presence of these deviations from protocol, which are not randomly assigned by the investigators but rather the result of subjects and treating physicians' decisions, the investigators can only record data as if they were conducting an observational study. The greater the proportion of subjects who deviate from the trial's protocol, the closer the resemblance between the randomized clinical trial and an observational study, and the more questionable the intention-to-treat analysis of the trial becomes.

Hence one can think of a continuum from an ideal randomized experiment in which subjects (perhaps laboratory rats) are fully compliant with the assigned treatment and never lost to follow-up to a purely observational study in which subjects' information is prospectively collected whenever it becomes available. The terms "large simple trial" or naturalistic trial have been coined to refer to a type of longitudinal study that shares, by design, characteristics of both randomized trials and observational studies. In a large simple trial, like in a conventional randomized clinical trial, the treatment is randomly assigned at baseline. However, large simple trials differ from conventional randomized clinical trials in their relative lack of restrictions on subject eligibility and their simplified data collection (Lesko and Mitchell, 2005). The idea is increasing the generalizability and the clinical relevance of the results by including subjects that represent the diversity existing in the actual patient population, and by explicitly allowing the treating physicians to modify the treatment regime depending on the subjects' response to the assigned treatment and their changing prognosis over the (often long) duration of the study.

This article is concerned with study designs that, like large simple trials, can be better characterized as longitudinal studies with baseline randomization than as either a pure randomized experiment or a purely observational study. We (i) discuss the intention-to-treat effect as an effect measure for randomized studies, (ii) provide a formal definition of causal effect for longitudinal studies, (iii) describe several methods — based on inverse probability weighting and g-estimation — to estimate such effect, (iv) present an application of these methods to a randomized study of antipsychotic therapy, and (v) discuss the relative advantages and disadvantages of each method. We start by describing the longitudinal study with baseline randomization that will be used as an example throughout the article.

2 Example: Antipsychotic medications and severity of schizophrenia symptoms

We analyzed a randomized, open-label, multi-center trial to compare the effect of antipsychotic medications on the symptom severity of schizophrenia. Details of the trial have been described elsewhere (Tunis *et al.*, 2006). Briefly, subjects were recruited within both academic and community treatment settings (primarily in mental health outpatient clinics) between May 1998 and September 2001, and were randomly assigned to one of three first-line treatments: olanzapine (N=229), risperidone (N=221), or conventional antipsychotics (N=214). Both olanzapine and risperidone are commonly known as atypical antipsychotics. Within the conventional group, the choice of antipsychotics (e.g., perphenazine, haloperidol) was at the discretion of the treating physicians. For simplicity, our analysis combined the olanzapine and risperidone arms to form an atypical antipsychotic arm (randomization arm R=1, N=450) and compared it with the conventional antipsychotic arm (R=0, N=214). Thus, for the purposes of this paper, we effectively assume that all types of atypical antipsychotic regimes are equivalent.

To be eligible for the study, subjects had to be at least 18 years old, meet the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for schizophrenia, schizoaffective disorder, or schizophreniform disorder, have no serious medical conditions or history of contraindication of the study medications, and have a psychotic symptom threshold of $\geq$18 on the Brief Psychiatric Rating Scale (BPRS). The BPRS score is commonly used to measure the symptom severity of schizophrenia (Overall and Gorham, 1962) and was rated by the clinicians in the current study. Each symptom in the scale ranges from 0 (not present) to 6 (extremely severe).

There was a randomization visit and five post-randomization visits at 2 weeks, and 2, 5, 8, and 12 months. Data on medication use, clinical symptoms, BPRS score, quality of life, and resource use were collected at the time of randomization and at each post-baseline visit. Subjects in the two arms had similar baseline characteristics (Table 1). The outcome of interest, Y, was the change in BPRS score between baseline and end of the study (i.e., 12 months post-baseline), with a negative value indicating a clinical improvement. For each subject, the treatment assigned at randomization could be changed (e.g., from conventional to atypical antipsychotics) during the study period based on the subject's response or other reasons. Only 7% of the person-visits reported no use of any antipsychotic therapy. For simplicity, our analyses do not differentiate between use of conventional antipsychotics and no use of any antipsychotics. We now describe the intention-to-treat approach and its application to this study.

Table 1. *Baseline characteristics by treatment arm **

Characteristics	*Atypical antipsychotic arm (N=450)*	*Conventional antipsychotic arm (N=214)*	*p-value †*
Age (in years): mean (SD)	42.4 (12.0)	43.6 (12.1)	0.24
Male: number (%)	277 (61.6)	143 (66.8)	0.19
Race: number (%)			0.79
White	241 (53.6)	108 (50.5)	
Black	138 (30.7)	74 (34.6)	
Other	54 (12.0)	24 (11.2)	
Missing	17 (3.8)	8 (3.7)	
Baseline BPRS: mean (SD)	32.1 (11.7)	31.2 (11.1)	0.37
Baseline GAF: mean (SD)	46.0 (12.9)	46.3 (12.8)	0.84

* SD: standard deviation; BPRS: Brief Psychiatric Rating Scale score; GAF: global assessment of functioning score

† Based on t-test for continuous variables, and χ^2 test for categorical variables.

3 The intention-to-treat effect

Randomized experiments, when analyzed using the intention-to-treat (ITT) principle, do not require any assumptions to consistently estimate causal effects. To briefly describe the ITT principle, suppose you want to estimate the causal effect of a dichotomous treatment A on a continuous outcome Y in certain population. You conduct a randomized experiment by randomly splitting the population into two groups, assigning one group (arm R=1) but not the other (arm R=0) to be treated, following all subjects for some fixed period (say, one year), and measuring every subject's outcome at the end of that period. Under the ITT principle, you compare the mean outcome between the group that you intended to treat (R=1) and the group that you intended to keep untreated (R=0), regardless of the treatment that each subject actually received. If the mean outcome differs, you can conclude that treatment A has an effect on the mean of Y because these two groups are expected to be exchangeable with respect to all measured and unmeasured characteristics at baseline. In contrast, causal inferences from observational studies are risky precisely because this exchangeability cannot be guaranteed.

When all subjects comply with their assigned treatment and there is no loss to follow-up, the mean outcome in group R=1, i.e., $E[Y|R=1]$, is also the mean outcome among the treated, i.e., $E[Y|A=1]$, which consistently estimates the mean outcome that would have been observed if all subjects in the population had been treated, i.e., $E[Y^{a=1}]$, where a subject's $Y^{a=1}$ is the counterfactual (or potential) outcome that would have been observed if the subject had, possibly contrary to the fact, received treatment (a=1). Similarly, the mean outcome among the untreated, i.e., $E[Y|R=0]= E[Y|A=0]$, is a consistent estimator of the mean outcome that would have been observed if no subject in the population had been treated, i.e., $E[Y^{a=0}]$. Thus, in the absence of noncompliance and loss to follow-up, the ITT difference of observed means

$$E[Y \mid R=1] - E[Y \mid R=0] = E[Y \mid A=1] - E[Y \mid A=0]$$

consistently estimates the difference of counterfactual (or potential) means

$$E[Y^{a=1}] - E[Y^{a=0}],$$

which is the effect of treatment on the outcome in the population on the scale of difference of means. For an introduction to counterfactual-based causal inference see, for example, Hernán (2004).

But things are rarely that easy in longitudinal studies with randomization at baseline. A key limitation of many randomized experiments, like our study of antipsychotics, is lack of compliance with the treatment assigned at baseline. That is, subjects may not adhere to their assigned treatment for the entire follow-up from randomization to the measurement of the outcome. Another key limitation of many randomized experiments is that some subjects do not complete the follow-up. In our study, only 46% (205/450) of the subjects assigned to atypical antipsychotics (R=1), and 26% (56/214) of the subjects assigned to conventional ones (R=0), stayed on their assigned treatment and completed the follow-up.

The problem of noncompliance highlights the crucial differences between assigned treatment and received treatment: the randomized assignment R is a baseline variable over which the investigators have direct control, but the treatment A_k received at visit k is actually a time-varying variable whose value at any given time is beyond the investigators' control. As a consequence, the ITT contrast $E[Y|R=1] - E[Y|R=0]$ does not estimate the effect of receiving the treatment but the effect of being assigned to the treatment. When, as in our study, all subjects initiate the treatment they were assigned to, regardless of whether they later continue taking it, we say the ITT effect is the effect of treatment initiation.

Because the ITT effect depends on the degree of noncompliance, it may be close to null in placebo-controlled experiments even if the treatment does actually have an effect, or it may be non null when comparing two active treatments even if the two treatments are equally effective. Despite this limitation, the ITT effect is often the only effect estimated in placebo-controlled randomized experiments because (i) it provides a valid test of the null hypothesis, and (ii) it is usually a conservative estimate (i.e., biased towards the null) of the effect of actually receiving the treatment. The conservativeness of the ITT effect in placebo-controlled experiments, however, makes it a risky effect measure when the goal is evaluating a treatment's safety: one could naïvely conclude that a treatment is safe because the ITT effect is null, even if treatment causes serious adverse effects. The explanation may be that many subjects stopped taking the treatment before developing the adverse effects.

In our study, we calculated the mean change in schizophrenia symptom severity, as measured by BPRS score, from baseline to the end of follow-up in each arm: $E[Y|R=1] = -10.55$ and $E[Y|R=0] = -9.50$. We then estimated the ITT effect as the difference $E[Y|R=1] - E[Y|R=0] = -1.05$ (95% confidence interval [CI]: –3.26, 1.16), i.e., the group assigned to atypical antipsychotics (R=1) had a slightly greater improvement in symptom severity at the end of follow-up than the group assigned to conventional antipsychotics (R=0). This estimate depends on the degree of noncompliance in our study. We defer to the next section a discussion about noncompliance and focus now on loss to follow-up.

Our ITT analysis seems straightforward. However, this is only because we cheated by ignoring that some subjects were lost to follow-up and thus did not have a measurement of BPRS score at one year from baseline. To obtain the estimate of –1.05, we restricted the analysis to subjects with at least one post-randomization visit during which their BPRS score was recorded (430 in the R=1 arm, 204 in the R=0 arm). We then computed, for each subject, the difference between her BPRS score at baseline and at her last available visit, which could have taken place at 2 weeks; or 2, 5, 8, or 12 months post-baseline. Thus, our –1.05 estimate is based on differences in symptoms severity measured at different times for each subject, depending on how long she was under follow-up. This "last available observation carried forward" approach, although commonly used in practice, is not a true ITT analysis. For example, imagine that treatment worsens some subjects' symptoms so much that they do not return after baseline (or that they return only at 2 weeks when the harmful effect of treatment is not apparent yet). Then this pseudo-ITT analysis would make treatment look better than it really is.

A naïve alternative to the pseudo-ITT analysis above is the so-called "complete-case" ITT analysis. That is, an ITT analysis restricted to subjects who completed the follow-up. In our study, the estimate for the "complete-case" ITT analysis was 0.42 (95% CI: –2.36, 3.19). Again, this estimate may be biased because those who completed the follow-up in arms R=1 and R=0 may not be exchangeable. Rather than a simple complete-case analysis, we need an analysis that still compares the mean outcomes between arms R=1 and R=0 but that adjusts for the potential bias induced by loss to follow-up. However, adjustment for loss to follow-up requires making uncheckable assumptions about the comparability between the subjects that were and were not lost to follow-up. This is too bad as the beauty of a pure ITT analysis is, precisely, that it does not require any assumptions about exchangeability.

In our study, we adjusted for censoring due to incomplete follow-up by using inverse probability weighting (IPW). IPW requires the untestable assumption that subjects with complete and incomplete follow-up are exchangeable, conditional on the measured variables. Because IPW is more formally described in Section 5, we only briefly outline the procedure here. First, for each subject, we estimated her probability of providing complete data at all visits from baseline until the first visit she missed, the first visit with incomplete information, or the end of follow-up at one year, whichever happened first. Second, we restricted the ITT analysis to those subjects who completed all visits (235 in R=1 and 130 in R=0), and assigned to each of them a weight proportional to the inverse of their estimated probability of complete follow-up. The inverse probability weighted ITT estimate was –0.86 (95% CI: –3.88, 2.15).

4 The effect of continuous treatment

As reviewed above, the ITT effect of baseline assignment R is often not satisfactory, even if no subject had been lost to follow-up, because of the presence of noncompliance. In those cases, we would rather estimate the effect of the time-varying treatment A_k. But what do we mean by the effect of a time-varying treatment A_k? The response to this question is not unique. One possibility is comparing the mean outcome if all subjects had been continuously treated during the entire follow-up with the mean outcome if no subject had been ever treated during the follow-up. We now provide a formal definition of this effect, which we will refer to as the “effect of continuous treatment”. For pedagogic purposes, we ignore the presence of incomplete follow-up until the end of this section.

We first need to introduce some notation. Let A_k be 1 if the subject is on atypical antipsychotics at visit k, and 0 otherwise. The baseline value A_0 equals the value of the randomized assignment R. For simplicity, let us assume that treatment status can only change at the time of a visit. We use the overbar notation $\bar{A}_k$ to denote a subject’s treatment history from baseline until visit k, i.e., $\bar{A}_k$={A_0, A_1,… A_k}. In our study k takes value 0 for the baseline (randomization) visit, and values 1 to 4 for the post-randomization visits. We use $\bar{A}$ to denote the subject’s treatment history over the entire follow-up. For example, a subject who was assigned to atypical antipsychotics (R=1) and initially complied with her assignment but stopped taking treatment at visit k=1 would have a treatment history $\bar{A}$={1, 1, 0, 0, 0}. We define a pre-specified (or static) treatment regime as $\bar{a}$={a_0, a_1, a_2, a_3, a_4}. For example, the treatment regime “always treated with atypical antipsychotics” can be represented as $\bar{1}$={1,1,1,1,1}, and the treatment “never treated with atypical antipsychotics” as $\bar{0}$={0,0,0,0,0}. A subject’s $Y^{\bar{a}}$ is the (possibly counterfactual) outcome that would have been observed at visit k=5 (i.e., one year post-baseline) if the subject had followed regime $\bar{a}$. Thus the effect of continuous treatment in the population can be expressed as

$$E[Y^{\bar{a}=\bar{1}}]-E[Y^{\bar{a}=\bar{0}}].$$

In our study, this is the mean counterfactual outcome that would have been observed if all subjects had been continuously treated with atypical antipsychotics minus the mean counterfactual outcome that would have been observed if all subjects had been never treated with atypical antipsychotics. If all subjects had been always on either atypical or conventional antipsychotic therapy, the effect of continuous treatment could also be conceptualized as “the effect had everybody stayed on their assigned treatment regime”, i.e.,

$$E[Y^{\bar{a}=\bar{1}} \mid R=1] - E[Y^{\bar{a}=\bar{0}} \mid R=0],$$

because in randomized experiments groups R=1 and R=0 are expected to be exchangeable.

Unlike the ITT effect, the effect of continuous treatment does not depend on the degree of noncompliance with the assigned treatment during the follow-up period. But, unlike the ITT effect in the absence of loss to follow-up, the effect of continuous treatment cannot be consistently estimated by a simple comparison. Specifically, we cannot simply compare the mean outcome between those subjects who happened to follow the regimes $\bar{1}$ "always treated with atypical antipsychotics" and $\bar{0}$ "never treated with atypical antipsychotics", because subjects who followed those regimes did so for some particular reasons (e.g., their response to treatment or the severity of their condition) and are generally not exchangeable. For example, in our study subjects randomized to atypical antipsychotics R=1 who did not require hospitalization during the follow-up were more likely to stay on their assigned treatment. As a result, the group of subjects who stayed on atypical antipsychotics $\bar{A} = \bar{1}$ is the selected sample of subjects that either had low severity to start with or that responded well to atypical antipsychotics. That is, the contrast

$$E[Y \mid R=1, \bar{A} = \bar{1}] - E[Y \mid R=0, \bar{A} = \bar{0}],$$

which equals the contrast

$$E[Y \mid \bar{A} = \bar{1}] - E[Y \mid \bar{A} = \bar{0}],$$

will generally result in a biased estimate of the effect of continuous treatment, even if no subjects had been lost to follow-up. This contrast is usually known as the "per protocol" analysis. (Note that the sample estimates of the differences $E[Y \mid R=1, \bar{A} = \bar{1}] - E[Y \mid R=0, \bar{A} = \bar{0}]$ and $E[Y \mid \bar{A} = \bar{1}] - E[Y \mid \bar{A} = \bar{0}]$ will generally differ.)

Estimating the effect of continuous treatment may require data on the time-varying treatment and joint predictors of compliance and the outcome, and some sort of adjustment for such predictors. In an attempt to identify the main predictors of noncompliance, Table 2 shows the association between several factors and compliance with the assigned treatment by randomized arm. To estimate the odds ratios in the table, we fit a pooled logistic model for the probability of staying on the assigned treatment

$$\Pr[A_k = r \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1} = \overline{r}, \overline{L}_k = \overline{l}_k, R = r]$$

in each arm R=r, where L_k is the vector of time-varying covariates measured at visit k, $\overline{L}_k$ ={L_0, L_1,… L_k } is the history of measured covariates measured by visit k, C_k is a time-varying censoring indicator that takes value 0 for subjects uncensored (i.e., with complete follow-up data) by time k, and value 1 otherwise, and $\overline{C}_{\mathrm{k}}$ ={C_0, C_1, …C_k} is a subject's censoring history through k. $\overline{C}_k = \overline{0}$ denotes complete follow-up through visit k. We assumed that L_k preceded A_k (our results were not sensitive to this assumption) and that $\overline{L}_k$ could be appropriately summarized by the following variables: age at baseline, sex, ethnicity, time since randomization, and the following time-varying covariates: baseline and most recent BPRS and global assessment of functioning (GAF, smaller score indicates more severe dysfunction) scores, and a moderate or severe adverse event (1: yes; 0: no) and hospitalization (1: yes; 0: no) since last visit. These time-varying covariates span the domains of symptom severity, functioning, tolerability, and resource utilization. Each person contributed as many observations to the logistic model as visits she was under complete follow-up until abandoning her assigned treatment. The model did not include covariates to summarize prior treatment history $\overline{A}_{k-1}$ because all subjects had the same treatment history while on their assigned treatment.

The analysis shown in Table 2 identifies some important predictors of noncompliance, such as recent symptom severity or hospitalization. Because these factors are also predictors of the outcome (which is, in fact, symptom severity at a later time), some sort of adjustment for these factors may be required. The need for an adjustment raises two issues.

First, adjustment for noncompliance often requires assumptions about the comparability of those who did and did not comply with the assigned treatment. In our example, some approaches would require the untestable assumption that all possibly time-varying factors (e.g., past BPRS score) that predict both treatment switching and symptoms severity at one year have been measured at all times and appropriately adjusted for. This reliance on empirically unverifiable assumptions makes the problem of causal inference from a longitudinal study with baseline randomization bear a striking resemblance to the problem of causal inference from an observational study.

Second, even if all the information required for noncompliance adjustment is available and even if all subjects had completed the follow-up, the use of standard adjustment methods (e.g., stratification, regression analysis, matching) may introduce selection bias when estimating the effect of continuous treatment. This selection bias will occur if the reasons for noncompliance at any time are affected by prior treatment received and by unmeasured determinants of the outcome (Hernán *et al.*, 2004). In this article we will use two analytic methods that do not introduce selection bias when adjusting for noncompliance: inverse probability weighting and g-estimation (Robins and Hernán, 2008). An additional advantage of some forms of g-estimation is that a consistent estimation of the effect of continuous treatment in a randomized study does not require untestable assumptions about the predictors of noncompliance. The next two sections describe these methods and their application to our study.

Table 2. *Odds ratio (95% confidence interval) of treatment discontinuation by randomized arm **

	Atypical antipsychotic arm (person-visits:1,240)	*Conventional antipsychotic arm (person-visits:593)*
Age (in years)	0.99 (0.97, 1.02)	1.00 (0.98, 1.02)
Female sex	1.02 (0.57, 1.83)	0.92 (0.53, 1.60)
White race	1.16 (0.66, 2.02)	1.03 (0.61, 1.74)
Baseline BPRS		
< 25	Reference	Reference
25 – 44	1.11 (0.50, 2.44)	0.99 (0.53, 1.81)
≥ 45	0.74 (0.21, 2.66)	0.64 (0.21, 1.99)
Baseline GAF		
< 50	1.79 (0.59, 5.37)	0.81 (0.36, 1.82)
50 – 60	1.18 (0.39, 3.52)	1.19 (0.53, 2.65)
≥ 60	Reference	Reference
BPRS at current visit		
< 25	Reference	Reference
25 – 44	2.17 (1.15, 4.20)	1.31 (0.70, 2.46)
≥ 45	1.11 (0.27, 4.62)	0.45 (0.09, 2.28)
GAF at current visit		
< 50	0.84 (0.31, 2.27)	0.90 (0.39, 2.07)
50 – 60	1.08 (0.44, 2.67)	1.28 (0.63, 2.59)
≥ 60	Reference	Reference
Adverse event since last visit	2.36 (1.34, 4.16)	1.98 (1.16, 3.37)
Hospitalization since last visit	2.40 (1.14, 5.05)	0.78 (0.29, 2.10)

* BPRS: Brief Psychiatric Rating Scale score; GAF: global assessment of functioning score

In this section we have so far ignored the problem of incomplete follow-up. Table 3 shows the association between loss to follow-up and the measured covariates in the trial. We now combine our previous discussion on incomplete follow-up with the current discussion on noncompliance to clarify that the effect of interest is not simply the effect of continuous treatment, but the *effect of continuous treatment under complete follow-up*. Let $\overline{C}$ a subject's censoring history from baseline until the end of follow-up (one year in our trial). Then the effect of continuous treatment under complete follow-up is defined as

$$E[Y^{\overline{a}=\overline{1},\overline{c}=\overline{0}}]-E[Y^{\overline{a}=\overline{0},\overline{c}=\overline{0}}].$$

That is, we would like to estimate the effect of continuous treatment if everybody had fully adhered to the follow-up procedures specified in the study protocol. The reasoning that we used for adjustment for noncompliance also applies to incomplete follow-up: we may need to identify the joint predictors of complete follow-up and the outcome, and adjust for them using methods that do not introduce selection bias, such as inverse probability weighting and g-estimation.

Table 3. *Odds ratio (95% confidence interval) of loss to follow-up by randomized arm **

	Atypical antipsychotic arm (person-visits:1,240)	*Conventional antipsychotic arm (person-visits:593)*
Age (in years)	0.99 (0.97, 1.00)	0.99 (0.97, 1.02)
Female sex	0.90 (0.62, 1.30)	1.21 (0.70, 2.09)
White race	0.64 (0.45, 0.91)	0.59 (0.34, 1.01)
Baseline BPRS		
< 25	Reference	Reference
25 – 44	0.67 (0.43, 1.05)	1.18 (0.58, 2.38)
≥ 45	0.47 (0.21, 1.04)	0.69 (0.21, 2.25)
Baseline GAF		
< 50	1.34 (0.68, 2.64)	1.52 (0.51, 4.57)
50 – 60	1.43 (0.75, 2.71)	1.79 (0.62, 5.19)
≥ 60	Reference	Reference
BPRS at current visit		
< 25	Reference	Reference
25 – 44	1.40 (0.93, 2.12)	2.32 (1.21, 4.42)
≥ 45	2.02 (0.90, 4.55)	3.42 (1.17, 9.98)

GAF at current visit		
< 50	1.36 (0.73, 2.53)	1.12 (0.41, 3.02)
50 – 60	1.19 (0.69, 2.06)	0.77 (0.31, 1.92)
≥ 60	Reference	Reference
Adverse event since last visit	1.24 (0.86, 1.81)	0.97 (0.55, 1.70)
Hospitalization since last visit	1.93 (1.09, 3.42)	0.69 (0.20, 2.46)

* BPRS: Brief Psychiatric Rating Scale score; GAF: global assessment of functioning score

5 Inverse probability weighting

The ideal procedure to estimate the effect of continuous treatment under complete follow-up would be to (i) randomize subjects to either R=1 or R=0, (ii) force all subjects in the R=1 arm to take treatment A_k=1 at all times k, (iii) similarly force all subjects in the R=0 arm to take treatment A_k=0 at all times k, and (iv) force all subjects to be under follow-up for the entire duration of the study. If this protocol were enforced, the effect would be consistently estimated by the contrast $E[Y \mid \overline{A} = \overline{1}] - E[Y \mid \overline{A} = \overline{0}]$, i.e., the mean outcome among those who were continuously treated with A=1 minus the mean outcome among those who were continuously treated with A=0. Note that we do not need to restrict the analysis to the uncensored subjects (i.e., we do not have to calculate the mean conditional on C_k being 0 for all times k) because, under the protocol described above, nobody is censored (i.e., everybody's C_k=0 for all k).

Unfortunately for investigators, but luckily for the study subjects, forcing the subjects in the study population to adhere to any given protocol is near impossible in human studies. In most randomized studies, subjects are indeed randomized to either R=1 or R=0, but their subsequent values of received treatment A_k (and censoring status C_k) depend on their evolving covariate history, as shown in Table 2. Inverse probability weighting (IPW) is a method that uses the data from the actual randomized study population to simulate a "pseudo-population" in which, under certain assumptions outlined below, subjects are randomly assigned to receive either treatment A_k=1 or A_k=0, and remain under complete follow-up C_k=0, at each visit k, irrespective of their evolving covariate history. In such pseudo-population, we can calculate the difference in mean outcomes between those who were continuously treated with each treatment, $E[Y \mid \overline{A} = \overline{1}] - E[Y \mid \overline{A} = \overline{0}]$, to consistently estimate the effect of continuous treatment under complete follow-up.

The key assumption for IPW to be able to create such pseudo-population is that the investigators have measured all joint determinants of received treatment

A_k (or, equivalently, of compliance, with the assigned treatment R) and the outcome Y. This assumption of no unmeasured confounding, or exchangeability assumption, can be stated in many equivalent ways. Here is one of them: consider the pooled logistic model that we fit to estimate the associations shown in Table 2. The assumption of no unmeasured confounding says that any risk factor for Y that is not included in the model would be unassociated with the received treatment (odds ratio of 1) if it were included in the model. The assumption of no unmeasured confounding is also known as the assumption of sequential randomization because it is equivalent to assuming that, within levels of past treatment and covariate history, the value of the received treatment A_k at each visit was selected at random. In other words, IPW uses the assumption that the treatment regime in the study population was randomly assigned conditional on covariate history to simulate a pseudo-population in which the treatment regime is randomly assigned unconditional on covariate history. A similar assumption regarding censoring status in the study population is required to ensure that nobody is censored in the pseudo-population. IPW is, conceptually, the general version of standardization (Hernán and Robins, 2006). The method was first described by Robins in the contexts of compliance adjustment (Robins, 1993; Robins and Finkelstein, 2000) and of models for causal inference from complex longitudinal data (Robins, 1997). For a less technical description of the conditions required by IPW, see Hernán and Robins (2006) for non time-varying exposures, and Robins and Hernán (2008) for time-varying exposures. These conditions include the exchangeability assumption sketched above, and the positivity condition. That is, the requirement that the conditional probability of receiving either treatment A_k=1 or A_k=0 and of remaining under complete follow-up C_k=0 is greater than zero (i.e., positive) for all covariate histories and at all visits. We assume that positivity holds throughout this paper.

The pseudo-population is simulated by reweighting the contributions of each study subject by a subject-specific inverse probability weight for treatment and censoring. We first describe the weights and then explain how we estimated them in our study.

The treatment weights are defined as

$$W^A = \prod_{k=0}^{4} \frac{f[A_k \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1}, R]}{f[A_k \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1}, \overline{L}_k, R]}$$

where $f[A_k|\cdot]$is the conditional density of A_k. These weights are referred to as inverse probability weights because the denominator of the weight is, informally, the probability that the subject received her own treatment history given her past treatment and covariate history. The numerator of the weights, which cannot be a function of the time-varying covariates in $\overline{L}_k$, is merely a stabilizing factor to reduce the variance of the estimator. Thus, these inverse probability weights are known as stabilized weights. Under the exchangeability assumption that all joint predictors of treatment A_k and the outcome Y are included in $\overline{L}_k$, IPW simulates a pseudo-population in which treatment was randomly assigned conditional, at most, on past treatment history $\overline{A}_{k-1}$ but in which the effect of treatment on the outcome is the same as in the original (unweighted) study population.

The censoring weights are defined as

$$W^C = \prod_{k=0}^{4} \frac{\Pr[C_{k+1} = 0 \mid \overline{C}_k = \overline{0}, \overline{A}_k, R]}{\Pr[C_{k+1} = 0 \mid \overline{C}_k = \overline{0}, \overline{A}_k, \overline{L}_k, R]}$$

for those subjects that completed the follow-up, and are set W^C equal to zero for the others. These weights are also inverse probability weights because the denominator of the weight is the probability that the subject completed the follow-up given her treatment and covariate history. The numerator, again not a function of the time-varying variables in $\overline{L}_k$, helps stabilize the weight. Under the exchangeability assumption that all joint predictors of incomplete follow-up C_k and the outcome Y are included in $\overline{L}_k$, IPW simulates a pseudo-population in which everybody completed the follow-up (i.e., in which censoring was abolished) but in which the effect of treatment on the outcome is the same as in the original (unweighted) study population. We used these censoring weights to obtain the inverse probability weighted ITT estimate under complete follow-up in section 2. The inverse probability weight used in our IPW analysis is the product $W^A \times W^C$.

The inverse probability weights are unknown but can be estimated from the data. In our study, we estimated the denominator of the treatment weights W^A by fitting, separately in each arm R=r, the pooled logistic model for $\Pr[A_k = r \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1} = \overline{r}, \overline{L}_k = \overline{l}_k, R = r]$ described above to generate Table 2, except that the categorical variables for the BPRS and GAF scores were replaced by linear and quadratic terms for the score. Note that the additional assumption of no misspecification of the model used to estimate the weights is necessary for the method to provide consistent estimates.

Because few subjects who did not adhere to the assigned arm switched back to the originally assigned class of antipsychotics, we assumed that the probability of staying on the non assigned drug was 1 for the reminder of the follow-up. We estimated the numerator of the treatment weights W^A by fitting, separately in each arm R=r, a similar pooled logistic model that included only time since randomization as a covariate. The denominator of the censoring weights W^C was estimated by fitting, separately in each arm, a pooled logistic model for

$$\Pr[C_{k+1} = 0 \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1} = \overline{r}, \overline{L}_k = \overline{l}_k, R = r]$$

that included the covariates listed above. Each person contributed as many observations to the logistic model as visits she was under complete follow-up. The numerator of the censoring weights was estimated by fitting a similar pooled logistic model except that it included only time since randomization and A_k as covariates.

The assignment of the estimated inverse probability weights $W^A \times W^C$ to every subject in the population results in a pseudo-population in which, under the assumptions of IPW, all subjects undergo complete follow-up and the treatment received at each visit is randomly assigned conditionally on prior treatment history, but in which the effect of treatment is the same as in the original study population. The effect of continuous treatment under complete follow-up can now be estimated by simply restricting the analysis to the members of the pseudo-population that always adhered to their assigned treatment at baseline. That is, by conducting a "per protocol" analysis in the pseudo-population. Note that, if our outcome of interest had been continuously measured during the follow-up (e.g., survival) — as opposed to the situation in our study in which the outcome was only measured at the end of follow-up — then we could have conducted an "on treatment" analysis in the pseudo-population, i.e., censoring subjects at the first time they deviated from their assigned treatment.

In our study, we computed the inverse probability weighted mean of the outcome in subjects assigned to R=1 who remained on atypical antipsychotics for the entire duration of the study (N=205), and in subjects assigned to R=0 who remained on conventional antipsychotics for the entire duration of the study (N=56). The estimated weights $W^A \times W^C$ had a mean of 0.98 (their expected mean is 1) and their values ranged from 0.37 to 2.77 in the censored population. Equivalently, we fit the weighted least squares model

$$E[Y \mid R] = \theta_0 + \theta_1 R$$

to these 261 subjects, in which the parameter θ_1 is the causal effect of interest (in the scale of the mean difference). The effect of continuous treatment that we estimated by applying this IPW approach to our study was –1.53 (95% CI: –5.46, 2.39). To account for the correlation induced by the use of inverse probability weights, we used a generalized estimating equation (Liang and Zeger, 1986) program (e.g., option "repeated" in SAS proc genmod) that outputs a robust variance estimator. The 95% CIs obtained from the robust variance are conservative (i.e., their coverage is at least 95% in large samples). Using either bootstrapping or a variance estimator that explicitly incorporates how the weights were estimated would have resulted in slightly narrower confidence intervals, as discussed by Robins (1999).

The IPW approach described in the previous paragraph used data from all subjects to estimate the inverse probability weights, but the final (weighted) contrast only included data from subjects who fully adhered to baseline treatment and remained uncensored through the end of the study. However, it may be reasonable to argue that the mean outcome in pseudo-population subjects who took atypical antipsychotics most, but not all, of the time will be closer to the mean outcome in subjects who took atypical antipsychotics all the time than to that in subjects who never took atypical antipsychotics all the time. It could further be argued that the mean outcome of subjects who took atypical antipsychotics half of the time would be somewhere in between the mean outcomes of subjects who took atypical antipsychotics most of the time and those who almost never took atypical antipsychotics. In other words, we may believe that there is a dose-response relation between duration of use of atypical antipsychotics and symptom severity of schizophrenia. In our study, we chose to represent our dose-response beliefs by the model

$$E[Y^{\bar{a},\bar{c}=\bar{0}}] = \psi_0 + \psi_1 \sum_{m=0}^{M} d(a)_m ,$$

where $d(a)_m$ is an indicator for use of atypical antipsychotics on day m (1; yes, 0; no), and $\sum_{m=0}^{M} d(a)_m$ is the duration of atypical antipsychotic treatment from baseline at day m=0 to the end of follow-up at day m=M (in our study M=365). The parameter ψ_1 from this model measures the increase (or decrease) in the mean outcome per each additional time period on atypical antipsychotic treatment, and $\psi_1 \times 365$ measures the effect of continuous atypical antipsychotic use compared with no use of atypical antipsychotic treatment. This model is referred to as a marginal structural model (MSM) (Robins *et al.*, 2000) because it models the marginal (unconditional) mean of the counterfactual outcomes, and models for functionals of counterfactual outcomes are often referred to as structural models. Our MSM provides a mapping from any static treatment regime $\bar{a}$ to the mean response $E[Y^{\bar{a},\bar{c}=\bar{0}}]$ under the assumptions that the mean outcome is a linear function of the duration of treatment over the entire follow-up. If necessary, this assumption can be relaxed by proposing a more complex model. For example, one may relax the assumption of linear dependence by adding a quadratic term for $\sum_{m=0}^{M} d(a)_m$, and the assumption of equal effect of treatment taken at different times during the follow-up by replacing $\sum_{m=0}^{M} d(a)_m$ by, say, $\sum_{m=180}^{M} d(a)_m$.

The parameters of our MSM can be consistently estimated by estimating the parameters of the regression model

$$E\left[Y \mid \overline{A}\right] = \beta_0 + \beta_1 \sum_{m=0}^{M} d(A)_m$$

in the pseudo-population, that is, by fitting a weighted least squares model in which each study subject receives the estimate of her inverse probability weight $W^A \times W^C$. The estimated weights in the entire pseudo-population had mean 1.00 (range: 0.16, 6.89). Our estimate of the effect of continuous treatment from the MSM with a robust variance estimator was: –2.52 (95% CI: –6.07, 1.04).

In summary, we used IPW to estimate the effect of continuous treatment under complete follow-up by using two strategies. First, we eluded making any assumptions about the dose-response relation between use of atypical antipsychotics and symptoms severity by restricting the analysis to those who fully complied with their assigned treatment during the follow-up. Second, we made assumptions about the dose-response by fitting a linear model. The model allowed us to estimate the effect of interest by "borrowing information" from subjects that did not fully comply with their assigned treatment.

6 G-estimation

We now provide a conceptual description of g-estimation, another method to estimate the parameter ψ_1 from the structural model above. G-estimation was first described by Robins (Robins, 1989; Robins, 1993). We describe g-estimation in four steps.

First, note that the counterfactual outcomes $Y^{\bar{a}}$ are unmeasured predictors of the observed outcome Y. To see why, let us pick the counterfactual outcome $Y^{\bar{a}=\bar{0}}$ under no treatment. In our study, subjects with large positive values of $Y^{\bar{a}=\bar{0}}$ are those who would have developed a large increase in symptoms severity at the end of follow-up had they not received atypical antipsychotics, that is, the subjects with a worse prognosis. Thus the counterfactual outcome $Y^{\bar{a}=\bar{0}}$ can be viewed as an individual's characteristic that, if known at baseline, would provide information about the individual's underlying predisposition for a bad outcome.

Of course, the value of $Y^{\bar{a}=\bar{0}}$ is missing for most subjects but, for the sake of the argument in this and the next paragraph, suppose that the value of $Y^{\bar{a}=\bar{0}}$ were known for all subjects at baseline.

Second, note that the random treatment assignment *R* is, by definition of randomization, expected to be unassociated with any baseline variable. In particular, *R* is expected to be independent of $Y^{\bar{a}=\bar{0}}$, a baseline marker for severity. More precisely, the parameter η_1 in the logistic model

$$\operatorname{logit}\Pr[R=1 \mid Y^{\bar{a}=\bar{0}}]=\eta_0+\eta_1 Y^{\bar{a}=\bar{0}}$$

will be zero.

Third, note that we cannot know the value of the counterfactual outcome but, given several candidates, we have a method to rule out some of then. Imagine that an omniscient friend of ours added several variables to the study dataset. Let us refer to this collection of variables as *H*(*p*=1), *H*(*p*=2),… where *p* is an arbitrary index. Our friend guarantees us that only one of them is the counterfactual outcome $Y^{\bar{a}=\bar{0}}$, and challenges us to identify it. No big deal: we simply fit the model

$$\operatorname{logit}\Pr[R=1 \mid H(p)]=\eta_0+\eta_1 H(p)$$

separately for each of the variables *H*(*p*=1), *H*(*p*=2), etc., and choose the variable that results in the estimate of η_1 that is closest to zero. For example, if we find that *H*(*p*=3) minimizes the absolute value of the estimate of η_1, then we would say that *H*(*p*=3) equals the counterfactual outcome $Y^{\bar{a}=\bar{0}}$. Only one more piece is needed to complete this conceptual description of g-estimation.

Fourth, let us assume that the structural model

$$Y_i^{\bar{a}}=\psi_{0,i}+\psi_1\sum_{m=0}^{M} d(a)_m$$

holds for every individual *i* in the study. This deterministic subject-specific model is stronger than the model used in the previous section because the subject-specific model assumes that ψ_1 is the treatment effect for every single individual whereas the model in last section assumes that ψ_1 is the treatment effect averaged over all subjects (i.e., the model in last section is a “mean model”). Below we explain how to estimate the parameter ψ_1 by using the subject-specific model rather than the mean model of the previous section. However, we do not believe that the subject-specific model holds. We use it only for pedagogic reasons: the g-estimation procedure is easier to explain for the subject-specific model than for the mean model. Fortunately, it turns out that the g-estimation procedure described below to consistently estimate the parameter ψ_1 under the subject-specific model also estimates the parameter ψ_1 under the mean model, which is our actual aim. See Robins and Hernán (2008) for technical details.

Clearly, the subject-specific parameter $\psi_{0,i}$ is the counterfactual outcome under no treatment so the model can be rewritten as

$$Y_i^{\bar{a}} = Y_i^{\bar{a}=\bar{0}} + \psi_1 \sum_{m=0}^{M} d(a)_m ,$$

or

$$Y_i^{\bar{a}=\bar{0}} = Y_i^{\bar{a}} - \psi_1 \sum_{m=0}^{M} d(a)_m .$$

If the model holds for all counterfactual outcomes then it also holds for the observed outcome $Y=Y^{\bar{A}}$, which is just the counterfactual outcome under the actual treatment regime $\bar{A}$. Thus, we can rewrite the model as

$$Y_i^{\bar{a}=\bar{0}} = Y_i - \psi_1 \sum_{m=0}^{M} d(A)_m .$$

Under this model, if we knew the true value of the parameter ψ_1, then we could calculate the value of $Y^{\bar{a}=\bar{0}}$ for all subjects. But if we knew the true value of ψ_1, we would not need to use g-estimation! The whole point of this section is describing a method to estimate ψ_1, so how does it help us having learned that the true value of ψ_1 can be used, under our modeling assumptions, to calculate the counterfactual outcome $Y^{\bar{a}=\bar{0}}$? We have reached the core of g-estimation: We can simply guess the value of ψ_1, use our guessed value to calculate a candidate for counterfactual outcome $Y^{\bar{a}=\bar{0}}$, and then check whether our guess was right by examining the estimate of η_1 for our candidate variable. If our guess was not right, we keep guessing until we find the true value of ψ_1 by checking the value of the estimate of η_1. More formally, we compute

$$H_i(p) = Y_i - p\sum_{m=0}^{M} d(A)_m$$

for a sufficiently wide and fine range of values of *p*. The g-estimate of ψ_1 is the value of *p* that results in an *H*(*p*) that results in an estimate of η_1 equal to zero. Although for a linear structural model, like the one considered in this example, a closed form estimator exists and thus g-estimation does not require an actual search over the range of *p*, such search will generally be necessary for more complex models (e.g., accelerated failure time models for survival analysis).

We now explain how to obtain a 95% confidence interval for ψ_1. For each value of *p* we conduct a test of the null hypothesis $\eta_1 = 0$. An (1–α)% confidence interval for ψ_1 is formed by the values of *p* that result in estimates of η_1 for which the null hypothesis $\eta_1 = 0$ cannot be rejected at the α level. Most standard software packages to estimate the parameters of a logistic model will automatically perform a Wald test for such null hypothesis and output the corresponding p-value, but any other large-sample test (e.g., score test) may be used. In fact, the estimating equations for ψ_1 described in more theoretical presentations of g-estimation (and used in software written specifically for g-estimation) correspond to the score test for $\eta_1 = 0$ from the logistic model. See, for example, the Appendix of Hernán *et al.* (2005).

We have so far ignored the fact that some subjects were censored by incomplete follow-up before their outcome *Y* was measured, and thus cannot participate in the g-estimation procedure. To adjust for the possible selection bias introduced by this censoring, we conducted g-estimation in a pseudo-population simulated by assigning the inverse probability of censoring weights W^C to all subjects who were uncensored in the study population. The estimated weights W^C had a mean of 0.99 (range: 0.63, 2.53). The effect of continuous treatment that we obtained by applying (inverse probability weighted) g-estimation to our study was –1.50 (95% CI: –6.84, 3.84).

The random treatment assignment *R* is an example of an instrumental variable or instrument. The method of g-estimation described above exploits the expected independence between the counterfactual outcome and the instrument to estimate the parameters of structural models. Hence g-estimation is the general version of instrumental variable estimation for time-varying treatments (Hernán and Robins, 2006). For applications of g-estimation in the analysis of randomized experiments, see Mark and Robins (1993) and Cole and Chu (2005).

Because a g-estimation analysis, like an ITT analysis, relies on the actual randomization and thus does not require the untestable assumption of sequential randomization of treatment given the measured covariates, it can be referred to as a *randomized analysis*. In contrast, an IPW analysis that requires the untestable assumption of sequential randomization of treatment given the measured covariates can be referred to as an *observational analysis*. (This classification applies to assumptions regarding treatment, not censoring, because both g-estimation and ITT analyses require IPW to adjust for censoring and thus both require the untestable assumption of sequential randomization of censoring given the measured covariates.) However, if one is willing to make the assumption of sequential randomization of treatment, then g-estimation can be easily modified to take advantage of this assumption. We now describe how to conduct an *observational analysis* based on g-estimation.

The assumption of sequential randomization of treatment given the measured covariates implies that no baseline predictors of the outcome, other than those measured and included in $\overline{L}_k$, will predict treatment A_k at any visit k. Consider the logistic model for $\operatorname{logit}\Pr[A_k = a_k \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1} = \overline{r}, \overline{L}_k = \overline{l}_k, R]$ that we fit to estimate the probabilities in the denominator of the inverse probability of treatment weight W^A. The assumption of sequential randomization says that the coefficient of any baseline risk factor that is added to the model as a covariate is expected to be zero (odds ratio equal to 1). In particular, if we add the covariate $Y^{\overline{a}=\overline{0}}$ and fit a logistic model for

$$\Pr[A_k = a_k \mid \overline{C}_k = \overline{0}, \overline{A}_{k-1} = \overline{r}, \overline{L}_k = \overline{l}_k, R, Y^{\overline{a}=\overline{0}}],$$

the parameter η_1 for $Y^{\overline{a}=\overline{0}}$ is expected to be zero (for comparability with the two models specified for IPW, one per randomization arm, we included product terms between R and all the other covariates except for $Y^{\overline{a}=\overline{0}}$ in the model above). Of course, we do not know the value of $Y^{\overline{a}=\overline{0}}$ for most subjects but we can use our structural model

$$Y_i^{\overline{a}=\overline{0}} = Y_i - \psi_1 \sum_{m=0}^{M} d(A)_m$$

to propose candidates

$$H_i(p) = Y_i - p\sum_{m=0}^{M} d(A)_m$$

for a sufficiently wide and fine range of values of p. Again, the g-estimate of ψ_1 is the value of p that results in an $H(p)$ that results in an estimate of η_1 equal to zero. In our study, the effect of continuous treatment that we obtained by using an observational analysis based on (inverse probability weighted) g-estimation to our study was –2.64 (95% CI: –6.12, 0.84).

Finally, a comment about statistical efficiency in g-estimation. The g-estimation procedures described above, and the analysis of our study, are based on adding the covariate $H(p)$ to the logistic model. But note that the rationale behind g-estimation would carry through if we added a function of $H(p)$ (say, its log), rather than $H(p)$ itself, to the model. In fact, it can be shown that using certain functions of $H(p)$ might result in a narrower confidence interval around the g-estimate compared with using $H(p)$. However, although g-estimation based on the estimating function $H(p)$ is possibly inefficient, it is also easy to carry out. On the other hand, the efficient g-estimator involves functions of $H(p)$ that are hard to compute and whose description is beyond the scope of this paper (Robins, 1993). In our study, estimates based on several simple functions of $H(p)$ (e.g., its log) were similar to the ones shown here using $H(p)$ (data not shown).

Table 4. *Estimates of the causal effect of atypical antipsychotics on change in score of the Brief Psychiatric Rating Scale (BPRS) at one year after randomization. See text for details.**

	Initiation vs. No initiation			***Continuous use vs. No use***			
Method	Pseudo-ITT	Complete-case ITT	ITT + IPW for censoring	IPW for censoring and treatment		G-estimation for treatment + IPW for censoring	
Assumptions							
Censoring							
Sequential randomization + correctly specified model	Un-conditional	Un-conditional	Conditional	Conditional	Conditional	Conditional	Conditional
Treatment							
Sequential randomization (conditional) + correctly specified model	No	No	No	Yes	Yes	No	Yes
Correctly specified structural (dose-response) model	No	No	No	No	Yes	Yes	Yes
No. of subjects in final contrast	634 †	365 ‡	365 ‡	261§	365 ‡	365 ‡	365 ‡
Effect estimate	–1.05	0.42	–0.86	–1.53	–2.52	–1.50	–2.64
95% confidence interval #	-3.26, 1.16	-2.36, 3.19	-3.88, 2.15	-5.46, 2.39	-6.07, 1.04	-6.84, 3.84	-6.12, 0.84

* ITT: intention-to-treat; IPW: inverse probability weighting

† Subjects with at least one post-randomization BPRS score recorded. The last available BPRS score was used

‡ Subjects with complete follow-up data

§ Subjects with complete follow-up data and full adherence to the assigned treatment

# Conservative 95% confidence intervals except for the pseudo-ITT and complete-case ITT analyses

7 Discussion

Table 4 summarizes all the effect estimates that we have presented throughout the article. A cursory examination of the table shows that none of the estimates reached traditional statistical significance (i.e., all 95% CIs include null value) and thus it can be argued that none of them is different from zero. However, for the purposes of this discussion, we will regard these point estimates as coming from a larger study with much narrower confidence intervals.

The table has two parts: The first 3 columns show different estimates of the ITT effect; the last 4 columns show different estimates of the effect of continuous treatment. Let us discuss the ITT effect estimates first. An ITT analysis is usually the primary, and often the only, analysis of randomized experiments. As discussed above, there are good reasons why the ITT effect needs to be reported. However, the ITT effect is often presented as a straightforward analysis even when that is not the case. For example, in our study many subjects did not complete the follow-up and thus their outcome was unknown. As a result, any ITT-like analysis requires some additional assumptions to estimate the ITT effect under complete follow-up. The first two columns of Table 4 present the estimates from two common ITT-like analyses: the "last available observation carried forward" or pseudo-ITT analysis, and the "complete-case" ITT analysis.

A pseudo-ITT analysis assumes that 1) those with and without complete follow-up are exchangeable, and 2) the ITT effect of treatment is the same whether we use a measurement of the outcome at 2 weeks or at 12 months since baseline. When applied to our study, the pseudo-ITT effect estimate was –1.05, which may be explained by either a sustained beneficial effect of atypical antipsychotics compared with conventional ones, or by an early beneficial effect of atypical antipsychotics followed by worsening of the symptoms leading to drop-out of subjects on atypical antipsychotics.

A complete-case analysis eliminates assumption 2) of the pseudo-ITT analysis but still assumes that those with and without complete follow-up are exchangeable. When applied to our study, the complete-case ITT effect estimate was 0.42, which may be explained by either a harmful effect of atypical antipsychotics, or by a differential drop-out of subjects doing badly on conventional antipsychotics.

Both ITT-like analyses make the assumption that there is no selection bias due to incomplete follow-up or, equivalently, that censoring by incomplete follow-up was randomly assigned during the follow-up. This assumption of unconditional sequential randomization of censoring is a strong assumption. We therefore considered a weaker assumption: censoring was sequentially randomized within levels of (conditionally on) the measured time-varying covariates. If investigators are willing to assume the weaker, conditional assumption — and they will if they were willing to assume the stronger, unconditional one — then they can use IPW to adjust for selection bias explained by the measured time-varying factors. The third column of Table 4 presents the ITT analysis with IPW adjustment for incomplete follow-up. The estimate of –0.86, which may be explained by a true beneficial ITT effect of atypical antipsychotics compared with conventional ones, can be affected by insufficient adjustment for selection bias. However, because the unadjusted estimate was 0.42 and the adjusted one is –0.86, it is possible that further adjustment (if it were possible) by unmeasured factors would have made the estimate even more negative, i.e., the true ITT effect would be stronger than our estimate, which can then be viewed as a conservative one. Of all three estimates of the ITT effect under complete follow-up that are shown in Table 4, the inverse probability weighted ITT analysis makes the weakest assumptions. Although in this particular study the inverse probability weighted ITT and the pseudo-ITT estimates happened to be very close, this coincidence cannot be generally expected.

The ITT effect is the effect of treatment assignment or initiation under a particular pattern of compliance. In our study, the ITT is the effect of initiating atypical antipsychotic therapy compared with conventional antipsychotic therapy. As discussed above, the effect of treatment initiation may be quantitatively or even qualitatively different from the effect of continuous treatment when many study subjects do not adhere to the treatment after initiation. This dependence of the magnitude on the ITT effect on the degree of compliance may make it hard to transport to other populations with different levels of compliance, or even to the same population at different times (for example, the publication of the study results may affect compliance in the same population in which the study was conducted). It also makes the ITT approach a dangerous one for identifying potential harmful effects. In placebo-controlled safety studies, one needs to be careful when presenting ITT effects because null effect estimates may merely reflect substantial noncompliance rather than the absence of adverse effects. In fact, the ITT effect may suggest that the treatment of interest is less toxic than the comparator even when both treatments have similar toxicity. Thus an ITT analysis cannot generally be the only analytic approach for a randomized experiment (e.g., a large simple trial) with a safety outcome or lack of a placebo control.

As a complement to the ITT effect under complete follow-up, Table 4 also shows our estimates of the effect of continuous treatment under complete follow-up, that is, the effect if all subjects had adhered to their assigned treatment for the entire follow-up. Unfortunately, to estimate this effect we need assumptions beyond those necessary to estimate the ITT effect under complete follow-up. At least one of two types of assumptions needs to be made: (i) sequential randomization of treatment within levels of the measured covariates, or (ii) a dose-response model. If only assumption (i) is made, then IPW is needed. If only assumption (ii) is made, then g-estimation is needed. If both assumptions are made, then either IPW or g-estimation can be used. Let us see each of these cases separately.

Under assumption (i), one can use IPW to simulate a pseudo-population in which treatment is given at random. In this pseudo-population, a subject's prognosis is unrelated to the treatment regime she receives during the follow-up. Thus, to estimate the effect of continuous treatment, one only needs to restrict the analysis to subjects who always adhered to their baseline assignment. In the pseudo-population of our study, we compared the average outcome of subjects assigned to atypical antipsychotics with that of subjects assigned to conventional antipsychotics. As expected, the IPW estimate of the effect of continuous treatment until complete follow-up (–1.53) suggests that atypical antipsychotics result in a greater symptomatic improvement than that suggested by the IPW estimate of the ITT effect (–0.86). The estimate of –1.53 will be biased if the covariates used to estimate the inverse probability weights do not include all important confounders of the treatment effect, or if they are measured with error, or if the weight model is misspecified. Note that, in our study, treatment may actually change in between visits while the adjusting factors are only measured at the visits, which may result in insufficient adjustment.

Under assumption (ii), one can use g-estimation to conduct a generalized instrumental variable analysis that does not require any assumptions regarding sequential randomization of treatment. That is, g-estimation may consistently estimate the effect of continuous treatment even in the presence of unmeasured confounding for the treatment effect. However, the method requires a structural model for the effect of treatment on the outcome. In our study, we assumed that the effect of atypical antipsychotic use on the outcome is a linear function of the duration of treatment. Our estimated effect of continuous treatment in the sixth column of Table 4, –1.50, depends critically on that dose-response assumption. To estimate the sensitivity of the estimate to the assumption of correct dose-response specification, we estimated the effect of continuous treatment under alternative models (data not shown) and found that the model used for the estimates in Table 4 were the closest to the null value. Thus our results are likely to be conservative.

Interestingly, we found that both the IPW-based observational analysis (column 4 of Table 4) and the g-estimation-based randomized analysis (column 6) yielded similar estimates of the effect of continuous treatment, even though the validity of each method rests on a qualitatively different assumption. Leaving aside sampling variability, this coincidence may reflect either that both assumptions were approximately correct, or that both were wrong in such a way that the bias was in the same direction and of the same magnitude.

Finally, one can combine assumptions (i) and (ii) to estimate the effect of continuous treatment by using either IPW or g-estimation. In our study, the estimates are –2.52 (column 5 of Table 4) and –2.64 (column 7), respectively. If either of assumptions (i) or (ii) does not hold, then both estimates will be invalid. On the other hand, if both assumptions hold, the estimates are expected to have narrower confidence intervals than the corresponding IPW and g-estimation ones that relied only on either assumption (i) or (ii).

Table 4 does not include any estimates from standard methods for bias adjustment, such as regression or matching. In our study, for comparison purposes, we fit a standard (unweighted) linear regression model for the mean outcome conditional on a summary of treatment history (number of days on atypical antipsychotics as in our structural linear model) and summaries of the baseline and time-varying factors (i.e., BPRS, functional status, hospitalizations, and toxicity). The estimate of continuous effect was 1.10 (95% CI: –3.29, 5.49). Unlike all estimates of continuous effect in Table 4, this standard estimate suggests that atypical antipsychotics are inferior to conventional ones. However, the validity of standard statistical techniques requires not only assumptions (i) and (ii), but also (iii) the assumption that the time-varying factors (e.g., the measured values of BPRS between baseline and the end of the study) are not affected by the treatment itself. If assumption (iii) does not hold, the standard estimate is expected to be biased (Hernán *et al.*, 2004). In most cases, like in our study, assumption (iii) will be hard to defend if we actually believe that treatment may affect the outcome.

The IPW and g-estimation methods presented here can be extended in a variety of directions. For example, as originally described by Robins, IPW can be applied to settings with non dichotomous treatments (Haight *et al.*, 2005; Cotter *et al.*, 2008) and with failure time outcomes (survival analysis) (Hernán *et al.*, 2001; Cole *et al.*, 2003), and to the estimation of the effect of dynamic treatment regimes (Hernán *et al.*, 2006). The extension to dynamic regimes is crucial because in some cases estimating the effect of continuous treatment (a non-dynamic regime) may be of little interest. For example, if many subjects stop taking the treatment because it causes serious adverse effects, one would not want to estimate the effect under the non-dynamic regimes "always adhere to the baseline treatment" but rather under the dynamic regimes "adhere to the baseline treatment unless adverse effects occur". Further, when certain types of patients will always discontinue treatment given certain adverse events, then estimating the effect under non-dynamic regimes like "always adhere to the baseline treatment" is problematic because the positivity assumption is violated. The consideration of dynamic regimes may make it more likely that the positivity assumption holds. In the analyses presented here, we chose the effect of continuous treatment for pedagogic, rather than clinical, reasons.

In summary, we recommend that a table similar to Table 4 is generated from randomized experiments with substantial noncompliance or loss to follow-up. Because each approach in the table has relative advantages and disadvantages, and depends on a different combination of assumptions, a general agreement among all estimates will strengthen our confidence in the results. On the other hand, the existence of serious discrepancies will provide some guidance regarding important sources of bias in the study that might not have been identified otherwise. Of course, implementing our recommendation would require major modifications to current practice, and to the protocols of randomized experiments. For example, to conduct the analyses that require the assumption of sequential randomization, the protocols of randomized experiments would need to include plans to measure post-randomization variables. To go beyond the ITT (or pseudo-ITT) analysis, the protocol would need to include a more complex statistical analysis plan and to collect more precise adherence information. To assess the sensitivity of the estimates to model specification in analyses that require the assumption of correct dose-response specification, the statistical plan would need to specify a variety of dose-response models. However, it seems to us that the added complexity is necessary to take full advantage of the substantial resources that are usually invested in a randomized experiment.

References

Cole, SR, Chu, H (2005): Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemp Clin Trials* 26:300-10

Cole, SR, Hernán, MA, Robins, JM, et al (2003): Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol* 158:687-94

Cotter, D, Zhang, Y, Thamer, M, et al (2008): The effect of epoetin dose on hematocrit. *Kidney Int* 73:347-53

Haight, T, Tager, I, Sternfeld, B, et al (2005): Effects of body composition and leisure-time physical activity on transitions in physical functioning in the elderly. *Am J Epidemiol* 162:607-17

Hernán, MA (2004): A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 58:265-71

Hernán, MA, Brumback, B, Robins, JM (2001): Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc* 96:440-448

Hernán, MA, Hernández-Díaz, S, Robins, JM (2004): A structural approach to selection bias. *Epidemiology* 15:615-25

Hernán, MA, Cole, SR, Margolick, J, et al (2005): Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmcoepidemiol Drug Saf* 14:477-91

Hernán, MA, Lanoy, E, Costagliola, D, et al (2006): Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol* 98:237-42

Hernán, MA, Robins, JM (2006): Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 60:578-86

— and — (2006): Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 17:360-72

Lesko, SM, Mitchell, AA (2005): The use of randomized controlled trials for pharmacoepidemiology studies, in Pharmacoepidemiology. Edited by Strom, BL. West Sussex, England, John Wiley & Sons Ltd, p.599-610

Liang, KY, Zeger, SL (1986): Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22

Mark, SD, Robins, JM (1993): A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Control Clin Trials* 14:79-97

Overall, JE, Gorham, DR (1962): The brief psychiatric rating scale. *Psychol Rep* 10:799-812

Robins, JM (1989): 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. Edited by Sechrest, L, Freeman, H, Mulley, A. U.S. Public Health Service, National Center for Health Services Research, p.113-159

— (1993): Analytic methods for estimating HIV treatment and cofactor effects, in Methodological Issues of AIDS Mental Health Research. Edited by Ostrow, DG, Kessler, R. New York, Plenum Publishing, p.213-290

— (1997): Marginal structural models. Proceedings of the American Statistical Association. Section on Bayesian Statistical Science, Alexandria, VA, American Statistical Association, p.1-10

— (1999): Marginal structural models versus structural nested models as tools for causal inference, in Statistical Models in Epidemiology: The Environment and Clinical Trials. Edited by Halloran, ME, Berry, D. New York, Springer-Verlag, p. 95–134

Robins, JM, Finkelstein, DM (2000): Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56:779-88

Robins, JM, Hernán, MA (2008): Estimation of the causal effects of time-varying exposures, in Longitudinal Data Analysis. Edited by Fitzmaurice, G, Davidian, M, Verbeke, G, Molenberghs, G. New York, Chapman and Hall/CRC Press, p.553-599

Robins, JM, Hernán, MA, Brumback, B (2000): Marginal structural models and causal inference in epidemiology. *Epidemiology* 11:550-60

Tunis, SL, Faries, DE, Nyhuis, AW, et al (2006): Cost-effectiveness of olanzapine as first-line treatment for schizophrenia: results from a randomized, open-label, 1-year trial. *Value Health* 9:77-89