

# Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation

Sander Greenland<sup>a</sup>, Stephan Lanes<sup>b</sup> and Michele Jara<sup>b</sup>

**Background** Randomized trials provide pivotal evidence for evaluation and approval of therapies. Nonetheless, such trials are often plagued by noncompliance, especially in the form of premature discontinuation of treatment. While intent-to-treat (ITT) analysis can provide valid tests of no-effect hypotheses, some trials may make ITT analysis impossible by ceasing follow-up when patients go off assigned treatment. Furthermore, estimates based on ITT, on-treatment, or per-protocol comparisons can seriously understate harm or benefit.

**Purpose** To show how g-estimation based on randomization status is a natural generalization of ITT null testing to estimating efficacy from trials with important discontinuation or noncompliance.

**Methods** We contrast with an analysis of the effect of a tiotropium inhaler on the occurrence of chronic obstructive pulmonary disease (COPD) events in a six-month double-blind placebo-controlled trial of 1829 patients with good but imperfect compliance.

**Results** The covariate-adjusted point estimates, 95% confidence limits (CL), and null *P*-values comparing expected COPD event times in placebo versus tiotropium patients were: ITT, 1.21, CL=1.02, 1.43, *P*=0.027; on-treatment, 1.27, CL=1.06, 1.52, *P*=0.009; per-protocol, 1.36, CL=1.13, 1.63, *P*=0.001; and g-estimation, 1.31, CL=1.03, 1.72, *P*=0.027. Thus g-estimation preserved the ITT test of the null, but exhibited more uncertainty about the size of the tiotropium effect than the other methods. In particular, it allowed for a much larger potential effect than did ITT analysis, but produced a much larger null *P* than exhibited by per-protocol analysis.

**Limitations** Like ITT analysis, g-estimation requires all patients be followed to the end of the trial protocol, regardless of whether they comply with the protocol. Like on-treatment and per-protocol analyses, it also requires accurate compliance information be recorded.

**Conclusion** G-estimation should become a standard procedure for the analysis of trials with noncompliance. Software to do so is available in major packages, and the procedure is easily coded for other packages. *Clinical Trials* 2008; 5: 5–13. <http://ctj.sagepub.com>

## Introduction

Randomized trials are crucial to the scientific evaluation of therapies, and are mandatory for most drug and device approvals. Such trials can be seriously compromised by noncompliance, however.

Premature discontinuation of treatment is a common example of the problem. Patients who discontinue before the assigned protocol ends often have a poorer prognosis than those who continue assigned treatment, and tend to be more common in the treatment arm experiencing more side effects

<sup>a</sup>Epidemiology and Statistics, University of California, Los Angeles, California, <sup>b</sup>Boehringer-Ingelheim Pharmaceuticals Inc., Ridgefield, Connecticut

**Author for correspondence:** Sander Greenland, Epidemiology and Statistics, University of California, Los Angeles California E-mail: [lesdomes@ucla.edu](mailto:lesdomes@ucla.edu)

**Potential Conflicts of Interest:** This work was supported by Boehringer-Ingelheim Pharmaceuticals. Dr. Greenland is a consultant for and Drs. Lanes and Jara are employees of Boehringer-Ingelheim Pharmaceuticals

or fewer benefits. As a consequence, patient groups remaining on assigned treatment will become increasingly imbalanced on prognostic factors as the trial progresses, with the result that analyses based only on patients completing assigned treatment can be severely confounded.

Intent-to-treat (ITT) analysis addresses such problems by comparing patients based on their assigned (intended) treatment, without regard to subsequent compliance. This comparison is done to preserve randomization of the compared groups. Under mild assumptions, ITT analysis will provide a valid test of the null hypothesis that the treatment is ineffective. Unfortunately, estimates from these comparisons represent a mix of compliance and treatment effects, and can seriously underestimate treatment effects when noncompliance is high.

Direct ITT estimates are often defended on the grounds that they reflect the effect of offering treatment, given that noncompliance is inevitable. This defense assumes that noncompliance patterns in practice will be quite similar to those seen in the trial, which is a questionable assumption. Patients that consent to enter trials are not typical with respect to motivation, and the extensive patient contact in a trial can enhance compliance. Furthermore, a trial patient will know from informed consent that the assigned treatment may be of no value (the new treatment is as yet unproven, and the control treatment may be a placebo); whereas in practice there is a presumption that the treatment is helpful. These knowledge differences could affect compliance behavior. Thus the generalizability of direct ITT estimates to clinical practice is doubtful.

In contrast to the difficulty of generalizing effects of offering treatment, generalization of physiologic effects of treatment from trials can often be argued based on the mechanism of action. The problem then is how to develop plausible estimates of these effects from randomized trials with noncompliance. Commonly recommended approaches include on-treatment (OT) analyses, in which patients are censored at the time they discontinue their assigned treatment; as-treated (AT) analyses, in which patients (or patient-times) are grouped according to their received treatment; and per-protocol (PP) analyses, in which only those patients who adhere to their assigned treatment are included in the analysis. Indeed, D'Agostino *et al.* [1] suggest 'that the PP analysis may be preferable [to ITT] in the non-inferiority trial setting,' although they also advise that an ITT analysis be conducted for reference.

Unlike ITT comparisons, OT and PP comparison groups are defined in part by post-randomization events affected by treatment and prognostic

factors, including discontinuation and other forms of noncompliance. Defining comparison groups based on post-randomization events can invalidate both tests and estimates [2–5]. It is thus surprising that some reports in major journals still emphasize OT or PP analyses [1]. Worse, some trials discontinue follow-up after treatment discontinuation, rendering ITT analysis impossible.

To summarize, noncompliance can invalidate effect estimates from conventional ITT, OT, and PP analyses. In response, several methods have been developed that can provide valid estimates under certain assumptions. Best known are classical instrumental-variable techniques, which were originally developed for econometric studies and later adapted to address noncompliance problems in clinical and field trials [6–9]. The reasoning behind these methods is subtle and the assumptions they require may seem restrictive compared to familiar direct comparisons of groups. Furthermore, their implementation is not straightforward when compliance corresponds to possibly unique and complex individual patterns of treatment.

We review here a less restrictive method based on *structural-nested modeling*, which involves only direct treatment group comparisons for estimation, thus preserving randomization, while allowing for arbitrarily complex compliance patterns [2,10,11]. This approach generalizes ITT testing and instrumental variables to estimate efficacy from trials with time-varying compliance. It has several advantages over conventional methods, but remains rare outside of statistical articles [12,13]. Our goal here is to explain the method in a way that will encourage broader use, especially as an alternative to PP analysis. The underlying model has a form similar to that of proportional-hazards models, and the model-fitting process (called g-estimation) employs familiar models for covariate adjustment. We illustrate the method with an analysis of the effect of tiotropium on chronic obstructive pulmonary disease (COPD) events in a large trial.

## A trial of tiotropium

Tiotropium is a once-daily anticholinergic bronchodilator. A parallel-group, double-blind, balanced randomized trial was used to study the effect of tiotropium versus placebo on COPD exacerbations among patients in the U.S. Veterans Administration (Trial 205.266). The study is described in detail elsewhere [14,15] and briefly summarized here.

The protocol was approved by local institutional review boards. Inclusion criteria were a clinical diagnosis of COPD, age at least 40 years, smoking history of at least 10 pack-years, and a forced expiratory volume in 1 s ( $FEV_1$ )  $\leq 60\%$  of predicted and  $\leq 70\%$  of the forced vital capacity (FVC). Exclusion criteria included asthma, COPD exacerbation in the preceding month, myocardial infarction within the prior six months, serious cardiac arrhythmia or hospitalization for heart failure within the prior year, known moderate to severe renal impairment, moderate to severe symptomatic prostatic hypertrophy or bladder-neck obstruction, or narrow-angle glaucoma.

Patients were provided written informed consent. Following a screening visit, participants were randomized to receive either tiotropium 18 mcg once daily or matching placebo capsules for 6 months, delivered via the HandiHaler dry-powder inhaler. Visits were scheduled at 3 and 6 months afterward. Patients were permitted to continue all previously prescribed respiratory medications other than inhaled anticholinergics, and were asked to attend all study visits and provide medical information even if study drug was discontinued prior to 6 months. Telephone contacts occurred monthly between visits. Drug compliance was assessed with counts of returned capsules and diary cards to record drug use. Spirometry was conducted before and 90 min after study drug administration at baseline, and at 3 and 6 months.

A COPD exacerbation was defined as 'a complex of respiratory symptoms (increase or new onset) of more than one of the following: cough, sputum, wheezing, dyspnea, or chest tightness with a duration of at least three days requiring treatment with antibiotics and/or systemic steroids and/or hospital admission' [14]. Hospitalizations were confirmed from available medical records. COPD exacerbations were reported on an exacerbation-specific case report form. Exacerbation events considered serious, such as hospitalizations (see following section for definition of serious), were also reported on a serious adverse event case report form. Data from both case report form sources were reconciled to insure accuracy and consistency. The adverse event terms reported by the investigational sites were coded according to the Medical Dictionary for Regulatory Activities while the trial remained blinded.

## Shared assumptions

All the methods we consider assume that if treatment has no effect, treatment assignment (intent) will be independent of the trial outcomes; in other words, there is no effect of assignment

other than through its impact on received treatment. This assumption is sometimes called 'exclusion restriction' [7] and is often enforced by double-blind assignment when the latter is successful (i.e., when patients and providers cannot identify the treatment). We also assume that violations of the assigned protocol (e.g., going off assigned treatment) are accurately recorded. Finally, the methods we discuss assume that the number of events per treatment arm is large enough for the statistical approximations to be adequate. The precise number required varies with method and with the level of accuracy desired, but the example numbers (over 250 events per arm) are more than adequate for the methods considered here.

## Conventional analyses of the trial

In this section, we contrast several approaches seen in published analyses of trials like the one we describe. There were several relevant overlapping endpoints to the trial. For simplicity we examine only time to first COPD exacerbation during the trial, or 'COPD time' for short. Other COPD endpoints such as time to first COPD hospitalization provide similar results. Table 1 presents basic data summaries from the trial. Table 2 provides a summary of statistical results from the methods described below.

### Conventional models

The ITT indicator  $R$  is defined as  $R=1$  if assigned tiotropium,  $R=0$  if not. The PP treatment indicator  $X$  is defined as  $X=1$  if the person receives tiotropium throughout the study, 0 if not (non-compliers are excluded).  $Z$  is the set of baseline covariates used for adjustment. We follow standard

**Table 1** Summary data from clinical trial of tiotropium versus placebo [14]

	Tiotropium arm ( $R=1$ )	Placebo arm ( $R=0$ )
No. assigned	914	915
No. with COPD event	251 (27%)	294 (32%)
No. discontinuing before 6 mos. (discarded by PP analyses)	208 (23%)	282 (31%)
Percentage of time off treatment (discarded by OT analyses)	5%	11%
No. censored before a COPD event <sup>a</sup>	33	37

<sup>a</sup>Deaths or drop-outs before first COPD event, up to 6 months (180 days).

**Table 2** Summary of analyses of the ratio of expected time to first COPD event after randomization in tiotropium and placebo groups, adjusted for baseline covariates using a proportional-hazards model with constant covariate-specific event rates<sup>a</sup>; under this model,  $e^{-\beta}$  is both the ratio of expected event times in treated versus untreated, and the ratio of event rates in untreated versus treated

Analysis method <sup>a</sup>	Estimate of time ratio $e^{-\beta}$		Adjusted 95% CL for time ratio	Adjusted null $P$ -value
	Unadjusted	Adjusted		
Intent-to-treat	1.21	1.21	1.02, 1.43	0.027
On-treatment	1.22	1.27	1.06, 1.52	0.009
Per-protocol	1.28	1.36	1.13, 1.63	0.001
G-estimation <sup>b</sup>	1.31	1.31	1.03, 1.72	0.027
– weighted <sup>b</sup>	1.31	1.31	1.02, 1.74	0.030

<sup>a</sup>Baseline adjustment covariates are age, sex, predicted FEV<sub>1</sub>, duration of COPD, race (black vs. other), current smoking indicator, and pack-years of smoking; indicators for inhaled and oral corticosteroids and xanthine use in previous month; and numbers of corticosteroid courses, antibiotic courses, and unplanned medical visits (to physician or health-care facility) in previous year.

<sup>b</sup>See Table 3.

practice and use Cox proportional-hazards models [16] to analyze the effects of having  $R=1$  versus  $R=0$  (the ITT effect) and of having  $X=1$  versus  $X=0$ . To simplify interpretations, we assume constant covariate-specific rates (hazards) of COPD events. This assumption makes the inverse rate ratios equal to COPD-time ratios, and may be a reasonable approximation given the short trial duration relative to the condition under study (COPD). Regardless, the model fits well by conventional tests, and dropping the assumption changes the estimated ratios by <1%.

Under this model, the ITT analysis assumes that persons with assignment  $R$  at level  $r$  and baseline covariates  $Z$  at level  $z$  have an expected time to a COPD event of

$$1/\exp(\alpha + \beta r + \gamma z) = \exp(-\alpha - \beta r - \gamma z). \quad (1)$$

As usual for a Cox model,  $e^{\beta}$  is the  $Z$ -adjusted ratio of the event rates when  $R=1$  versus when  $R=0$ . In addition,  $1/e^{\beta} = e^{-\beta}$  is the  $Z$ -adjusted ratio of expected time to a COPD event when  $R=1$  versus when  $R=0$ . Because  $R$  is randomized, we can interpret  $e^{-\beta}$  as the factor by which assignment to tiotropium extended or contracted the average time to the COPD event relative to placebo assignment within levels of  $Z$ .

In a parallel fashion, PP analyses assume that, among persons who comply with their assigned treatment, those with actual treatment  $X$  at level

$x$  and baseline covariates  $Z$  at level  $z$  have a constant event rate of  $\exp(\alpha + \beta x + \gamma z)$ , and an expected time to a COPD event of

$$1/\exp(\alpha + \beta x + \gamma z) = \exp(-\alpha - \beta x - \gamma z), \quad (2)$$

where  $e^{-\beta}$  is the  $Z$ -adjusted ratio of expected time to a COPD event when  $X=1$  versus when  $X=0$ . Unlike for  $R$ , however, we do not have randomization of  $X$ . Thus we cannot be sure that the ratio  $e^{-\beta}$  reflects only the effect of  $X$  on the expected time;  $e^{-\beta}$  may also reflect effects of uncontrolled covariates that influence both  $X$  and expected time (confounding). On-treatment analyses also assume model (2), but use data from partial compliers up to their end of compliance to help estimate  $e^{-\beta}$ .

The rationale for ITT analysis (i.e., using  $R$  instead of  $X$ , or Model 1 instead of Model 2) can now be stated as follows: If the only way that  $R$  can affect time to event is through its effect on  $X$  (treatment received), then any association of  $R$  with time to event (i.e., having  $\beta \neq 0$  in Model 1) must be due to an effect of  $X$  on time to event. In contrast, because of the confounding problem just described, we may have  $\beta \neq 0$  in Model 2 even if there is no effect of  $X$  on time to event.

All  $P$ -values below are from score tests of  $\beta=0$  in the models; the logrank test for comparing Kaplan–Meier survival curves is a special case [16]. To better connect with g-estimation, we will present results in terms of estimates of the time ratio  $e^{-\beta}$ , which is also the ratio of the rates in the untreated versus treated.

### Censoring issues

With censoring a further assumption is needed, for example that censoring is independent of treatment assignment, or that censoring is independent of COPD time. While proper randomization makes treatment independent of censoring determined before treatment assignment (such as censoring by end of study), it cannot undo dependencies generated after assignment, as when treatment affects death rates.

A common strategy for dealing with this problem is to adjust for baseline covariates that predict both censoring and COPD, in the hope that censoring will be independent of COPD time within strata of these covariates. The extent of the problem and the effectiveness of adjustment depend on how much data are censored for the analysis, how much treatment affects that censoring, and how strongly that censoring predicts COPD after adjustment. In the present ITT analyses, only 4% of subjects in each arm are censored

(representing death or drop-out), limiting the problem and the impact of adjustment.

In the OT and PP analyses, however, censoring also arises from discontinuation, which is frequent and strongly affected by treatment, as can be seen in Table 1 (time lost from discontinuation is 5% in the tiotropium arm, as opposed to 11% in the placebo arm,  $P < 0.001$ ). Furthermore, discontinuation strongly predicts COPD. These strong relations should raise concerns about the validity of the OT and PP analyses [15].

### Intent-to-treat analyses

Conventional ITT analysis compares the COPD-time distributions in the assigned groups using conventional methods, ignoring noncompliance. The ITT analysis from the Cox model without covariate adjustment gives an estimated time ratio of 1.21, 95% confidence limits (CL) of 1.02 and 1.43,  $P = 0.029$ . Adjusting for the covariates listed in Table 2, the estimate remains 1.21 with CL of 1.02, 1.43,  $P = 0.027$ . Thus the ITT analysis suggests that assignment to tiotropium resulted in an increased time to COPD events relative to assignment to placebo, but that increase might be only a few percent and was not likely much more than 40%.

Because the ITT  $P$ -value requires the fewest assumptions for validity, it could reasonably be taken as a gold standard against which to judge other  $P$ -values. On the other hand, for estimating efficacy, we would expect the ITT estimate to suffer bias toward the null in proportion to the degree of noncompliance; in other words, the potential for benefit from treatment is understated [2,6–8,10] (e.g., the upper confidence limit of 43% benefit is too low). A good method would correct this bias without overstating the significance of the result, which in this example would imply that the upper limit should encompass larger effects, but the lower limit should change very little because it is close to the null.

### On-treatment analyses

We next repeat the above survival analysis, but censoring patients (terminating follow-up) when they go off their assigned treatment. This censoring results in discarding an average of 5% of follow-up time among the tiotropium patients and 11% of follow-up time among placebo patients, reflecting better compliance among tiotropium patients (presumably because of beneficial treatment effects). The differential censoring introduces a potential bias in the comparison, insofar as the extra

proportion of time retained in the tiotropium arm is unlikely to be random. The main change from the ITT analysis should be greater impact of covariate adjustment. Indeed, the OT Cox model without covariates gives an estimated time ratio of 1.22 with CL of 1.02 and 1.45,  $P = 0.031$ , close to the ITT results; in contrast, the covariate-adjusted estimate is 1.27 with CL of 1.06 and 1.52,  $P = 0.009$ , shifted upward from the ITT results. Nearly identical results were obtained for tiotropium versus placebo from an AT analysis with three groups: tiotropium, placebo, and no treatment (discontinued).

The changes from the ITT results are as expected. Nonetheless, the adjusted OT results rely heavily on the success of the covariate adjustment in removing bias due to the nonrandomized comparison. Given the relation of discontinuation to both treatment and outcome, the adequacy of this adjustment is uncertain if not doubtful. As the interval estimates and  $P$ -value do not account for this uncertainty, they must overstate precision (i.e., the intervals are too narrow to ensure at least 95% confidence), with too high a lower limit, and the  $P$ -value must overstate statistical significance.

### Per-protocol analyses

Per-protocol analysis discards entire records of patients that go off treatment. The PP Cox model without covariates gives an estimated rate ratio of 1.28 with CL of 1.08 and 1.54,  $P = 0.006$ , while the covariate-adjusted estimate becomes 1.36 with CL of 1.13 and 1.63,  $P = 0.001$ . There is a strong relation of discards to treatment (23% for tiotropium vs. 31% for placebo), and the higher retention of patients in the tiotropium arm is unlikely to be random. Thus the analysis is highly dependent on covariate adjustment to control bias, and is of even more doubtful validity than the OT analysis. In particular, because of uncertainty about whether the adjustment is adequate, the interval estimates overstate precision, with lower limits too high, and the  $P$ -values severely overstate significance.

### Structural nested models and g-estimation

Structural nested modeling accommodates discontinuation by direct modeling of COPD time as a function of time on treatment [2,10]. Suppose that each patient has a time (possibly unique to them)  $T_0$  at which they would have their first

COPD event after assignment, had they never used tiotropium, and another time  $T_1$  at which they would have their first COPD event after assignment had they always used tiotropium as directed before the event. This pair of times may be unique to each patient.  $T_0$  is not observed unless the patient never uses tiotropium before the COPD event, whereas  $T_1$  is not observed unless the patient always uses tiotropium before the event.

The simplest structural-nested model for the physiologic effect of tiotropium says that these times are proportional:  $T_1 = e^{-\beta}T_0$ , or equivalently  $T_0 = e^{\beta}T_1$ . We can rewrite this model as  $T_x = \exp(-\beta x)T_0$ , where  $x$  indicates usage rather than assignment, as in the PP analysis. This simple model is an accelerated-failure-time model [16] in which  $e^{\beta} = T_0/T_1$  is the ratio of the time to a COPD event when never treated ( $X=0$  always) versus when always treated ( $X=1$  always). If the event time with no tiotropium ( $T_0$ ) follow a log-linear model  $T_0 = \exp(-\alpha - \gamma z)$ , the structural-nested model becomes

$$T_x = \exp(-\beta x) \exp(-\alpha - \gamma z) = \exp(-\alpha - \beta x - \gamma z),$$

and the event rate while on treatment  $X=x$  becomes

$$\frac{1}{T_x} = \exp(\alpha + \beta x + \gamma z).$$

The latter formula is the constant-rate Cox model used in the conventional analyses.

To generalize this model to partial usage, suppose a patient spends time  $T_{\text{on}}$  on tiotropium and  $T_{\text{off}}$  off tiotropium before the first COPD event, which occurs at time  $T_{\text{tot}} = T_{\text{on}} + T_{\text{off}}$ . The structural-nested model then says that  $T_1 = T_{\text{on}} + e^{-\beta}T_{\text{off}}$ , or equivalently that  $T_0 = e^{\beta}T_1 = e^{\beta}T_{\text{on}} + T_{\text{off}}$ . Patients who never use tiotropium have  $T_{\text{on}} = 0$ ; for those patients,  $T_0 = T_{\text{off}} = T_{\text{tot}}$ , regardless of the tiotropium effect  $\beta$ . Similarly, patients who always use tiotropium have  $T_{\text{off}} = 0$ ; for those patients,  $T_1 = T_{\text{on}} = T_{\text{tot}}$ .

Several mechanisms lead to this model. As an example, suppose a tissue or organ system (e.g., the lungs) declines in functionality at a constant rate if no tiotropium is used, leading to an event at  $T_0$  when functionality declines below a certain critical level. The model would then arise if, during the times tiotropium is used, the decline rate is multiplied by  $e^{\beta}$ ; note that  $\beta < 0$  if treatment slows decline. The same model would arise if the organ cells (e.g., the cells of the substantia nigra) die at a constant rate, which would lead to an event (e.g., Parkinsonian tremors) at  $T_0$  without intervention, and the active treatment multiplied this

rate by  $e^{\beta}$ . The decline rate and threshold for the outcome event may vary across patients; only the treatment effect on the decline rate  $e^{\beta}$  is assumed constant. The model is thus analogous to the Cox model, which assumes the treatment effect on the hazard is a constant  $e^{\beta}$ . The two model effects will be equal when the underlying event (hazard) rate is constant (as assumed here), and are proportional under a Weibull survival models [17]. (It is possible to formulate structural Cox models directly [18,19] but estimation is not as simple.)

### G-testing and g-estimation

Fitting of Cox models requires a special procedure (partial-likelihood maximization) that is largely hidden from users within packaged software. Similarly, fitting of structural nested models requires a special procedure called *g-estimation*. It is based on *g-testing*, in which the hypothesis that  $\beta = b$  in a structural-nested model is tested by comparing those assigned to tiotropium ( $R=1$ ) with those assigned to placebo ( $R=0$ ), thus adhering to the ITT principle.

Suppose treatment assignment is randomized, with the assignment (ITT) indicator being  $R=1$  if assigned to tiotropium, 0 otherwise, as before. The logic of *g-estimation* is then as follows:

- 1) The time of the event if the patient never uses tiotropium,  $T_0$ , exists and is defined regardless of and unaffected by whatever treatment the patient is assigned or receives.
- 2) Therefore, assignment  $R$  should have only a random association with  $T_0$ .
- 3) If  $b$  were the true value of  $\beta$ , for each patient the true value of  $T_0$  would be  $T_0(b) = e^{bT_{\text{on}}} + T_{\text{off}}$ .
- 4) Therefore, if  $b$  were the true value of  $\beta$ ,  $T_0(b)$  should have only a random association with assignment  $R$ .
- 5) It follows that the  $P$ -value for the association of assignment  $R$  with  $T_0(b)$  is a valid, randomization-based (ITT) test of the hypothesis that  $\beta = b$  (called the *g-test* of  $\beta = b$ ). It can also be viewed as a test of the hypothesis that the ratio  $e^{-\beta} = T_1/T_0$  (the event time when always treated divided by the event time when never treated) is equal to  $e^{-b}$ .
- 6) The value  $e^{-b}$  for  $e^{-\beta}$  that has a two-sided  $P$  closest to 1 for its *g-test* is a valid estimate of  $e^{-\beta}$  (called the *g-estimate* of  $T_1/T_0$ ). Similarly, the values  $e^{-b}$  for  $e^{-\beta}$  that give a two-sided  $P$  closest to 0.05 are valid approximate 95% confidence limits for  $e^{-\beta}$ .

Following on this final step, a simple way to find the g-estimate and confidence limits for the time ratio  $T_1/T_0 = e^{-\beta}$  is to test the association of  $T_0(b)$  with treatment assignment  $R$  for many  $b$ , and tabulate the results. We then take the  $e^{-b}$  with  $P$ -values nearest 1 and 0.05 as our point estimate and confidence limits for  $T_1/T_0$ . This repeated-testing process is basic g-estimation.

Typically, the  $P$ -value will jump suddenly as  $e^{-b}$  is varied, and there may be no value for  $e^{-b}$  with  $P$  exactly equal to 1 or 0.05. Hence one should use a very fine search grid, in order to locate the jump points nearest 1 and 0.05. The Appendix provides further details on how censored observations can be incorporated into the process.

The null hypothesis of no treatment effect is  $\beta = 0$ . Under that hypothesis we test the association of assignment  $R$  with  $T_0(0) = \exp(0)T_{on} + T_{off} = T_{on} + T_{off} = T_{tot}$ , the actual time of the outcome event. In other words, the g-test of the null examines the association of treatment assignment with the observed times of the outcome event. Thus the g-test of the null (no effect) is just a standard ITT test of the effect of treatment comparing the  $R=1$  and  $R=0$  groups (usually done with a logrank test). This equivalence shows how g-estimation generalizes ITT analysis from null testing to testing other hypotheses, and to estimation. In doing so, g-estimation maintains the original randomized-group definitions (and thus helps preserve test validity), which OT, AT, and PP analyses do not do.

Because g-estimation is based on the original randomized treatment assignment (ITT) and there is little censoring in the example, we should expect covariate adjustment to have only slight impact. Table 3 summarizes the g-testing results (the final column is described in the Appendix). From the table, we can see that both the unadjusted and adjusted estimates of the time ratio  $e^{-\beta}$  (the ratio with  $P$  nearest 1) is 1.31, with 95% limits (the ratios with  $P$  nearest 0.05) of 1.03 and 1.72, with null  $P$ -values of 0.029 and 0.027, identical to the ITT values. G-estimation has thus met our key goals:

- 1) By avoiding nonrandomized comparisons that distort the tests and limits from OT, AT, and PP analyses, it agrees with the ITT test of the null and lower confidence limit, indicating that the time increase due to tiotropium may be just a few percent (null  $P=0.03$ ), as opposed to the lower limit of 13% from PP (null  $P=0.001$ ).
- 2) At the other end, it has corrected the null bias of the conventional ITT estimates, suggesting that the time increase achieved by treatment may exceed 70% (as opposed to the ITT upper limit of 43%).

**Table 3** Results from g-testing different values for the time ratio  $T_1/T_0 = e^{-\beta}$  where  $T_1$  is the time the first COPD event would have occurred if tiotropium had always been used and  $T_0$  is the time the first COPD event would have occurred if tiotropium had never been used, assuming with constant covariate-specific event rates; under this model,  $e^{-\beta}$  is also the ratio of placebo and tiotropium event rates

Test hypothesis $T_1/T_0 = e^{-\beta} =$	Unadjusted $P$ -value	Adjusted $P$ -value <sup>a</sup>	Adjusted $P$ -value from weighted analysis <sup>b</sup>
1 (no effect)	0.029	0.027	0.030
1.02	0.039	0.036	0.040
1.03	0.055 <sup>c</sup>	0.053 <sup>c</sup>	0.058 <sup>c</sup>
1.30	0.95	0.96	0.92
1.31	0.99 <sup>c</sup>	0.99 <sup>c</sup>	0.97 <sup>c</sup>
1.32	0.93	0.92	0.96
1.71	0.075	0.067	0.089
1.72	0.043 <sup>c</sup>	0.039 <sup>c</sup>	0.055
1.74	0.037	0.034	0.048 <sup>c</sup>

<sup>a</sup>Baseline adjustment covariates are as in footnote of Table 2.  
<sup>b</sup>From inverse-probability-of-censoring weighted analysis [25].  
<sup>c</sup>G-estimate and 95% limits (effects in column with  $P$ -values nearest 1 and 0.05, to three-digit accuracy).

These accomplishments may be summarized by saying that g-estimation has properly accounted for uncertainty at both ends of the spectrum, providing an appropriately wide confidence interval that neither understates nor overstates the likely treatment benefit.

## Discussion

Compliance was high in our example: only 8% of trial time was off treatment. Thus, one might think that noncompliance would be of little importance. But different methods for dealing with noncompliance gave very different answers. Relative to ITT, the PP analysis grossly overstated the significance of the results. Furthermore, relative to g-estimation, all the conventional intervals (including ITT) are overly precise, insofar as they appear to rule out treatment benefits that are quite compatible with the data at the 95% confidence level.

For simplicity, our structural model assumed there was no placebo effect, which is not always realistic. There are several ways to allow for such effects, for example by adding a model parameter for the effect. No single analysis is ideal, however, and our main point is that multiple analyses of the same data can be of benefit in interpreting study results, provided the limitations of each analysis are recognized. ITT provides a 'gold standard' for testing the null, which addresses the question

'can we reliably conclude there is an effect of actual treatment?' Nonetheless, conventional ITT can be deceptive for estimating treatment effects, and is particularly biased for answering the question of 'how large might the effect be?' Such questions require more thoughtful use of the ITT indicator than just pretending it represents actual treatment.

Different analyses are based on different assumptions, models, and patient exclusions, and may estimate different effects. A spectrum of results thus displays how sensitive conclusions are to these choices. Differing results call for explanation in terms of possible assumption failures. Nonetheless, PP analysis is extremely sensitive to apparently small sources of bias. Relative to PP analysis, OT analysis is just as easy to conduct, requires no stronger assumption, preserves more information, and stays closer to the original randomization. Thus the rationale for bothering with PP analysis is obscure. The rationale is all the more doubtful if one employs g-estimation, which preserves the randomization and gives back the ITT null *P*-value, without the bias of the ITT estimate and confidence interval.

Like ITT analysis, g-estimation primarily requires that an ITT *design* should be employed, in which follow-up continues regardless of treatment compliance. Computation is not difficult; several articles have presented SAS and Stata code for g-estimation, [13,20–22] and our appendix gives an algorithm for g-testing that can be easily coded into any package. These algorithms cover only the basic structural nested model, however. As with conventional methods, there are many generalizations that include time-varying treatment effects, heterogeneous treatment effects ('effect modification'), random effects, multiple treatment arms, continuous treatments (dose-dependent effects), and repeated outcomes, as well as improved methods for separating direct from indirect effects and for controlling time-varying confounders [11,17,23,24].

## Acknowledgments

The authors thank Babette Brumback, PhD, and Marshall Joffe, MD, PhD, and the referees for helpful comments.

## References

1. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Stat Med* 2003; 22: 169–86.
2. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank-preserving structural failure-time models. *Commun Stat* 1991; 20: 2609–31.
3. Scheiner LB, Rubin DB. Intent-to-treat analysis and the goals of clinical trials. *Clin Pharm Therap* 1995; 57: 6–15.
4. Pearl J. *Causality: Models, Reasoning, and Inference* Ch. 8, Cambridge University Press, Cambridge, New York, 2000.
5. Dunn G, Goetgebheur E. Analysing compliance in clinical trials. *Stat Methods Med Res* 2005; 14: 325–6.
6. Sommer A, Zeger S. On estimating efficacy from clinical trials. *Stat Med* 1991; 10: 47–52.
7. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Stat Assoc* 1996; 91: 444–55.
8. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000; 29: 722–9.
9. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Stat Methods Med Res* 2005; 14: 369–95.
10. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Control Clin Trials* 1993; 14: 79–97.
11. Robins JM. Correcting for noncompliance in randomized trials using structural nested mean models. *Commun Stat Theory Meth* 1994; 23: 2379–412.
12. Korhonen PA, Laird NM, Palmgren J. Correcting for non-compliance in randomized trials: an application to the ATBC study. *Stat Med* 1999; 18: 2879–97.
13. Cole SR, Chu H. Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemp Clin Trials* 2005; 26: 300–10.
14. Niewoehner DE, Rice K, Cote C, et al. Prevention of exacerbations of chronic obstructive pulmonary disease with tiotropium, a once-daily inhaled anticholinergic bronchodilator. *Ann Intern Med* 2005; 143: 317–26.
15. Kesten S, Plautz M, Piquette C, et al. Premature discontinuation of patients most afflicted from the disease under study: a potential bias in COPD clinical trials. *European Respiratory Journal* 2007; 30: 898–906.
16. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data* (2nd ed.) Wiley, New York, 2002.
17. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; 3: 319–36 [errata: *Epidemiology* 1993; 4: 189].
18. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; 11: 561–70.
19. Loeys T, Goetghebeur E, Vandebosch A. Causal proportional hazards models and time-constant exposure in randomized clinical trials. *Lifetime Data Analysis* 2005; 11: 435–49.
20. Witteman JCM, D'Agostino RB Sr, Stijnen T, et al. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Study. *Am J Epidemiol* 1998; 148: 390–401.
21. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2002; 2: 164–82.
22. White IR, Walker S, Babiker A. Strbee: Randomization-based efficacy estimator. *The Stata Journal* 2002; 2: 140–50.
23. Robins JM, Greenland S. Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *J Am Stat Assoc* 1994; 89: 737–49.

24. Hernán MA, Cole SR, Margolick J, *et al.* Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology Drug Safety* 2005; **14**: 477–91.
25. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**: 106–21.
26. Robins JM, Finkelstein D. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**: 779–88.
27. Joffe MM. Administrative and artificial censoring in censored regression models. *Stat Med* 2001; **20**: 2287–304.

## Appendix: g-testing with standard software

Suppose that for each patient there is a known time  $C$ , unaffected by treatment, and at this time the patient will be censored if the event under study has not yet occurred ( $C$  is typically time from randomization until the end of the study). Let  $Y=1$  if the patient was observed to have the outcome event (not censored),  $Y=0$  if the patient is censored. Recall that  $T_{\text{on}}$  and  $T_{\text{off}}$  are the times that the patient spends on and off active treatment; if the patient is not censored (i.e., if  $Y=1$ ) they are both observed, and  $T_0(b) = e^b T_{\text{on}} + T_{\text{off}}$  can be computed.

Suppose reaching time  $C$  without the event is the only source of censoring, so that  $Y$  indicates whether a patient had an event before  $C$ . Then a G-test of the hypothesis  $\beta=b$

(equivalently, the test of  $T_1/T_0 = e^{-b}$ ) in a structural-nested model can be conducted with standard survival-analysis software by creating a new censored-survival time  $U_0(b)$  and outcome indicator  $Y_0(b)$  for each patient. These new variables are defined from the observed data and  $e^b$  as follows:

$$\begin{aligned} &\text{If } Y = 1; \\ &\quad U_0(b) = \text{minimum of } T_0(b), e^b C, \text{ and } C; \\ &\quad Y_0(b) = 1 \text{ if } U_0(b) = T_0(b); \\ &\quad Y_0(b) = 0 \text{ otherwise;} \\ &\text{If } Y = 0; \\ &\quad U_0(b) = \text{minimum of } e^b C \text{ and } C; \\ &\quad Y_0(b) = 0. \end{aligned}$$

The  $P$ -value for g-testing  $\beta=b$  is then just the  $P$ -value for the association of the assignment (ITT) indicator  $R$  with the new survival time  $U_0(b)$  and outcome indicator  $Y_0(b)$ .

The  $P$ -values in the text and the first two columns of Table 3 were generated from the above algorithm taking  $C$  as the minimum of end of study, death time, or drop-out time. They thus assumed that death and drop-out (as well as end of study) were unaffected by treatment even if treatment affected COPD time. This assumption is not realistic, and could lead to bias if death and drop out are common and affected by treatment. For less restrictive methods see [24–27]. The final column of Table 3 (weighted analysis) applies these methods to the example; they here have negligible impact because only 70 patients (4%) were censored by death or drop-out, and the latter events showed little association with treatment (Table 1).