1. INTRODUCTION

In a typical clinical trial, patients are randomized to one of the treatment groups and each patient is expected to receive that treatment throughout the follow-up to assess the effect of the treatment on some outcome. However, most clinical trials are not ideal; hence, patients often fail to adhere to their assigned treatment and switch to another trial treatment. Such non-compliance with assigned treatment is a common feature of clinical trials. Recently there has been much interest in methods for analyzing randomized clinical trials of treatments to which the subject are not compliant [1–3].

One approach for analyzing data for non-compliance is the as-treated (AT) analysis, which compares outcomes based on the treatment that patients actually received. When non-compliance is completely at random, that is, independent of both (observed and unobserved) baseline and time-dependent factors, the AT analysis can give a valid test for the null hypothesis of no treatment effect and can also give an unbiased estimate of treatment effect. In most clinical trials, however, patients who comply with their assigned treatment are not comparable with those who do not with respect to some important prognostic factors. In this case, both the decision to comply and the outcome may well depend on underlying possibly unmeasured health status. Thus, when non-compliance is non-random, the AT analysis will not be valid even under the null hypothesis because of the comparison of selected groups [3, 4].

The more commonly used analytic approach is an intention-to-treat (ITT) analysis, which compares outcomes based on the treatment groups randomized by design regardless of whether the patients complied with their assigned treatment. Because the comparability of the treatment groups is guaranteed by randomization, the null hypothesis of no treatment effect for all patients (sharp causal null hypothesis) is preserved in the ITT analysis. That is, successful randomization insures that the ITT comparison provides a valid test for the sharp causal null hypothesis of no treatment effect even in the presence of non-random non-compliance. Moreover, p-values have a randomization interpretation when design-based (randomization-based) analyses are used [5]. Furthermore, the ITT estimate would correspond to the overall treatment effect that would be realized if the treatment were actually adopted and practiced in the community, provided the rate of non-compliance and the factors influencing non-compliance that are observed in the trial are identical to those that would occur in the community. A point against the ITT analysis is that the ITT parameter does not measure the true biological effect of treatment, but rather a mixture of the effect on the compliers with the absence of effect on the non-compliers, because the ITT estimate is the average effect of treatment assignment. Hence, the ITT analysis gives estimates that are biased toward the null when treatment crossover is present, and the ITT measure of treatment effect will diminish as non-compliance increases. Moreover, the rate of non-compliance in the community, once the treatment is adopted, may not be the same as the rate in the original clinical trial.

Therefore, in the analysis of non-compliance data, it is important to estimate the causal effect of treatment, that is, the effect that would be realized if all patients complied with the treatment to which they were assigned. Robins [6–8] has proposed a structural nested mean model (SNMM) to estimate such causal effect in the presence of non-random non-compliance. Under the assumption that non-compliance at each time is at random, given the observed histories that influence a patient's decision to comply, that is, the assumption of no unmeasured confounders, the causal parameter in a SNMM can be estimated by the technique of g-estimation.

Recently, Brumback *et al.* [9] proposed the intensity score approach for the analysis of time-varying treatments in the presence of time-dependent confounding. They provided conditions

DOI: 10.1002/sim

under which the intensity score approach consistently estimates a treatment effect in a SNMM. The intensity score is cumulative differences over time between treatment actually received and treatment predicted by prior observed medical history. The SNMM treatment effect can be obtained by regressing outcomes on the intensity score. Thus, the intensity score approach can provide an easy implementation of g-estimation for the analysis of non-random non-compliance. Since the intensity score approach was originally proposed for continuous outcomes, we extend its use to time-to-event outcomes with censoring. This extension is useful, because censoring due to end of scheduled follow-up requires special care when using g-estimation based on the structural accelerated failure time (SAFT) model [10–16], while the intensity score approach can treat the censoring within the framework of standard regression models. Furthermore, the intensity score approach has the advantage of providing estimates of parameters in a SNMM that allows the treatment effects to vary across time, while it has been difficult to apply such a model in practice using the technique of g-estimation [9].

This article is organized as follows. In Section 2, we describe the motivating study from a large randomized primary prevention study for coronary events, the Management of Elevated Cholesterol in the Primary Prevention Group of Adult Japanese (MEGA) study [17, 18]. In Section 3, we develop the intensity score approach for event times. In Section 4, simulation studies are conducted to compare the performances of the proposed intensity score method with those of the AT, ITT, and g-estimation (semi-parametric randomization-based) analysis [10, 12]. Section 5 presents the analysis results of the MEGA study data. Finally, Section 6 provides some discussions.

2. THE MEGA STUDY

We will briefly describe the motivating study and the data (MEGA study). Full details on the design, conduct, and main clinical results have been reported [17, 18]. The MEGA study is a randomized controlled trial conducted in Japan to evaluate the primary preventive effect of a statin against coronary heart disease (CHD) in daily clinical practice. In this prospective, randomized, open-labeled, blinded-endpoints design study, men and postmenopausal women aged 40–70 years with hypercholesterolemia (total cholesterol (TC) level: 220–270 (mg/dL)) and no history of CHD or stroke were randomized to diet (diet group) or diet plus pravastatin 10–20 mg daily (pravastatin group).

Between February 1994 and March 1999, a total of 15 210 persons visiting outpatient clinics were registered throughout Japan. Of the 15 210 subjects who met the inclusion criteria regardless of their TC levels and who provided signed informed consent, 8214 who met the TC criterion were randomized to either diet or diet plus pravastatin treatment using the permuted block method with stratification according to gender, age, and medical institution. After the exclusion of 382 patients (94 withdrew consent, 224 exclusion criteria violation, and 64 no recorded data after randomization), the remaining 7832 patients were analyzed (3966 diet group; 3866 pravastatin group).

Table I shows the baseline characteristics of the analyzed patients. There was no clinical difference between the two groups in baseline characteristics. Women accounted for 68.4 per cent (5356 patients) of the study population. Mean body mass index (BMI) was 23.8 (kg/m^2). Mean TC, low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C) levels were 242.6, 156.6, and 57.5 (mg/dL), respectively. Median triglyceride (TG) level was 127.5 (mg/dL). Of the study patients, 41.8 and 20.8 per cent had hypertension and diabetes mellitus based on physician diagnosis, respectively.

DOI: 10.1002/sim

Table I. Baseline characteristics of analyzed 7832 patients.

Characteristics	Diet group ($N=3966$)		Diet+pravastatin group ($N=3866$)	
Age (years), mean (SD)	58.4	(7.2)	58.2	(7.3)
Women, no. (per cent)	2718	(68.5)	2638	(68.2)
BMI (kg/m^2), mean (SD)	23.8	(3.0)	23.8	(3.1)
Current smoker, no. (per cent)	572	(14.4)	612	(15.8)
Current drinking, no. (per cent)	1183	(29.8)	1180	(30.5)
Hypercholesterolemia medication history, no. (per cent)	621	(15.7)	586	(15.2)
Hypertension, no. (per cent)	1664	(42.0)	1613	(41.7)
Diabetes, no. (per cent)	828	(20.9)	804	(20.8)
TC (mg/dL), mean (SD)	242.6	(12.1)	242.6	(12.0)
TG (mg/dL), median (inter-quartile range)	127.5	(95.0–179.0)	127.4	(95.7–176.5)
HDL-C (mg/dL), mean (SD)	57.5	(15.1)	57.5	(14.8)
LDL-C (mg/dL), mean (SD)	156.5	(17.3)	156.7	(17.6)

SD, standard deviation; BMI, body mass index; TC, total cholesterol; TG, triglyceride; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol.

After randomization, patients were followed at months 1, 3, and 6 and thereafter every 6 months. At each visit, data on treatment compliance, use of concomitant drugs, onset of events, occurrence of adverse events, and laboratory tests including serum lipids were collected by the investigators. Additionally, an ECG (electrocardiogram) was obtained and evaluated annually. The follow-up period was initially scheduled for 5 years; however, on the basis of the recommendation of the Data and Safety Monitoring Committee, the study was continued for an additional 5 years to increase the number of events, and thus, patients who provided written consent at 5 years to continue the study were followed until the end of March 2004 [17, 18].

The primary endpoint was the first occurrence of CHD, comprised of fatal and non-fatal myocardial infarction, angina, cardiac and sudden death, and a coronary revascularization procedure. One of the secondary endpoints was the first occurrence of stroke events. All endpoints were reviewed strictly by the blinded Endpoint Committee and additional information obtained from the physician as needed [17]. A total of 7832 patients were followed by 2658 physicians in 1320 hospitals. The follow-up period was 41 195 person-years (mean follow-up period 5.3 years). CHD events occurred in 101 of 3966 patients in the diet group (2.55 per cent) and 66 of 3866 patients in the pravastatin group (1.71 per cent). Figure 1 shows the Kaplan–Meier curves for CHD events. The ITT analysis indicated that the incidence of CHD was significantly lower by 33 per cent in the pravastatin group than in the diet group (The ITT hazard ratio $=0.67$; 95 per cent confidence interval (CI): 0.49–0.91; $p=0.01$ for the log-rank test) [18].

However, many patients changed to the other trial treatment frequently during the study period (treatment crossover). This was because the protocol in the MEGA study stated that patients in the diet group could be switched to pravastatin treatment when a reduction of TC level was not observed, while patients in the pravastatin group could discontinue pravastatin treatment when the reduction of TC level was observed. The treatment decisions for changing the treatment or increasing the dose of paravastatin were determined by each treating physician. Patients who changed to another trial treatment even once in the first 5 years were 19.9 per cent ($n=790$) in the diet group and 53.4 per cent ($n=2064$) in the pravastatin group. These numbers for the whole

DOI: 10.1002/sim

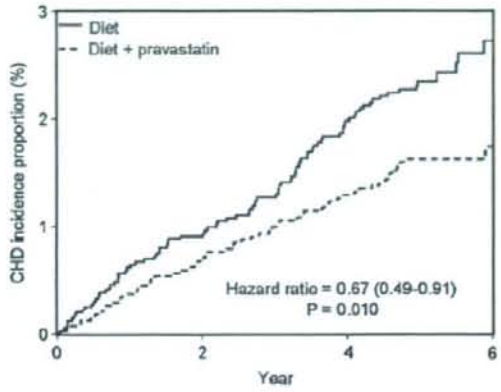

Figure 1. Incidence proportion for CHD events.

10 years were 21.3 per cent ($n=844$) and 63.1 per cent ($n=2441$), respectively. The effect of patients from one treatment to the other is to make the treatment profiles of the two randomized groups more similar than they otherwise would have been, and therefore to move the ITT hazard ratio toward the null.

3. INTENSITY SCORE METHOD

3.1. The multiplicative structural nested mean model

We consider a randomized clinical trial in which two groups (test and control treatment) are compared with respect to time-to-event outcomes and each patient i ($i=1,\ldots,N$) receives one of the treatments at the start of each time t ($t=0,\ldots,M-1$; time zero is the randomization time and the start of the first treatment). However, some patients fail to comply with their assigned treatment and cross over to the other treatment at each time t.

Suppose we have repeated measures on treatment $S_i(t)$ ($S_i(t)=1$ if test treatment, $S_i(t)=0$ if control treatment) and covariates $L_i(t)$ at time t. Let $H_i(t)$ be the observed history of treatment and the covariates prior to treatment at time t, i.e. $H_i(t)=(L_i(0), S_i(0), \ldots, L_i(t-1), S_i(t-1), L_i(t))$, with $H_i(0)=(L_i(0))$. Let $T_i(\overline{S}_i(t), 0)$ denote the potential event times in response to the hypothetical treatments $(S_i(0),\ldots,S_i(t), S_i(t+1)=0,\ldots,S_i(M-1)=0)$. That is, $T_i(\overline{S}_i(t), 0)$ represents the event time we would have observed if, possibly contrary to fact, the patient had his/her actual treatment history up to time t but was then switched to control treatment at time $t+1$ and remained at that treatment until the event occurred. Our notation for the potential outcomes implicitly assumes Rubin's stable unit treatment value assumption, which implies that potential outcomes of patient i do not depend on the treatment received by any other patient [19]. We will also assume that the potential outcomes satisfy the consistency assumption [7] that serves to link the potential outcomes with the observed outcomes T_i. This assumption states that $T_i=T_i(\overline{S}_i(t), 0)$ for all t when actually $S_i(t+1)=\cdots=S_i(M-1)=0$ occurred.

DOI: 10.1002/sim

We introduce a simple multiplicative SNMM [6–8]

$$\log E[T_i(\overline{S}_i(t),0)|H_i(t),S_i(t)]-\log E[T_i(\overline{S}_i(t-1),0)|H_i(t),S_i(t)]=\beta_0 S_i(t) \quad (1)$$

where β_0 is the constant (across t) incremental causal effect of a final treatment $S_i(t)$ at time t on the potential outcome $T_i(\overline{S}_i(t-1),0)$ following a patient's actual treatment through times $0,\ldots,t-1$ and control treatment after $t-1$. Under this constant treatment effect model (1), β_0 multiplied by M (number of visit time) can be interpreted as the average causal treatment effect that would be realized if all patients had continued to comply with the treatment to which they were assigned. Robins [6–8] proposed the estimation method of β_0, the so-called, g-estimation method, under the assumption of sequential conditional independence for any t and k with $k\leqslant t-1$

$$T_i(\overline{S}_i(k),0)\perp\!\!\!\perp S_i(t)|H_i(t) \quad (2)$$

which states that, when $k\leqslant t-1$, treatment $S_i(t)$ is independent of the potential outcomes $T_i(\overline{S}_i(k),0)$, given the observed history up to time t, $H_i(t)$. In practice, we would not expect this assumption to be precisely true, but given a rich collection of prognostic factors that influence a patient's decision to comply at time t recorded in $H_i(t)$, it may well be approximately true. Robins [6–8] has referred to (2) as the assumption of no unmeasured confounders.

3.2. *Estimation of β_0 via the intensity score method*

Brumback *et al.* [9] proved that the SNMM treatment effect, that is, g-estimator of β_0, can be obtained by the intensity score method, in which outcomes are regressed on the cumulative intensity score. We utilize their results and consider the accelerated failure time (AFT) model to obtain the consistent estimator of β_0 in the multiplicative SNMM (1). Here, we assume that the observed event time T_i is subject to independent random censoring such as an end-of-study censoring, where T_i for censored subjects is either the time until dropout or the time until end of study.

We assume the following exponential regression model (log-linear model) for T_i [20]:

$$\log T_i=\mu+\beta_I\sum_{t=0}^{M-1}\hat{I}_i(t)+\varepsilon_i \quad (3)$$

where μ is the intercept parameter, $I_i(t)=S_i(t)-E[S_i(t)|H_i(t)]$ is the intensity score at time t, and ε_i follows the extreme value distribution. For binary treatment $S_i(t)$, the time-dependent intensity score $I_i(t)$ measures departures of actual treatment from the propensity score $\Pr[S_i(t)|H_i(t)]$ of Rosenbaum and Rubin [21]. Since the propensity score is usually unknown, it must be estimated from the data. If we assume a parametric model for $\Pr[S_i(t)|H_i(t)]$ such as

$$\text{logit}\Pr[S_i(t)=1|H_i(t)]=\theta^{\mathrm{T}}H_i(t) \quad (4)$$

then the intensity score at time t can be estimated by $\hat{I}_i(t)=S_i(t)-E[S_i(t)|H_i(t);\hat{\theta}]$, where $\hat{\theta}$ is the maximum likelihood estimator of θ under model (4). Here we assume that the intensity score at time t is not equal to zero with probability 1 for each patient, that is, $\hat{I}_i(t)\neq 0$ for any t. This assumption will be satisfied unless there is a covariate level $H_i(t)$ such that all patients with that level of the covariate are certain to receive the treatment.

The estimate for β_I in model (3) can be obtained *via* the ordinary-weighted least-squares (WLS) method. However, the cumulative intensity score is generally uncorrelated with the cumulative propensity score, although $E[I_i(t)E(S_i(t)|H_i(t))]=0$ for any t. Therefore, as Brumback *et al.* [9]

DOI: 10.1002/sim

have pointed out in the case of linear model, to obtain a consistent estimator of β_0 *via* the WLS method for model (3), the correction term $N\beta_I C$ must be subtracted from the WLS estimating function, where $C=(1/N)(\sum_{t=0}^{M-1}\hat{I}_i(t))\omega_i(\sum_{t=0}^{M-1}E(S_i(t)|H_i(t);\hat{\theta}))$ and $\omega_i=\exp(-\mu)\cdot T_i\cdot\exp(-\beta_I\sum\hat{I}_i(t))$. The corrected estimating function for model (3) is

$$U(\mu,\beta_I)\equiv\sum_i^N(-d_i\sum\hat{I}_i(t)+\sum\hat{I}_i(t)\exp[\log t_i-\mu-\beta_I\sum\hat{I}_i(t)])-N\beta_I C=0 \tag{5}$$

where d_i is the event indicator that takes the value of one if the subject failed and zero if the subject is censored. In Appendix A, we show the proof that the correction term must be subtracted from the WLS estimating function to obtain a consistent estimator.

Our estimating function has the form $\sum_i^N U_i(\gamma)=0$, where $\gamma=(\mu,\beta_I,\theta)$ represents the intercept, coefficient of the intensity score, and parameter used to model the propensity score. To correct for having the estimated θ, the asymptotic variance of $\hat{\gamma}$ was obtained by using a sandwich estimator, which was computed as

$$[\hat{E}(\partial U_i/\partial\gamma)]^{-1}[\hat{E}(U_iU_i^{\mathrm{T}})\hat{E}(\partial U_i/\partial\gamma)]^{-1,\mathrm{T}}/N$$

where the estimated expectations were computed using the empirical distribution of the sample.

3.3. Time-dependent treatment effects

An extension of the multiplicative SNMM (1) is to allow the treatment effects to vary across time,

$$\log E[T_i(\overline{S}_i(t),0)|H_i(t),S_i(t)]-\log E[T_i(\overline{S}_i(t-1),0)|H_i(t),S_i(t)]=\beta_0(t)S_i(t) \tag{6}$$

where $\beta_0(t)$ is the causal parameter at each time t. Since $\beta_0(t)$ is the incremental causal effect of a final treatment $S_i(t)$ at time t, the cumulative effect $\sum_{k=0}^{t}\beta_0(k)$ is the average causal treatment effect that would be realized if all patients had continued to comply with the treatment to which they were assigned until time t. Assuming the consistency assumption and the sequential conditional independence (2), the time-dependent causal parameters in model (6) are consistently estimated by fitting the following model [9]

$$\log T_i=\mu+\sum_{t=0}^{M-1}\beta_I(t)\hat{I}_i(t)+\varepsilon_i \tag{7}$$

where the correction term $\sum_i[\hat{I}_i(t)\cdot\omega_i\cdot\sum_{t=0}^{M-1}\{\beta_I(t)E(S_i(t)|H_i(t);\hat{\theta})\}]$ must be subtracted from the WLS estimating function for model (7).

4. SIMULATION STUDY

4.1. Simulation design

To evaluate the performance of the AT, ITT, g-estimation (see Section 4.2) and intensity score methods, we carried out simulation studies under non-random non-compliance. We simulated data from two treatment groups, coded as $R=0$ (control treatment) or $R=1$ (test treatment). About equal sample size of 1000 for each group was randomly generated (total sample size was 2000).

DOI: 10.1002/sim

The simulations were based on 1000 replications so that the estimated coverage probability of a true 95 per cent CI would have a simulation accuracy of approximately 1.35 per cent.

For each subject i ($i=1,\ldots,2000$), a baseline covariate L_i was generated from the normal distribution with mean of 2 and variance of 1. The potential baseline failure time U_i was generated from the following exponential model:

$$U_i = U_0 \exp(\alpha_0 + \alpha_1 L_i) \tag{8}$$

where U_0 was an exponential random number with mean of 1 and $(\alpha_0, \alpha_1) = (3.2, -0.5)$ so that the larger the value of L_i, the shorter the baseline failure time U_i. We evaluated the treatment actually received $S_i(t)$ at three time points $t=0$, 2, and 4, where all subjects were assumed to take the assigned treatment at $t=0$ ($S_i(0)=R_i$) and the treatment crossover occurred at $t=2$ and 4 according to the following model:

$$\text{logit Pr}[S_i(t)] = \gamma_0 + \gamma_1 L_i + \gamma_2 S_i(t-2) \tag{9}$$

where $(\gamma_1, \gamma_2) = (1.2, 4.5)$ so that patients with poor prognosis and taking the test treatment at previous time point tended to receive the test treatment. The non-compliance rate was adjusted by the value of the intercept parameter γ_0, where two settings were considered: 45 per cent ($R=0$) *versus* 15 per cent ($R=1$) and 30 per cent ($R=0$) *versus* 10 per cent ($R=1$). In this non-compliance rate, the subject was considered as a non-complier when the subject received another treatment even once during the study period.

The observed failure time T_i was calculated from the SAFT model

$$U_i = \int_0^{T_i} \exp[-\psi_0 S_i(t)]\,dt \tag{10}$$

where ψ_0 is the causal treatment effect, which was set to $\psi_0 = 0.5$. The observed failure time T_i was censored at the fixed censoring time C, where $C=5$, 70, and ∞, so that the overall censoring proportion was nearly 90, 30, and 0 per cent, respectively.

Simulations were evaluated in terms of the bias, mean-squared error (MSE), mean length of the 95 per cent CI (length), 95 per cent coverage probability (CP), power for rejecting the null hypothesis, and α-error.

4.2. *g-estimation (semi-parametric randomization-based analysis)*

A semi-parametric randomization-based approach to estimate the causal effect has been developed by Robins and coworkers [10, 12]. For time-to-event outcomes, their approach is based on the causal AFT model (10), which relates a patient's observed event time T_i to the potential baseline event time U_i, that would have been observed if no treatment had been given, and the treatment actually received $S_i(t)$ *via* a causal parameter ψ_0. Note that if $S_i(t) \equiv 0$, then equation (10) gives $T_i = U_i$ as expected, while if $S_i(t) \equiv 1$, (10) gives $T_i = U_i \exp(\psi_0)$. Therefore, equation (10) implies that for continuous treatment the potential event time U_i is prolonged by the factor $\exp(\psi_0)$. A positive value of ψ_0 represents a beneficial treatment effect.

To estimate the causal parameter ψ_0, they choose to avoid all assumptions about both observed and unobserved factors that influence an individual's decision to comply such as (2), while comparing outcomes based only on the treatment groups randomized by design, that is, their analyses are randomization-based analysis. The key to understanding their estimation method

DOI: 10.1002/sim

(g-estimation) is to realize that U_i is a baseline variable identically distributed across the randomized groups. We define $U_i(\psi)$ to equal the right-hand side of (10) for given ψ. We also define $Z(\psi)$ to be a test statistic comparing the distribution of $U(\psi)$ in the two randomized groups, where we will use the log-rank test. The point estimate of ψ_0 is the value for which $Z(\psi)=0$, and this can be found by a search over a grid. A $100(1-\alpha)$ per cent confidence interval for ψ_0 is the range of values for which $|Z(\psi)|<z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. One attractive point of this approach is that at the null value, it is non-parametric, because $U_i(0)=T_i$; hence, $Z(0)$ is the usual ITT log-rank test statistic.

However, if T_i is a censored time, then $U_i(\psi)$ is censored at

$$D_i(\psi)=\int_0^{C_i}\exp[-\psi S_i(t)]\,dt$$

where C_i is defined as the time between subject i's randomization and the fixed end of the follow-up date. Although C_i is known for uncensored as well as censored subjects, $D_i(\psi)$ is a function of $S_i(t)$ and may depend on the underlying prognosis. Therefore, even when censoring on the T-scale is non-informative, that is, an administrative censoring, censoring on the U-scale is likely to be informative, if $\psi_0\neq 0$ and there is non-random non-compliance. Thus, we cannot replace T_i by $X_i=\min(T_i, C_i)$ to calculate the pseudo-treatment-free event time.

To avoid this problem, Robins and Tsiatis [10] defined a new censoring time $C_i(\psi)=C_i$ if $\psi\leqslant 0$ and $C_i(\psi)=C_i\exp(-\psi)$ if $\psi>0$, according to the direction of treatment effect. For given ψ, let $X_i(\psi)=\min[C_i(\psi), U_i(\psi)]$ and $\Delta_i(\psi)=I[U_i(\psi)>C_i(\psi)]$ to be the new follow-up time and censoring indicator, respectively. $X_i(\psi)$ is observable since $T_i\geqslant C_i$ implies $U_i(\psi)>C_i(\psi)$. Because any function of $\{U_i(\psi), C_i\}$ is independent of random treatment assignment R_i, we have $\{U_i(\psi_0), \Delta_i(\psi_0)\}\coprod R_i$, where the symbol $\coprod$ means independent.

4.3. Results of simulations

Table II shows the results. The constant treatment effect model (3) with $M=3$ was applied in the intensity score analysis, where a logistic regression model (9) was used for the estimation of the propensity score at $t=2$ and 4. From Table II we see that both the AT and ITT estimates were largely biased toward the null in all situations (true value of treatment effect=0.5). The biases increased as the difference of non-compliance proportion between groups increased and as the censoring proportion decreased. The α-errors for the ITT estimate were close to the nominal level of 5 per cent, reflecting that the ITT approach provides a valid test for the null hypothesis of no treatment effect even in the presence of non-random non-compliance.

As expected, the g-estimates performed well in all situations, because the data generation process was based on the SAFT model (10). Note that the powers for the ITT and g-estimate were about the same, reflecting that even though the g-estimation approach uses non-compliance information it does not increase the power against the null hypothesis when compared with the ITT approach.

The intensity score estimates were nearly unbiased and their coverage probabilities were close to the nominal level of 95 per cent in all situations. The α-errors were controlled under the correctly specified parametric model (9). The intensity score estimates gave smaller MSE and narrower confidence intervals than those of the g-estimates, except in the censoring proportion=90 per cent. The powers were slightly increased compared with the g-estimates.

DOI: 10.1002/sim

Table II. Results of simulation studies for AT, ITT, g-estimation, and IS method.

Method	Non-compliance	Censoring (per cent)	Bias	MSE	CI length	95 per cent CP	Power	α Error
AT	45 *versus* 15 per cent	0	−0.415	0.175	0.183	0.0	44.3	100.0
		30	−0.405	0.170	0.216	0.0	41.4	100.0
		90	−0.230	0.074	0.564	64.0	47.5	30.4
	30 *versus* 10 per cent	0	−0.363	0.134	0.180	0.0	84.3	95.0
		30	−0.353	0.127	0.212	13.7	77.7	100.0
		90	−0.190	0.056	0.559	73.7	57.8	22.2
ITT	45 *versus* 15 per cent	0	−0.305	0.095	0.176	0.0	99.4	5.6
		30	−0.295	0.090	0.209	0.1	97.0	5.8
		90	−0.138	0.041	0.567	82.6	71.1	4.8
	30 *versus* 10 per cent	0	−0.253	0.066	0.177	0.0	99.9	4.3
		30	−0.243	0.062	0.209	0.4	99.8	6.1
		90	−0.108	0.033	0.566	87.2	78.4	4.6
g-estimation	45 *versus* 15 per cent	0	0.006	0.013	0.541	95.1	99.4	5.6
		30	0.008	0.018	0.615	94.3	97.0	5.8
		90	0.001	0.037	0.854	94.6	70.9	4.9
	30 *versus* 10 per cent	0	0.001	0.009	0.454	95.4	99.9	4.3
		30	0.003	0.015	0.509	97.0	100.0	6.1
		90	−0.001	0.032	0.825	96.5	78.3	4.6
Intensity score	45 *versus* 15 per cent	0	−0.046	0.005	0.274	95.7	100.0	5.2
		30	−0.019	0.005	0.287	94.5	100.0	4.6
		90	0.060	0.061	0.959	97.8	74.3	4.7
	30 *versus* 10 per cent	0	−0.045	0.004	0.257	95.2	100.0	4.5
		30	−0.018	0.005	0.262	95.2	100.0	4.6
		90	0.053	0.046	0.840	98.0	82.4	4.7

AT, as-treated; MSE, mean-squared error; CI, confidence interval; CP, coverage probability.

5. ANALYSIS OF MEGA STUDY DATA

In the analysis of the MEGA study data, we divided the follow-up period into 10 time intervals with equal space (1 year). Patients were classified as a non-complier in a time interval if he/she switched to the other trial treatment at least once during the interval.

5.1. Estimation of the propensity score

To estimate the propensity score at each time t ($t=0,\ldots,9$), the logistic regression model (4) was used, in which four time-dependent factors as well as 12 baseline factors shown in Table I were included as covariates $H_i(t)$. For the four time-dependent factors, the most recent recorded values were included as covariates $H_i(t)$ in model (4). All TC values were excluded accounting for the multicollinearity of covariates. Among baseline factors, missing data were observed in the values of BMI (0.24 per cent), current smoking (0.18 per cent), and drinking (0.17 per cent). The missing values of BMI were imputed by the mean value of 23.8 (kg/m^2). The latter two factors were imputed by zero (no smoking and no drinking, respectively). Four time-dependent factors were three lipids (TG, HDL-C, and LDL-C) and treatment actually received before time t. For the

DOI: 10.1002/sim

Table III. Predictors of receiving the pravastatin treatment at $t=3$.

Predictors	Odds ratio	95 per cent CI
Baseline covariates		
Assigned treatment	4.645	3.536, 6.102
Age (years)	1.008	0.991, 1.026
Women	0.916	0.663, 1.264
BMI (kg/m^2)	1.008	0.968, 1.050
Current smoker	1.262	0.884, 1.800
Current drinking	0.932	0.684, 1.271
Medication history	1.484	1.086, 2.029
Hypertension	1.169	0.915, 1.493
Diabetes	1.247	0.938, 1.658
TG (mg/dL)	1.001	0.998, 1.003
HDL-C (mg/dL)	0.992	0.974, 1.010
LDL-C (mg/dL)	1.013	1.003, 1.023
Time-dependent covariates		
TG (mg/dL) at $t=2$	1.003	1.001, 1.005
HDL-C (mg/dL) at $t=2$	1.030	1.014, 1.046
LDL-C (mg/dL) at $t=2$	1.010	1.001, 1.015
Treatment received at $t=2$	240.2	179.2, 321.7

CI, confidence interval; medication history: hypercholesterolemia medication history.

missing data of lipid values (21.5 per cent), the regression imputations were separately conducted, where 11 baseline factors, allocation group, and the last observed lipid value were included as covariates in each prediction model.

Table III shows the odds ratio of each factor associated with receiving the pravastatin treatment at time $t=3$. The results for other time points (not shown) were essentially similar to those shown in Table III. For the baseline covariates, patients who were assigned to the pravastatin group and have hypercholesterolemia medication history tended to receive the pravastatin treatment. As expected, the previous use of pravastatin also predicted the use of pravastatin subsequently.

5.2. *Estimation of treatment effect adjusting for treatment changes*

Table IV shows the estimates of treatment effect by several methods. Hazard ratios for stroke event, which was one of the secondary endpoints in the MEGA study, were also presented. Analysis models for stroke were the same as those for CHD events, and similar results for factors associated with receiving the pravastatin treatment were observed (not shown) as shown in Table III. For both CHD and stroke events, two analyses were conducted, where each endpoint was evaluated at 5 or 10 years, respectively. Two intensity score estimates were obtained: one (intensity score 1) was the constant treatment effect by applying model (3) and the other (intensity score 2) was the cumulative treatment effect by applying model (7).

Both the intensity score and g-estimation methods gave the larger treatment effects for pravastatin than the ITT ones for all endpoints. The adjustment effects were larger in the stroke events. The statistically significant effect in the stroke event at 10 years was observed by the intensity score 1 (hazard ratio = 0.51; 95 per cent CI: 0.28–0.95). The results from intensity score 2, in particular

DOI: 10.1002/sim

Table IV. Estimates of treatment effect for CHD and stroke events.

	CHD				Stroke			
	5 years		10 years		5 years		10 years	
Method	HR	95 per cent CI	HR	95 per cent CI	HR	95 per cent CI	HR	95 per cent CI
ITT	0.70	0.50, 0.97	0.67	0.49, 0.91	0.65	0.43, 0.97	0.83	0.57, 1.21
Intensity score 1	0.68	0.44, 1.05	0.59	0.36, 0.99	0.44	0.25, 0.79	0.51	0.28, 0.95
Intensity score 2	0.68	0.46, 1.02	0.66	0.27, 1.60	0.53	0.31, 0.90	0.45	0.17, 1.21
g-estimation	0.65	0.30, 0.91	0.64	0.39, 0.83	0.54	0.26, 0.87	0.63	0.33, 1.26

HR, hazard ratio; CI, confidence interval; intensity score 1, constant treatment effect from model (3); intensity score 2, cumulative treatment effect from model (7); g-estimation, semi-parametric randomization-based analysis using model (10).

at 10 years, gave the wider confidence intervals than those from intensity score 1, which probably reflects the sparse data problems in estimating $\beta_I(t)$. The confidence intervals for the g-estimates contained the null value of 1 whenever the ITT result was not significant.

6. DISCUSSION

In this paper, we developed the intensity score approach for time-to-event outcomes with censoring to estimate the causal treatment effect in the presence of non-random non-compliance. The proposed approach has three major advantages over the g-estimation based on the SAFT model (10). The first advantage is that an artificial recensoring scheme (Section 4.2) is necessary requirement for the g-estimation to account for administrative censoring correctly, while the proposed approach can treat the censoring uniquely within the framework of standard regression models. The rationale for recensoring in the g-estimation is that if the potential baseline failure time U_i is independent of treatment assignment, the same should be true for any function of U_i and C_i since C_i is a baseline covariate. Therefore, there are several choices for an observable random variable that is a function of $\{U_i, C_i\}$ as a basis for inference [13, 16, 22].

The second major advantage of the proposed approaches is that they can be easily extended to the estimation of time-dependent treatment effects such as (6), where the technique of g-estimation has been difficult to apply in practice to the multi-parameter model. Although the constant treatment effect model (1) is very simple, model (6) is more robust to the estimation of dynamic sequential treatments conditional on past medical history. This robustness property of model (6) will be compromised with the sacrifice of the precision as shown in Table IV. To avoid the sparse-data problems, Brumback *et al.* [9] proposed the use of parametric constraints among the $\beta_I(t)$ such as $\beta_I(t)=a_0+a_1t$ depending on context.

The third advantage is its ease of application, that is, the g-estimate can be obtained in three steps: we compute propensity scores, derive intensity scores, and fit an ordinary regression model for any outcome variable, although the correction term must be subtracted from the estimating function to obtain the consistent estimator.

Nevertheless, the g-estimation has a number of advantages over the proposed approach. First one is that it is a semi-parametric randomization-based approach, that is, it preserves the validity of tests of the null hypothesis regardless of what determinants of outcome have influenced a

DOI: 10.1002/sim

patient's decision to comply. Furthermore, the g-estimation provides estimated effects that are of the same sign as the ITT effect and that are only statistically significant if the ITT analysis is statistically significant. In relation to this point, a major drawback of the intensity score approach is that one must be able to specify a correct model for the conditional probability of treatment, $\Pr[S_i(t)|H_i(t)]$, for each t up to the end of follow-up, although the increase of power will be anticipated. Unfortunately, the assumption of no unmeasured confounders (2) is a non-identifiable assumption and is not testable from the observed data. Furthermore, even when assumption (2) is approximately true, we require strong modeling assumptions, since there are many covariates in $H_i(t)$. It is unlikely that these modeling assumptions would be precisely correct. In the MEGA study, many clinically important prognostic factors were measured and all of them were used as covariates to estimate the propensity score at each time. In addition to the prediction model shown in Table III, the analyses based on other prediction models, such as a parsimonious model using a variable selection procedure or full models in which time-dependent covariates, were entered as the difference from the baseline or the absolute past two values, and the intensity score estimates were shown to be insensitive to the selection of the prediction models conditional on the measured covariates.

Another advantage of the g-estimation over the intensity score approach is that one can use the SAFT model (10) to estimate the effect of a treatment on outcome in studies, where at each time t there is a covariate level such that all patients with that level of the covariate are certain to receive the identical treatment. For example, this circumstance implies that the intensity score approach should not be used for the analysis of non-compliance data, in which treatment switching was observed in only one group, because the intensity score at each time will be zero for patients in the complete compliance group. Robins [23] and Robins *et al.* [24] discussed a similar problem, that is, structural zero, for the adjustment of time-dependent confounding and showed that the IPTW (inverse probability of treatment weighted) estimators, which are based on the propensity score, are biased for the data with structural zero.

As Brumback *et al.* [9] have discussed, the intensity score approach resembles the IPTW estimation method based on the marginal structural model (MSM). Although the MSM is useful for estimating the causal effect of the pre-specified treatment regime such as always treat or treat on alternate month [23, 24], it is much less useful for modeling the interaction of treatment with a time-dependent covariate and for estimating the effect of a dynamic treatment plan in which the treatment on a visit depends on a subject's evolving covariate history. It is important to recognize that actual medical treatment regimes including non-compliance data are usually dynamic, and the SNMM is more suitable for parametrizing such dynamic effects. Another difference between the SNMM and the MSM is that the latter makes fewer assumptions than the former by not requiring treatment effects to be constant across strata of covariate history, because the IPTW estimators can be interpreted as standardized parameters [24, 25]. Thus, in theory, the IPTW estimator is more robust than the intensity score one.

In the analyses of the MEGA study data, we observed the larger adjustment effects in the stroke events in spite of the fact that factors associated with non-compliance were nearly the same for CHD and stoke events in each group. The explanatory analyses among the non-compliant cases were conducted to investigate the relation between the non-compliance rate (/year) of each case and the occurrence of each event. These analyses showed that, in the diet group, the effect of non-compliance rate on the non-occurrence of stroke events (odds ratio = 144; 95 per cent CI: 1.3–∞; 5 stroke events among 865 non-compliant cases) was larger than that of CHD events (odds ratio = 14.5; 95 per cent CI: 1.7–150; 19 CHD events among 844 non-compliant cases), while,

DOI: 10.1002/sim

in the pravastatin group, the effect of non-compliance rate on the occurrence of stroke events (odds ratio=5.7; 95 per cent CI: 1.2–26; 16 stroke events among 2440 non-compliant cases) was also larger than that of CHD events (odds ratio=1.3; 95 per cent CI: 0.3–5.3; 20 CHD events among 2441 non-compliant cases). These facts may explain the larger discrepancy between the ITT estimate and the causal one observed in stroke events.

In the MEGA study, like any other clinical trial, dropout of patients during the study period was observed. In addition to the usual loss to follow-up cases, there was another problem of dropouts due to the refusal of further follow-up at 5 years [17, 18]. In this paper, we considered all these dropout cases as non-informative censoring cases. Because observed dropout proportions were not different among treatment groups (loss to follow-up: 546/3966=0.14 in diet group and 594/3866=0.15 in pravastatin group; refusal of follow-up by patients: 278/3966=0.07 in diet group and 270/3866=0.07 in pravastatin group), the effect of these dropouts on the comparison of treatment group may seem to be small. However, these non-administrative censorings may be informative and hence a source of selection bias. To adjust for selection bias due to non-administrative censoring, the IPCW (inverse probability censoring weighted) method has been proposed [26–28]. The underlying idea of the IPCW method is to base estimation on the observed outcomes but weight them to account for the probability of being uncensored. We analyzed the MEGA study data using the IPCW method which can adjust for some types of dependent censorings, and confirmed that there were no large differences between the ITT estimates and the IPCW ones for both CHD and stroke events [29]. Our intensity score method can also incorporate the IPCW method, and this will be a future work.

APPENDIX A

We show that the correction term $N\beta_I C$ must be subtracted from the WLS estimating function to obtain a consistent estimator of β_0 in (1), where $C=(1/N)(\sum_{t=0}^{M-1}\hat{I}_i(t))\omega_i(\sum_{t=0}^{M-1}E(S_i(t)|H_i(t);\hat{\theta}))$ and $\omega_i=\exp(-\mu)\cdot T_i\cdot\exp(-\beta_I\sum\hat{I}_i(t))$. We define the 'estimated' potential outcome under no treatment:

$$\log\hat{T}_{0i}\equiv\log T_i-\beta_0\sum_{t=0}^{M-1}S_i(t)$$

Under model (1) and assumption (2), the estimated potential outcome is mean independent of future treatment given past history, which implies that $E[\hat{I}_i(t)\cdot w_i\cdot(\log\hat{T}_{0i}-\mu)]=0$, $t\leqslant M-1$. Therefore,

$$E\left(\sum_{t=0}^{M-1}\hat{I}_i(t)\cdot w_i\cdot\left[\log T_i-\mu-\beta_0\sum_{t=0}^{M-1}S_i(t)\right]\right)=0 \tag{A1}$$

Now, the WLS estimating equation that $\hat{\beta}_I$ solves under model (3) has unconditional mean zero if and only if $E(\sum_{t=0}^{M-1}\hat{I}_i(t)\cdot w_i\cdot[\log T_i-\mu-\beta_I\sum_{t=0}^{M-1}\hat{I}_i(t)])=0$. Substituting $\hat{I}_i(t)=S_i(t)-E[S_i(t)|H_i(t);\hat{\theta}]$ yields

$$E\left(\sum_{t=0}^{M-1}\hat{I}_i(t)\cdot w_i\cdot\left[\log T_i-\mu-\beta_I\sum_{t=0}^{M-1}S_i(t)+\beta_I\sum_{t=0}^{M-1}E[S_i(t)|H_i(t);\hat{\theta}]\right]\right)=0 \tag{A2}$$

DOI: 10.1002/sim

Comparing (A1) with (A2), it follows that $\hat{\beta}_I$ is consistent for β_0 if for any t, $\hat{I}_i(t) \neq 0$ and $E(\sum_{t=0}^{M-1} \hat{I}_i(t) \cdot w_i \cdot \sum_{t=0}^{M-1} E[S_i(t)|H_i(t); \hat{\theta}]) = 0$.

ACKNOWLEDGEMENTS

MEGA Study is supported by Sankyo Co. Ltd. We also thank reviewers for their comments, which led to a much improved version of the paper.

This research was supported in part by Grant-in-Aid for Scientific Research (A) No. 16200022. Research funds for MEGA Study were provided by the Japanese Ministry of Health, Labor and Welfare for the first 2 years of the study, and thereafter the study was funded by Sankyo Co Ltd, Tokyo.

REFERENCES

1. Efron B, Feldman D. Compliance as an explanatory variable in clinical trials (with Discussion). *Journal of the American Statistical Association* 1991; **86**:9–26.
2. Sommer A, Zegar SL. On estimating efficacy from clinical trials. *Statistics in Medicine* 1991; **10**:45–52.
3. Angrist JD, Imbens GW, Rubin DB. Identification of casual effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–455.
4. Fisher LD, Dixon DO, Herson J, Frankowski RK, Hearron MS, Pease KE. Intention-to-treat in clinical trials. *Statistical Issues in Drug Research and Development*. Marcel Dekker, Inc.: New York, 1990; 331–350.
5. Koch GG, Edwards S. Clinical efficacy trials with categorical data. In *Biopharmaceutical Statistics for Drug Development*, Peace K (ed.). Marcel Dekker, Inc.: New York, 1988; 403–457.
6. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics* 1994; **23**:2379–2412.
7. Robins JM. Causal inference from complex longitudinal data. In *Latent Modelling with Applications to Causality*, Berkane M (ed.). Springer: New York, 1997; 69–117.
8. Robins JM. Correcting for non-compliance in equivalence trials. *Statistics in Medicine* 1998; **17**:269–302.
9. Brumback B, Greenland S, Redman M, Kiviat N, Diehr P. The intensity score approach for adjusting for confounding. *Biometrics* 2003; **59**:274–285.
10. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics* 1991; **20**:2609–2631.
11. Robins JM, Blevins D, Ritter G, Wolfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**:319–336.
12. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Controlled Clinical Trials* 1993; **14**:79–97.
13. Keiding N, Filiberti M, Esbjerg S, Robins JM, Jacobsen N. The graft versus leukemia effect after bone marrow transplantation: a case study using structural nested failure time models. *Biometrics* 1999; **55**:23–28.
14. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: examples from the concorde trial. *Statistics in Medicine* 1999; **18**:2617–2634.
15. Korhonen PA, Laird NM, Palmgren J. Correcting for non-compliance in randomized trials: an application to the ATBC study. *Statistics in Medicine* 1999; **18**:2879–2987.
16. Hernan MA, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* 2005; **14**:477–491.
17. Management of Elevated Cholesterol in the Primary Prevention Group of Adult Japanese (MEGA) Study Group. Design and baseline characteristics of a study of primary prevention of coronary events with pravastatin among Japanese with mildly elevated cholesterol levels. *Circulation Journal* 2004; **68**(9):860–867.
18. Nakamura H, Arakawa K, Itakura H *et al.* Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA Study): a prospective randomised controlled trial. *Lancet* 2006; **368**:1155–1163.
19. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978; **6**:34–58.
20. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
21. Rosenbaum PR, Rubin DM. The central role of the propensity score in observational studies of casual effects. *Biometrika* 1983; **70**:41–55.

DOI: 10.1002/sim

22. Witteman JC, D'Agostino RB, Stijnen T *et al.* G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham heart study. *American Journal of Epidemiology* 1998; **148**:390–401.
23. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Halloran ME, Berry D (eds). Springer: New York, 1999; 95–133.
24. Robins JM, Hernan MA, Brumback BA. Marginal structural models and casual inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
25. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003; **14**:680–686.
26. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate makers. *American Statistical Association Proceedings of the Biopharmaceutical Section* 1993; 24–33.
27. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
28. Robins JM, Finkelstein DH. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with Inverse Probability of Censoring Weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
29. Yoshida M, Matsuyama Y, Ohashi Y for the MEGA Study Group. Estimation of treatment effect adjusting for dependent censoring using the IPCW method: an application to a large primary prevention study for coronary events (MEGA study). *Clinical Trials* 2007; **4**:318–328.

DOI: 10.1002/sim

Japanese Journal of Biometrics Vol. 29, No. 1, 19–34 (2008)

Original Article

A Modification of the 50 %-Conditional Power Approach for Increasing the Sample Size Based on an Interim Estimate of Treatment Difference

Kohei Uemura, Yutaka Matsuyama and Yasuo Ohashi

Department of Biostatistics, School of Health Sciences and Nursing, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
e-mail:uemura@epistat.m.u-tokyo.ac.jp

Recently, flexible approaches with updating of sample size during the course of clinical trials have been proposed; the weighted Z-statistic approach and the 50 %-conditional power approach. In this paper, we propose a modification of the 50 %-conditional power approach, which increases the sample size only when the conditional power based on the unblinded interim results is greater than 50 %. Our method can control the type I error rate due to the restriction on the minimum required sample size ratio under the decision of increasing sample size. Simulation studies showed that the proposed method increased power about 10 % compared with the fixed sample size design and attained higher power than the original 50 %-conditional power approach. Compared with the weighted Z-statistic approach, the proposed method had several promising operating characteristics; a substantial gain in conditional power given the decision of sample size adjustment, a low probability of reaching the maximum sample size, a substantial decrease in the conditional type II error rate given the maximum sample size, and a conservative property of not increasing sample size erroneously under no treatment effect.

Key words: adaptive design; conditional power; interim look; sample size re-estimation; type I error.

1. Introduction

The sample size calculation is an important element in the design of a clinical trial. Typically, determination of sample size rests on knowledge of the expected treatment effect size, which is a function of the expected treatment difference and the variance of an outcome variable. These are usually obtained from previously completed small size clinical trials or historical data. If the actual treatment difference is smaller and/or the actual variance is much larger than expected, the planned sample size will be severely underestimated and consequently it may fail to detect a treatment effect of clinical interest and waste limited resources. It is thus appealing to use

Received December 2007. Revised February 2008. Accepted March 2008.

the information from the current trial at interim stages, updating the initial assumptions and adjusting the sample size if necessary, to ensure adequate power to detect a clinically meaningful treatment difference while maintaining the type I error rate (Chow and Chang, 2007).

If sample size re-estimation is based on estimates of nuisance parameters such as within-group variance, the type I error rate will not be materially inflated (Wittes and Brittain, 1990; Gould, 1992; Wittes et al., 1999; Zucker et al., 1999). However, if sample size re-estimation is based on the observed treatment difference, the type I error rate could be substantially inflated and an appropriate statistical adjustment may be needed to control it (Gould, 2001; Proschan and Hunsberger, 1995; Shun et al., 2001). Mid-course sample size modification methods based on the observed treatment difference have been developed over the last decade by many authors. These include Bauer and Kohne (1994), Proschan and Hunsberger (1995), Fisher (1998), Shen and Fisher (1999), Cui, Hung, and Wang (1999), Lehmacher and Wassmer (1999), and Chen, DeMets, and Lan (2004), and some of them are briefly reviewed in the next section.

In this paper, we will focus our attention on the 50 %-conditional power (CP) approach of Chen, DeMets, and Lan (2004), which is the only method with no need of statistical adjustments to control the type I error rate. The final analysis of their approach is conducted as usual, and that is the merit for a clinician to understand it easily. Furthermore, some authors point out the problem that the weighted Z-statistic approach of Cui, Hung, and Wang (1999) including the methods listed above other than the 50 %-CP approach can reject the null hypothesis even when the usual test used in a fixed sample design fails to reject (Denne, 2001; Posch, Bauer, and Brannath, 2003; Burman, and Sonesson, 2006). Such inconsistency of rejection region occurs, because those methods essentially allocate unequal weights to equally informative observations. Thus, we propose a modification of the 50 %-CP approach using the usual test statistic. Simulation studies are conducted to compare several operating characteristics among the original CP, our proposed CP, and the weighted Z-statistic approach.

2. Sample size re-estimation methods based on the observed treatment difference

2.1 Problems posed by re-estimation using an interim estimate of treatment difference

We consider a situation which is common in phase III clinical trials that involve the comparison of a new treatment with a placebo or standard therapy. A statistical design is specified in the protocol based in part on the specification of a type I error rate α and a power $1-\beta$ at a given effect size δ. At some intermediate point during the course of the trial, researchers examine the outcome data collected so far and decide they wish to modify the original design. For example, the choice of design effect size δ has been over-optimistic, whereas it is now apparent that the benefit of the new treatment is liable to be somewhat less than δ and it is unlikely that a significant result will be achieved at the planned end of the trial. Even so, the estimated effect may still be large enough to be deemed clinically significant and worthwhile.

Consider an outcome variable Y_{ij} for subject i in group j ($j = 1,2$), which is normally distributed with their mean, μ_1 and μ_2, respectively, and common within variance σ^2. The variance σ^2 is assumed known or otherwise it can be estimated from the data. Without loss of generality, we assume $\sigma^2 = 1$, and thus, the effect size is $\delta = \mu_1 - \mu_2$. For our purpose, we consider a two sample test H_0: $\delta = 0$ versus H_1: $\delta = \delta_{pre}$ (> 0) using the one-sided two sample mean test with the significance level $\alpha = 0.025$. Let N_0 denote the initial planned sample size per group for detecting a pre-assumed effect size δ_{pre} with a desired power $1 - \beta$. Thus,

$$N_0 = 2\{(z_\alpha + z_\beta)/\delta_{pre}\}^2, \tag{1}$$

where z_u denotes the $(1-u)$th quantile of the standard normal distribution.

Now suppose the data are examined at an intermediate stage of the trial when n out of N_0 subjects have been collected. Denote the estimate of δ computed from the n subjects per group accumulated so far by

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - Y_{i2}).$$

Consider the situation where $\hat{\delta}$ is positive but somewhat smaller than the planned effect size δ_{pre} at which power $1 - \beta$ was specified. If the true value of δ is close to $\hat{\delta}$, it is unlikely that H_0 will be rejected, that is, the conditional power at $\delta = \hat{\delta}$ is low. However, the researchers now realize that the magnitude of $\hat{\delta}$ is clinically meaningful and the original target effect size δ_{pre} was over-optimistic. This would have required the larger sample size $\lambda^2 N_0$, which can be obtained by substituting $\hat{\delta}$ for δ_{pre} in (1), where $\lambda = \delta_{pre}/\hat{\delta}$.

A naïve approach to this trial would be simply to increase the number of remaining subjects on each group from $(1-t)N_0$ to $\gamma(1-t)N_0$, where $t = n/N_0$ and $\gamma = \dfrac{\lambda^2 - t}{1-t}$, and proceed to use the naïve final test statistic

$$Z_{naive} = \frac{1}{\sqrt{2(\lambda^2 N_0)}}\sum_{i=1}^{\lambda^2 N_0}(Y_{i1} - Y_{i2}).$$

However, since the random variable λ is a function of the first stage data, this Z-statistic does not follow a $N(0,1)$ distribution under H_0 and the test that rejects H_0 when $Z_{naive} > z_\alpha$ does not have type I error rate α. Cui, Hung, and Wang (1999) shows that, typically, the type I error rate of such a test is inflated by 30% to 40%; using other rules to determine the second-stage sample size, it can more than double (Proschan and Hunsberger, 1995; Shun et al., 2001).

2.2 The weighted Z-statistic approach

The weighted Z-statistic approach assigns less weights to subjects enrolled after the decision of increasing the sample size than to those enrolled before the decision (Fisher, 1998; Shen and Fisher, 1999; Cui, Hung, and Wang, 1999). Here, the two-stage version of their approach is shown. Let N denote the re-planned sample size per group and let $M = \lambda^2 N_0$. In practice, there is always an upper limit for the number of available subjects in a trial, then let $N_{\max}$ be the

maximum number of re-planned sample size per group. We will not decrease sample size when the interim result $\hat{\delta}$ is larger than what was expected (Shih, 2001). Thus, re-planned sample size N is determined as,

$$N = \begin{cases} N_0 & if \quad M \leq N_0 \\ M & if \quad N_0 < M < N_{\max} \\ N_{\max} & if \quad M \geq N_{\max} \end{cases}. \tag{2}$$

If the sample size is increased to N, which may depend on the observed value of test statistic based on the first stage n subjects, $Z^{(n)}$, the weighted Z-statistic is defined as,

$$Z_w^{(N)} = \sqrt{t}Z^{(n)} + \sqrt{1-t}Z^{(N-n)}, \tag{3}$$

where $Z^{(n)} = (2n)^{-1/2}\sum_{i=1}^{n}(Y_{i1} - Y_{i2})$, $Z^{(N-n)} = (2(N-n))^{-1/2}\sum_{i=n+1}^{N}(Y_{i1} - Y_{i2})$ are the test statistics based on n and $(N-n)$ subjects, respectively. Under H_0, $Z^{(n)}$ and $Z^{(N-n)}$ are two independent standard normal variables. Note that the $(N-n)$ subjects enrolled after the decision to increase sample size contribute to a constant amount of information fraction $(1-t)$ regardless of N. H_0 is rejected if the weighted Z-statistic $Z_w^{(N)} > z_\alpha$ and the type I error rate is controlled exactly at the nominal level α (Cui, Hung, and Wang, 1999). The weighted Z-statistic approach can be easily extended to a group sequential trial (Cui, Hung, and Wang, 1999), which is equivalent to a variance-spending approach proposed by Fisher (1998) and Shen and Fisher (1999).

2.3 The 50 %-conditional power approach

The basic idea of the 50 %-conditional power (50 %-CP) approach is that the initial planned sample size is increased if and only if the interim result is promising, where a treatment effect is said to be promising if the conditional power under the current trend is greater than 50 %, or the sample size increment to achieve a desired power is no more than a prespecified upper bound $N_{\max}$ (Chen, DeMets, and Lan, 2004). This approach is intended to save the marginally significant result based on N_0 subjects.

In this paper, we define the conditional power $CP(t, z, \delta = \hat{\delta})$ based on the estimate of effect size at an intermediate stage of the trial when n out of N_0 subjects have been collected, such as,

$$CP(t, z, \delta = \hat{\delta}) = \Pr(Z^{(N_0)} > z_\alpha | Z^{(n)} = z; \delta = \hat{\delta}) \times 100 = \Phi((z/\sqrt{t} - z_\alpha)/\sqrt{1-t}) \times 100, \tag{4}$$

where $Z^{(N_0)} = (2N_0)^{-1/2}\sum_{i=1}^{N_0}(Y_{i1} - Y_{i2})$ is the test statistic at the end of study based on the initial planned N_0 subjects and $\Phi(\cdot)$ is the cumulative distribution function for a standardized normal variable. In this approach, if $CP(t, z, \delta = \hat{\delta}) \geq 50$, $M = \lambda^2 N_0$ is calculated at the intermediate stage and the re-planned sample size N is determined using the rule (2). If $CP(t, z, \delta = \hat{\delta}) < 50$, N is set to N_0. The final test statistic is a simple one based on N subjects,

$$Z^{(N)} = \frac{1}{\sqrt{2N}}\sum_{i=1}^{N}(Y_{i1} - Y_{i2}), \tag{5}$$

where H_0 is rejected if the Z-statistic $Z^{(N)} > z_\alpha$. Chen, DeMets, and Lan (2004) showed that increasing the sample size without any adjustment to the test statistic or the final critical value will not inflate the type I error rate if the interim result is promising, that is, $CP(t,z,\delta=\hat{\delta}) \geq 50$. However, their simulation results suggested that the actual type I error rate of the 50 %-CP approach was strictly less than the nominal level for all scenarios and that the type I error rate reduction was substantial, especially in the case of late stage sample size re-estimation, such as $t = 0.8$. Thus, their approach will be too conservative for sample size re-estimation and the power will not be increased so much.

2.4 The proposed modified conditional power approach

To improve the conservative property in power of the 50 %-CP approach, we consider a modification of it, where the conditional power boundary to decide whether or not to increase the initial sample size is set to a lower value Q % than 50 %. If one increases the sample size at $Q \leq CP(t,z,\delta=\hat{\delta}) < 50$, the type I error rate may be inflated. However, the type I error reduction at $CP(t,z,\delta=\hat{\delta}) \geq 50$ might be able to compensate for such inflation. The CP approach including 50 % one determines the re-planned sample size per group according to the following rule,

$$N = \begin{cases} N_0 & \text{if} \quad CP(t,z,\delta=\hat{\delta}) < Q \\ N_0 & \text{if} \quad CP(t,z,\delta=\hat{\delta}) \geq Q \quad \text{and} \quad M \leq N_0 \\ M & \text{if} \quad CP(t,z,\delta=\hat{\delta}) \geq Q \quad \text{and} \quad N_0 < M < N_{\max} \\ N_{\max} & \text{if} \quad CP(t,z,\delta=\hat{\delta}) \geq Q \quad \text{and} \quad M \geq N_{\max} \end{cases}, \tag{6}$$

where $M = \lambda^2 N_0$.

To investigate the effect of lowering the value of Q on the type I error rate, we calculated the actual type I error rate under the rule (6) with different conditional power boundaries $Q = 5$, 10, 15 and 20 %. To calculate the actual type I error rate without any statistical adjustment, we define the change in the type I error rate conditional on the observed data as,

$$\begin{aligned} \Delta(r,t,z) &= \Pr(Z^{(N)} > z_\alpha | Z^{(n)} = z, r, \delta = 0) - \Pr(Z^{(N_0)} > z_\alpha | Z^{(n)} = z, r, \delta = 0) \\ &= \Phi((z\sqrt{t/r} - z_\alpha)/\sqrt{1-t/r}) - \Phi((z\sqrt{t} - z_\alpha)/\sqrt{1-t}) \end{aligned}$$

where $r = N/N_0$. The change depends on the magnitude of sample size ratio (r), the information fraction at which the interim decision is made (t) and the observed test statistic at an intermediate stage (z). The actual type I error rate is $\alpha + E\{I(r>1)\Delta(r,t,z)\}$, where for any proposition A, $I(A)$ equals one if A is true and zero otherwise.

Table 1 gives the actual type I error rates in various scenarios, where the desired powers $1-\beta = 0.8$, 0.9, the nominal level $\alpha = 0.025$, the maximum number of re-planned sample size $N_{\max} = 1.25N_0$, $1.5N_0$, $1.75N_0$, $2N_0$, $2.5N_0$, and the information fraction of intermediate stage $t = 0.2$, 0.5, 0.8 (see Appendix A). Table 1 shows that lowering the value of Q less than 50 % does