Published in final edited form as:
Stat Med. 2014 August 15; 33(18): 3078–3088. doi:10.1002/sim.6149.

Conditional Monte Carlo Randomization Tests for Regression Models

Parwen Parhat, **William F. Rosenberger**[*,†], and **Guoqing Diao**
Department of Statistics, George Mason University, 4400 University Drive, MS 4A7, Fairfax, VA 22030, U.S.A.

Abstract

We discuss the computation of randomization tests for clinical trials of two treatments when the primary outcome is based on a regression model. We begin by revisiting the seminal paper of Gail, Tan, and Piantadosi (1988), and then describe a method based on Monte Carlo generation of randomization sequences. The tests based on this Monte Carlo procedure are design-based, in that they incorporate the particular randomization procedure used. We discuss permuted block designs, complete randomization, and biased coin designs. We also use a new technique by Plamadeala and Rosenberger (2012) for simple computation of conditional randomization tests. Like Gail, Tan, and Piantadosi, we focus on residuals from generalized linear models and martingale residuals from survival models. Such techniques do not apply to longitudinal data analysis, and we introduce a method for computation of randomization tests based on the predicted rate of change from a generalized linear mixed model when outcomes are longitudinal. We show, by simulation, that these randomization tests preserve the size and power well under model misspecification.

Keywords

Generalized linear models; Generalized linear mixed models; Linear rank test; Longitudinal data; Martingale residuals; Time-to-event data

1. Introduction

In randomized clinical trials comparing two treatments, randomization tests provide a distribution-free test of the treatment effect. Under the null hypothesis, data are assumed to be exchangeable, unaffected by random treatment assignment. In clinical trials, this assumption of exchangeability can typically be ensured by the fact that differences due to patient characteristics should be mitigated by the act of randomization. It is not uncommon for the FDA to require a “re-analysis” of data using a randomization test. Typically Monte Carlo simulation is used to sample randomization sequences under a specified randomization procedure, and the *p*-value is computed as the proportion of sequences that yield a test statistic as extreme or more extreme than the observed test statistic once it is assumed that large values are evidence against H_0. The test statistic can be any metric, such

[*]Correspondence to: Department of Statistics, George Mason University, 4400 University Drive, MS 4A7, Fairfax, VA 22030, U.S.A.
[†]wrosenbe@gmu.edu.

as the difference of means or proportions, or a linear rank statistic. In this paper we describe how one might compute randomization tests in the context of regression modeling. Such modeling might result in the primary outcome of a clinical trial, such as from a survival model, a generalized linear model, or a generalized linear mixed model. Such models can lead to randomization tests for adjusted or unadjusted treatment effects, which might be continuous, binary, ordinal, time-to-event or rates of change in longitudinal assessment. In most cases, we can rank residuals from the model as our outcome measure, as described in the seminal paper by Gail, Tan, and Piantadosi [1]. These residuals can be score residuals or, in the case of survival models, martingale residuals. For longitudinal data, we rank the predictors of the random effects from a generalized linear mixed model (GLMM) to use as our outcome measure. Our methods can be applied to complete randomization or common restricted randomization procedures, such as the permuted block design, or Efron's biased coin design (BCD) [2]. Such tests are "design-based", where inference is conducted according to how the trial was designed. This differs from Gail, Tan, and Piantadosi's approach, which are based on equiprobable treatment assignments. We apply techniques of Plamadeala and Rosenberger [3] to compute conditional randomization tests. The flexibility of the approach allows investigators to compute a randomization test for virtually any type of primary outcome measure encountered in modern phase III clinical trials.

2. Background

Let $T_1,\ldots, T_n$ be a randomization sequence, where $T_i = 1$ if subject i is randomized to treatment A; $T_i = 0$ if B, $i = 1,\ldots, n$. Let $N_A(j)=\sum_{i=1}^{j} T_i$ be the number of subjects randomized to treatment A after j assignments, and $N_B(j) = j - N_A(j)$. Rosenberger and Lachin [4, Chapter 3] describe complete and restricted randomization procedures that are used in clinical trials. The procedures can be defined by $p_j = P(T_j = 1|T_1,\ldots, T_{j-1})$. In complete randomization, $T_1,\ldots, T_n$ follow i.i.d. Bernoulli(1/2) distributions. In restricted randomization, prior information on imbalances are used to adaptively balance treatment assignments. Restricted randomization induces correlation among the treatment assignments. The permuted block design (PBD) is probably the most commonly used, and divides the n subjects into B blocks of size $b = n/B$, assuming n/B is an integer. In practice, the last block may go unfilled, resulting in $N_A(n)$ being random. If n/B is an integer and all the blocks are filled, $N_A(n) = n/2$, and $N_A(n)$ is no longer a random variable. One can use either the random allocation rule (RAR) or truncated binomial rule (TBD) to fill each block. Within a block of size b (even), a random allocation rule can be thought of as an urn with $b/2$ balls representing treatment A and $b/2$ balls representing treatment B. Patients are randomized by drawing balls without replacement. The truncated binomial design uses complete randomization to fill each block. Once $b/2$ patients are assigned to one of the treatments, all further patients in the block are assigned to the opposite treatment with probability 1. Biased coin designs were introduced by Efron [2] and extended by Wei [5], Smith [6], and Baldi Antognini and Giovagnoli [7], among others. In this paper, we consider only Efron's original design. Let $D_j = N_A(j) - N_B(j)$ and let $p \in [0.5, 1]$. Then the BCD(p) randomization procedure is given by

$$\begin{aligned} p_j &= 1/2, \quad \text{if} \quad D_{j-1}=0, \\ &= p, \quad \text{if} \quad D_{j-1}<0, \\ &= 1-p, \quad \text{if} \quad D_{j-1}>0. \end{aligned}$$

For both complete randomization and the biased coin design, $N_A(n)$ is a random variable.

Let $X_1,\ldots, X_n$ be some outcome measure, which could be residuals or some other derived quantities, and let $a_{1n},\ldots, a_{nn}$ be a score function of $X_1,\ldots, X_n$, which could be simple rank scores, van der Waerden scores, Savage scores, or other scores (see Rosenberger and Lachin [4, Chapter 7]). The linear rank test can be written as

$$S=\sum_{i=1}^{n}(a_{in}-\overline{a}_n)T_i, \quad (1)$$

where $\overline{a}_n=\sum_{i=1}^{n} a_{in}/n$. A valid test of the null hypothesis that there is no treatment effect can be carried out by permuting the realizations of $(X_1,\ldots, X_n)$ in all possible ways (e.g., Lehmann [8, Chapter 5]) and computing the test statistic under each permutation. Under a randomization procedure, the sequences $(T_1,\ldots, T_n)$ may not be equiprobable, and each sequence will have an associated probability under the randomization procedure of π_k, $\sum_{k=1}^{2^n}\pi_k=1$. An exact p-value can be computed by summing the probabilities of the 2^n sequences where the test statistic $|S_k|$ equals or exceeds the observed statistic $|S_{obs.}|$ for a two-sided test (Rosenberger and Lachin [4, Chapter 7]). We call this the unconditional p-value (e.g., Rosenberger and Lachin [4, Chapter 7]). A Monte Carlo resampling procedure can be used by generating K sequences under the particular randomization procedure used, computing S_k, $k = 1,\ldots, K$. A strongly consistent estimator of the p-value, $\hat{p}$, is the proportion of sequences where $|S_k|$ equals or exceeds $|S_{obs.}|$. A value of K around 2500 will bound the MSE of the p-value by 0.0001.

Complete and restricted randomization can result in imbalances in the treatment assignments. Since $N_A(n)$ is an ancillary statistic, and the unconditional reference set contains sequences of little value such as $N_A(n) = n$ or 0, the conditional reference set is often used in practice with only $\begin{pmatrix} n \\ n_A \end{pmatrix}$ possibilities instead of 2^n, which finds probabilities conditional on $N_A(n) = n_A$, where n_A is the number of subjects assigned to A (e.g., [9, 10]). While N_A is ancillary, many statisticians argue that the realized design should be a part of the analysis (see, for example, [10]). Plamadeala and Rosenberger [3] present a Monte Carlo method for generating randomization tests with respect to this conditional reference set, which requires the same number of sequences be generated as for an unconditional test. When $n_A \neq 0.5n$, this leads to a massive reduction of computational complexity, and allows these tests to be computed quickly. The idea they use is to generate sequences directly from the conditional reference set. Let $p_j(m_{j-1}) = P(T_j = 1|N_A(j-1) = m_{j-1})$. The Plamadeala and Rosenberger method generates sequences via the rule

$$\phi_j = P(T_j = 1 | T_1, \ldots, T_{j-1}) = \begin{cases} p_j(m_j - 1)\frac{P(N_A(n) = n_A | N_A(j) = m_j)}{P(N_A(n) = n_A | N_A(j-1) = m_{j-1})}, & 1 < j \leq n, \\ \frac{P(N_A(n) = n_A | T_j = 1)}{2P(N_A(n) = n_A)}, & j = 1. \end{cases}$$

We call the resulting *p*-value the conditional *p*-value. We use this technique to generate conditional randomization tests for complete randomization, the permuted block design with an unfilled final block, and Efron's biased coin design.

3. Generalized Linear Models: Revisiting Gail, Tan, and Piantadosi (1988)

Gail, Tan, and Piantadosi [1] describe a method to compute randomization tests based on the score residuals of the generalized linear model. Let Y be the outcome variable, X the covariate, and T the treatment indicator (centered with values 1 and −1) which is independent of X; and let $h(\cdot)$ be a known function, and η be the parameter representing the linear relationship. The generalized linear model (GLM) is built as follows:

$$\eta = \mu + T\alpha + X\beta;$$

$$E(Y | X, T) = h(\eta).$$

Under the population model, the null hypothesis of no treatment effect is $\alpha = 0$. In the case of linear regression, where $h(\eta) = \eta$, the estimate of α will be unbiased whether X is included or not. But for certain non-linear regressions (e.g., Poisson regression with count data), the estimator of α based on the traditional population model may lead to inflated type I error rate when X is omitted because the estimated variance of the score function for α is biased. Gail, Tan and Piantadosi [1] proposed to use the randomization distribution of score residuals to estimate the variance; they employed three ways to examine the effect of model misspecification on the population test and their proposed test: the size of the population tests under the false population model, variance estimate ratios between the true variance and the false variance, and the relative efficiency between their proposed test under the false model and the test under the true model. Their proposed test is considered to be efficient most of the time.

Define $\theta(\cdot)$ to be the canonical parameter of an exponential family with first derivative $\theta'(\cdot)$, φ as a scale parameter, $d(\cdot)$ and $b(\cdot)$ as functions identifying the distribution of Y, and $R(\cdot)$ is a function that does not contain any unknown parameters. The log likelihood of Y conditional on X and T is as follows:

$$\ell = \{d(\phi)\}^{-1}[Y\theta(\eta) - b\{\theta(\eta)\} + R(Y)] + c(Y, \phi).$$

The regression model is thus $E(Y|X, T) = h(\eta) = h(\mu + T\alpha + X\beta)$, where $h(\eta) = b'(\theta(\eta))$ and $b'(\cdot)$ is the first derivative. Under the null hypothesis H_0 of the randomization test, there is no

difference between treatment groups *A* and *B*. Gail, Tan and Piantadosi [1] proposed to construct randomization tests based on the randomization distribution of the residuals *r* from the fitted model obtained from maximum likelihood estimation. Note that the use of residuals from the fitted model may lead to exact conditional solutions if the parameter estimates are permutationally invariant [11, 12, 13]. Suppose φ is known, let $\hat{\eta}_0=\hat{\mu}_0+X\hat{\beta}_0$, where $\hat{\mu}_0$ and $\hat{\beta}_0$ represent maximum likelihood estimators of μ and β under the null hypothesis $\alpha = 0$. The score function then is

$$\begin{aligned} U &= \frac{\partial \ell}{\partial \alpha} \\ &= \{d(\phi)\}^{-1}\sum \left(Y\frac{\partial \theta}{\partial \eta}\frac{\partial \eta}{\partial \alpha} - \frac{\partial b}{\partial \theta}\frac{\partial \theta}{\partial \eta}\frac{\partial \eta}{\partial \alpha}\middle|\eta=\hat{\eta}_0 \right) \\ &= \{d(\phi)\}^{-1}\sum T\theta'(\hat{\eta}_0)\{Y-h(\hat{\eta}_0)\}. \end{aligned}$$

The residuals *r* have the following form:

$$r=(d(\phi))^{-1}\theta'(\hat{\eta}_0)(Y-b'(\theta(\hat{\eta}_0)))=(d(\phi))^{-1}\theta'(\hat{\eta}_0)(Y-h(\hat{\eta}_0)).$$

Under the canonical link, $\theta'(\eta) = 1$.

As *T* and *X* are independent, under complete randomization, $E(T) = 0$ and $Var(T) = 1$. Given the condition that $E(\partial \ell^2/\partial \alpha \partial \mu) = E(\partial \ell^2/\partial \alpha \partial \beta) = E(\partial \ell^2/\partial \mu \partial \beta) = 0$ for each element of β, the variance of the score function can be computed under the randomization of *r* and has the form $Var(n^{-1/2}U)=n^{-1}\sum r^2$. This provides an asymptotically normal unconditional randomization test, provided that complete randomization is employed. The test based on this variance is efficient even when the covariates are omitted [1]. However, the variance is difficult to compute under restricted randomization procedures which are often applied in clinical trials. Instead, we directly perform the randomization test by ranking the residuals and computing the linear rank test in (1) and applying the Monte Carlo approach described in Section 2 for both conditional and unconditional tests. Such tests should be considered "design-based" because they are conducted under the randomization procedure used in the trial. We performed the simulation with the same settings as found in Table 1 of Gail, Tan, and Piantadosi [1], and found that our Monte Carlo randomization tests preserve size under model misspecification in the same way that their asymptotic test does, for complete randomization, the permuted block design, and Efron's biased coin design. Figure 1 demonstrates the type I error preservation for eight different generalized linear models, including the Poisson, exponential, and logistic regression models, where *X* is removed. We used van der Waerden scores for normally distributed data and Wilcoxon scores for all others in the linear rank test (Rosenberger and Lachin [4, Chapter 7]). Note that the randomization test preserves the size of the test under each model, as did Gail, Tan, and Piantadosi's asymptotic test. Since our results are design-based, there are minor variations for different randomization procedures.

4. Survival Data

We now apply our methods to time-to-event outcome data, using survival models and ranking martingale residuals. Assume in a clinical study a patient with treatment T and covariate X has failure time W. Because of potential right censoring, we observe $Y = \min(W, C)$ and $\Delta = I(W \leq C)$, where C is the censoring time. The censoring indicator Δ takes value 1 if we observe the failure time and takes value 0 otherwise. The survival function is given by $S(t) = P(W > t)$ and the hazard function by $\lambda(t) = \lim_{\delta \to 0} P(t \leq W < t + \delta | W \geq t)/\delta$.

We consider the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model. Let $\lambda_0(\cdot)$ be an unspecified baseline hazard, and $\Lambda_0(\cdot)$ be the corresponding baseline cumulative hazard function. Given the linear predictor $\eta = T\alpha + X\beta$, the hazard function and survival function at time t for the PH model are given by $\lambda(t|T, X) = \lambda_0(t) \exp(\eta)$, and $S(t|T, X) = \exp(-\Lambda_0(t) \exp(\eta))$. We maximize the partial likelihood function based on n independent and identically distributed observations $\{(Y_i, \Delta_i, T_i, X_i), i = 1,\ldots n\}$ to obtain the maximum likelihood estimators of (α, β), denoted $(\hat{\alpha}, \hat{\beta})$. Inference can be conducted by applying Wald, score, or likelihood ratio tests.

The AFT model is used when the failure time W is assumed to have a certain distribution, and the impact of the covariates is proportional with respect to the time t. Common distributions under the AFT model include exponential, Weibull, lognormal and log-logistic. Let $S_0(t)$ be the baseline survival function. The AFT model specifies that the conditional survival function of W, given T and X takes the form $S(t|T, X) = S_0\{t/\exp(\eta)\}$, where η is the linear predictor defined above.

For a given distribution, if the hazard function satisfies both assumptions of PH model and the AFT model, such as the Weibull and exponential distribution, then these two models are equivalent; the difference is only in the way we parameterize the regression and distribution parameters. However, for some distributions like the log-logistic, only the AFT model can be accommodated. The AFT model for the log-logistic distribution is also called the proportional odds model, as its odds ratio is constant over time but not hazard ratio.

Following the strategy for performing randomization tests under the GLM, here we propose to use residuals from the PH and AFT models as the outcome variable. It can be shown [14] that $E(\Delta|T, X) = \Lambda(Y|T, X)$, where $\Lambda(t|T, X)$ is the conditional cumulative hazard function of W, given T and X. Assuming $\alpha = 0$ (i.e., there is no treatment effect), the martingale residuals under the PH model and the AFT model are given by $\Delta - \hat{\Lambda}_0(Y) \exp(\hat{\eta})$ and $\Delta - (-\log \tilde{S}_0(Y/\exp(\tilde{\eta})))$, where $(\hat{\Lambda}_0, \hat{\eta})$ and $(\tilde{S}_0, \tilde{\eta})$ are the maximum likelihood estimators of (Λ_0, η) and b(S_0, η), respectively.

5. Longitudinal Data

In this section we provide a new algorithm for computing randomization tests for the generalized linear mixed model (GLMM). In particular, we investigate the treatment effect variation over the repeated measures, such as whether a treatment has time varying effect for a patient. Here the residuals from a traditional GLMM model are not appropriate for

evaluating a rate of change as the primary outcome of a clinical trial because they are not a direct metric for the slope. Instead, we propose to use a linear rank test based on the predictors of the random slopes from the GLMM. This novel approach differs from the approaches used by Gail, Tan, and Piantadosi [1] outlined in Sections 3 and 4.

Let Y_{ij} be the response variable for *j*th measurement occasion of the *i*th subject, where $i = 1, \ldots n$, $j = 1, \ldots n_i$. Conditional on a $q \times 1$ vector of random effects $\boldsymbol{b}_i$, $\boldsymbol{Y}_i = (Y_{i,1}, \ldots Y_{i,n_i})$ are independent and each belongs to the exponential family. Let $\boldsymbol{X}_i$ be a $n_i \times p$ matrix, $\boldsymbol{Z}_i$ be a $n_i \times q$ design matrix, μ be a constant intercept, $T_i = 1$ if patient *i* is assigned to group *A*, 0 if patient *i* is assigned to group *B*. Let $h(\cdot)$ be a known function, $\upsilon(\cdot)$ be a variance function and φ be a scale parameter. Then we have

$$\eta_{ij} = \mu + T_i\alpha + \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \boldsymbol{Z}_{ij}'\boldsymbol{b}_i,$$

where

$$E(Y_{ij}|X_{ij}, T_i, \boldsymbol{b}_i) = h(\eta_{ij})$$

and

$$var(Y_{ij}|X_{ij}, T_i, \boldsymbol{b}_i) = v(h(\eta_{ij}))\phi.$$

Here μ, α, $\boldsymbol{\beta}$ are considered as fixed-effects parameters. In the above equations, X_{ij} and η_{ij} denote the covariates and linear predictor of the *j*th occasion for the *i*th subject.

As it is of interest to relate the treatment effect to the repeated measures, let $\boldsymbol{X}_{ij}$ be the vector of independent variables and interactions given by $\boldsymbol{X}_{ij} = (Q_i, \tau_{ij}, \tau_{ij} \times T_i)$, $\boldsymbol{Z}_{ij} = (1, \tau_{ij})$, $\boldsymbol{b}_i = (b_{0i}, b_{1i})'$, where Q_i is some covariate, τ_{ij} is the *j*th time point for the *i*th patient, and b_{0i} and b_{1i} are the random intercept and the random slope for the *i*th patient, respectively. We assume that the random effects $\boldsymbol{b}_i$ follow a bivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{G}$. Then the linear predictor can be written as

$$\eta_{ij} = \mu + T_i\alpha + Q_i\beta_1 + \tau_{ij}\beta_2 + \tau_{ij}T_i\beta_3 + b_{0i} + \tau_{ij}b_{1i}. \quad (2)$$

Under the null hypothesis of the population test, there is no varying treatment via repeated measures, and we have $\beta_3 = 0$. As the random effects are not observed, we can estimate ($\boldsymbol{\beta}$, $\boldsymbol{G}$) based on the likelihood function, given by

$$L(\boldsymbol{\beta}, \phi, \boldsymbol{G}) = \prod_{i=1}^{n} \int f_Y(\boldsymbol{Y}_i|\boldsymbol{b}_i) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i$$

where f_Y is the conditional density function for the response variables and f_b is the density function of the random effects. This likelihood function is obtained by integrating out $\boldsymbol{b}_i$; it depends on the variance-covariance of $\boldsymbol{b}_i$ but not the unobserved $\boldsymbol{b}_i$ [15]. Numerical

integration techniques have to be used to obtain the solution of the likelihood function. There is also no simple solution for the variance of $\hat{\beta}$ which depends on the variance-covariance of the random effects. For the population test, the test statistic $\hat{\beta}_3/\hat{SE}(\hat{\beta}_3)$ approximately follows a *t*-distribution. The robustness of the test for β_3 then depends on the accuracy of the specified variance-covariance structure of the random effects, the distribution of the response, and whether all other assumptions are valid.

To avoid the distributional and variance-covariance assumptions inherent in the traditional population-based inference procedures, we propose to use a randomization test to detect the change of the treatment effects with repeated measures considering the predicted random slope as the fixed outcome. Given maximum likelihood estimators of $\boldsymbol{\beta}$, φ and $\boldsymbol{G}$, the random effects $\boldsymbol{b}_i$ can be predicted as follows:

$$\hat{\boldsymbol{b}}_i = E(\boldsymbol{b}_i | \boldsymbol{Y}_i, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\boldsymbol{G}}). \quad (3)$$

This is the conditional mean of $\boldsymbol{b}_i$ given the responses for the *i*th subject, the estimated fixed effects and variance parameters. Under the null hypothesis of the randomization test, for each subject there is no difference in the repeated measure changes regardless of which treatment is assigned. The algorithm for estimating *p*-values for unconditional tests and conditional tests under the GLMM is then similar to the algorithm under the GLM. Following (2), for a given data set, one performs the regression analysis assuming $\beta_3 = 0$ and obtain the estimates of $\boldsymbol{\beta}$, φ and $\boldsymbol{G}$. Then one predicts the random slope b_{1i} by solving (3) with adaptive Gaussian quadrature, and computes the linear rank test statistic by ranking these $\hat{b}_{1i}$'s. Monte Carlo techniques are then applied as for the GLM.

6. Simulation Results

In this section, we generate survival data and longitudinal data and compare the proposed randomization test with the traditional population test. We compute size and power under both correctly specified and misspecified models. Throughout the simulation studies, we used the same randomization procedure for the generation of the observed statistic and for the analysis.

6.1. Survival Models

To compare the efficiency of the randomization and parametric tests, we consider three different cases of simulated survival data. In the first simulation, we generate the failure time from

$$\lambda(t|T, X) = \exp(T\alpha + X\beta),$$

where $T = 1$ for the treatment group and $T = 0$ for the control group, and X is a Bernoulli random variable with success probability 1/2. We compare the size and power of parametric and randomization tests when X is omitted under the PH model. In the second simulation, we replace 10 percent of the survival times generated from the above model by standard

lognormal outliers and analyze data under the PH model. In the third simulation, we generate failure time data from the proportional odds model

$$S(t|T,X)=\frac{1}{1+\exp(3+T\alpha+X\beta)t},$$

and analyze the data under the PH model.

In all simulations, we set $\beta = 2$ and the treatment effect α is set to be 0 under the null hypothesis and varies from 1 to 3 under the alternative. We generate the censoring time from an Exponential distribution with mean 1. The censoring rate varied from 5 to 20 percent under different scenarios. Throughout the simulation studies in this section, we use Wilcoxon scores in the linear rank test. For each simulation setting, we generate 5000 replications each with 60 subjects. The hypothesis testing results are based on two-sided significance level of 0.05. Each randomization test is performed with $K = 2500$ sequences. For all conditional randomization tests, we use $n_A = 30$.

Table 1 gives the size and power of the traditional and randomization tests when there is an omitted covariate. All tests exhibit type I error control at the nominal level of 0.05. When $\alpha = 1$, the complete conditional randomization test and the model-based test have comparable powers and both of them are more powerful than other randomization tests. When $\alpha = 1.5$, the model-based test performs better than all randomization tests. Table 2 gives the size and power of the traditional and randomization tests when there are lognormal contaminants under the proportional hazards model. Here the traditional test has greatly inflated size and loss of power. However, the randomization tests preserve the size of the test, and power is slightly lower, but much better than that of the traditional tests. Table 3 gives the size and power of the traditional and randomization tests when the model is misspecified. Failure time data are generated from the proportional odds model, but the data are analyzed according the PH model. One can see that all tests perform similarly except for the inflated size of the model-based test. It appears that randomization tests (particularly the conditional randomization tests) are most useful when there are outliers in survival times.

6.2. Longitudinal Models

We now simulate longitudinal data and compare the traditional population-based tests to our Monte Carlo randomization tests. The simulation is performed assuming an identity canonical link function $h(\cdot)$, corresponding to the classical linear mixed model. Specifically, we generate data from the model (2) in Section 5 and assume the residual errors for the ith patient e_i are normally distributed with mean zero and $n_i \times n_i$ variance-covariance matrix $\boldsymbol{R}_i$. The confounding covariate Q_i takes value 0 or 1 with equal probability. We use sample size $n = 100$. For each patient, the repeated time τ varies from 1 to 5; for other parameters we have $\mu = -3$, $\alpha = 2$, $\beta_1 = 2$, and $\beta_2 = 1.5$. The random effects have the multivariate normal distribution with $E(b_{0i}) = E(b_{1i}) = 0$, $var(b_{0i}) = 4$, $var(b_{1i}) = 4$, and $cov(b_{0i}, b_{1i}) = 0$. The error term has the standard normal distribution where $E(e_i) = 0$ and $\boldsymbol{R}_i$ is an identity matrix. Under the null hypothesis, there is no time-varying treatment effect, i.e., $\beta_3 = 0$. All tests are

two-sided with significance level 0.05. For all conditional randomization tests, we use n_A = 46.

There are five scenarios considered as follows:

1. We analyze the data under the correct model specification (Table 4).
2. The conditional normal distribution of $\boldsymbol{Y_i}$ is misspecified; by doing so, we replace 10 percent of the data with $\boldsymbol{Y_i^* = -\eta_i + e_i}$ and analyze data assuming the response having a normal distribution (Table 5).
3. The variance-covariance matrix of the error term is misspecified. Here, instead of generating standard normal distributed errors, the variance-covariance matrix $\boldsymbol{R_i^*}$ has a first-order autoregressive structure, with ρ = 0.5. And we analyze data assuming the error term has a standard normal distribution. The traditional and conditional randomization tests have comparable performance, so we do not present the results.
4. The distribution of $\boldsymbol{b_i}$ is misspecified. We replace 20 percent of b_{1i}'s with Cauchy distributed random variables. We analyze the data assuming the b_i's are normally and independently distributed (Table 6).
5. The variance-covariance matrix of the random effects is misspecified. Instead of using the original variance-covariance structure, we use $var(b_{0i}^*)=4$, $var(b_{1i}^*)=4$, and the variance-covariance matrix $\boldsymbol{G^*}$ has correlation 0.8. We analyze the data assuming b_{0i} and b_{1i} are independent. The traditional and conditional randomization tests have comparable performance, so we do not present the results.

Table 4 demonstrates that there is little loss of efficiency by using a randomization test instead of the traditional model-based population test under the correct GLMM model. Table 5 suggests that the randomization test improves power over the traditional test when the response distribution is misspecified. Table 6 indicates that the population-based test is conservative when the random effects distribution is misspecified. On the other hand, the proposed randomization tests maintain the appropriate size and are substantially more powerful than the population-based test.

7. Examples

In this section we use simple user-friendly SAS macros to compute randomization tests from actual clinical trials data. The SAS macros are available from the first author upon request.

7.1. Example 1. Survival Outcomes

The data for the first example is the remission data [16], which is from a study that assessed the impact of a maintenance therapy 6-Mercaptopurine for prolonging the duration of steroid-induced remission in patients under the age of 20 with acute leukemia. Patients in remission were assigned to either treatment group with 6-Mercaptopurine therapy or placebo using an unknown randomization procedure. There were 92 patients who entered the study, and the study was stopped after the analysis of the first 42 patients (21 pairs). The outcome

variable of this data is the remission survival time of patients; covariates include the log form of white blood count (WBC) and the treatment indicator. The censoring rate was 0.3.

For the traditional model-based population test, we assume patients are randomized to treatment and placebo using complete randomization. We fit the PH model. Under the null hypothesis, $\alpha = 0$. Using the PHREG procedure in SAS, we obtain the MLE $\hat{\alpha}$ the *p*-value for the test is computed based on comparing a Wald test statistic to critical values from the chi squared distribution. The SAS macro computes the randomization test based on the martingale residuals from the PH or AFT models under any of the randomization procedures we have discussed. Both conditional and unconditional tests can be computed. For the randomization test in this example, the PBD with block size 2 is used as the randomization procedure, and blocks are filled using the random allocation rule. The *p*-value for the traditional test based on the PHREG procedure is 0.002. The *p*-value from the randomization test is 0.001 (S.E.=0.0002).

7.2. Example 2. Longitudinal Outcomes

The data set is from a randomized, double-blind and controlled clinical trial of AIDS patients [17]. The purpose of this study is to compare the survival benefits of using different HIV-1 inhibitors therapies. The benefit is measured by CD4 counts. A total number of 1309 patients were randomized into four groups with different reverse transcriptase inhibitors therapies using a permuted block design of unknown block size. Measurements of CD4 counts were recorded based on 8 week intervals. The number of measurements ranges from 1 to 9. Since there are missing values due to skipped visits and dropouts, we only consider the records of 729 patients with the first four visits. We combine groups 2, 3, and 4. The response variable is the log form of the CD4 counts $\log(CD4 + 1)$; covariates include age, gender, treatment, and visit time.

For the traditional test, we assume patients are randomly assigned to the two treatment groups using complete randomization. Using the GLIMMIX procedure, we obtain parameter estimates using restricted maximum likelihood, and compute the *p*-value of the test $H_0 : \beta_3 = 0$ from (2) based on a *t* statistic.

The SAS macro computes the randomization test based on ranking the best linear unbiased predictor of the b_{1i}'s, and can be applied to any randomization procedure mentioned in the paper, as well as for conditional and unconditional tests. In this example, the randomization procedure used was not specified. As only 197 out of 729 patients were in the treatment group, we apply the conditional test based on complete randomization and BCD(2/3). All tests demonstrate a significant time-varying treatment effect. Under the traditional model, the *p*-value for a *t* test of the treatment effect on the rate of change is less than 0.0001; under the randomization model, the *p*-value for the conditional complete randomization test is 0.0002 (S.E.=8.9×10^{-5}) and less than 0.0001 for the conditional BCD(2/3) (S.E.< 6.3×10^{-5}).

8. Discussion

We provide a uniform approach for testing the primary outcome of a clinical trial based on a regression model. This approach is based on Monte Carlo resampling of randomization procedures, and is therefore fully design-based. Both unconditional and conditional tests can be computed for various restricted randomization procedures. The regression approach allows us to deal with most any type of outcome that would be analyzed in a clinical trial of two treatments. We have not discussed test for more than two treatments. These outcomes include adjusted treatment affects from a generalized linear model, time-to-event outcomes, and the rate of change from a longitudinal study. While the metrics used in the linear rank test arise from a parametric or semi-parametric model, the procedure itself is nonparametric, since it is based on the randomization distribution induced by the particular randomization sequence. Such tests can be considered an alternative to fully parametric tests, or as a stand-alone test of the primary outcome when a randomization test is desired.

In this paper, we have shown by simulation that the randomization test tends to preserve size and power when the model assumptions are violated. In particular, for survival data where there are outliers due to lognormal contaminants, the randomization test appears unaffected, but the traditional population test based on the proportional hazards model has inflated size and reduced power. For longitudinal data, the randomization test is robust with respect to misspecification of response distribution and is significantly better in terms of size and power when the distribution of the random effects are misspecified.

Stratified randomization can be incorporated easily into randomization tests by simply computing a randomization test within each stratum, and then summing over strata (e.g., Rosenberger and Lachin [4, Chapter 8]). We have not addressed the issue of missing data. Imputing worst ranks for missing data is one conservative analysis option. See Kennes, Hilgers, and Heussen [18] for a simulation study.

We repeat that a series of SAS macros computes randomization tests for each of these scenarios, and these macros are available upon request from the first author.

Acknowledgments

Professor Rosenberger was supported in part by a grant from the National Science Foundation under the American Recovery and Reinvestment Act 2009. Professors Rosenberger and Diao were supported in part by a grant from the National Cancer Institute under the American Recovery and Reinvestment Act 2009.

References

1. Gail MH, Tan WY, Piantadosi S. Tests for no treatment effect in randomized clinical trials. Biometrika. 1988; 75(1):57–64.
2. Efron B. Forcing a sequential experiment to be balanced. Biometrika. 1971; 58(3):403–417.
3. Plamadeala V, Rosenberger WF. Sequential monitoring with conditional randomization tests. Annals of Statistics. 2012; 40(1):30–44.
4. Rosenberger, WF.; Lachin, JM. Randomization in Clinical Trials: Theory and Practice. John Wiley and Sons; New York: 2002.
5. Wei LJ. An application of an urn model to the design of sequential controlled clinical trials. Journal of the American Statistical Association. 1978; 73(363):559–563.

6. Smith RL. Sequential treatment allocation using biased coin designs. Journal of the Royal Statistical Society B. 1984; 46(3):519–543.
7. Baldi Antognini A, Giovagnoli A. A new 'biased coin design' for the sequential allocation of two treatments. Journal of the Royal Statistical Society C. 2004; 53(4):651–664.
8. Lehmann, EL. Nonparametrics: Statistical Methods Based on Ranks. Springer; New York: 2006.
9. Cox DR. A remark on randomization in clinical trials. Utilitas Mathematica A. 1982; 21:245–252.
10. Berger VW. Pros and cons of permutation tests in clinical trials. Statistics in Medicine. 2000; 19(10):1319–1328. [PubMed: 10814980]
11. Commenges D. Transformations which preserve exchangeability and application to permutation tests. Journal of Nonparametric Statistics. 2003; 15(2):171–185.
12. Langsrud Ø. Rotation tests. Statistics and computing. 2005; 15(1):53–60.
13. Pesarin, F.; Salmaso, L. Permutation tests for complex data: theory, applications and software. John Wiley & Sons; 2010.
14. Breslow N. The proportional hazards model: applications in epidemiology. Communications in Statistics A. 1978; 7(4):315–332.
15. Fitzmaurice, G.; Laird, N.; Ware, J. Applied Longitudinal Analysis. John Wiley and Sons; New York: 2004.
16. Freireich EJ, Gehan E, Frei EF, Schroeder LR, Wolman IJ, Anbari R, Buget EO, Milld SD, Pinkel D, Selaway OS, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia. Blood. 1963; 21(6):699–716.
17. Henry K, Erice A, Tierney C, Balfour HH, Fischl MA, Kmack A, Liou SH, Kenton A, Hirsch MS, Phair J, et al. A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced aids. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology. 1998; 19(4):339–349. [PubMed: 9833742]
18. Kennes LN, Hilgers RD, Heussen N. Choice of the reference set in a randomization test based on linear ranks in the presence of missing values. Communications in Statistics–Simulation and Computation. 2012; 41(7):1051–1061.

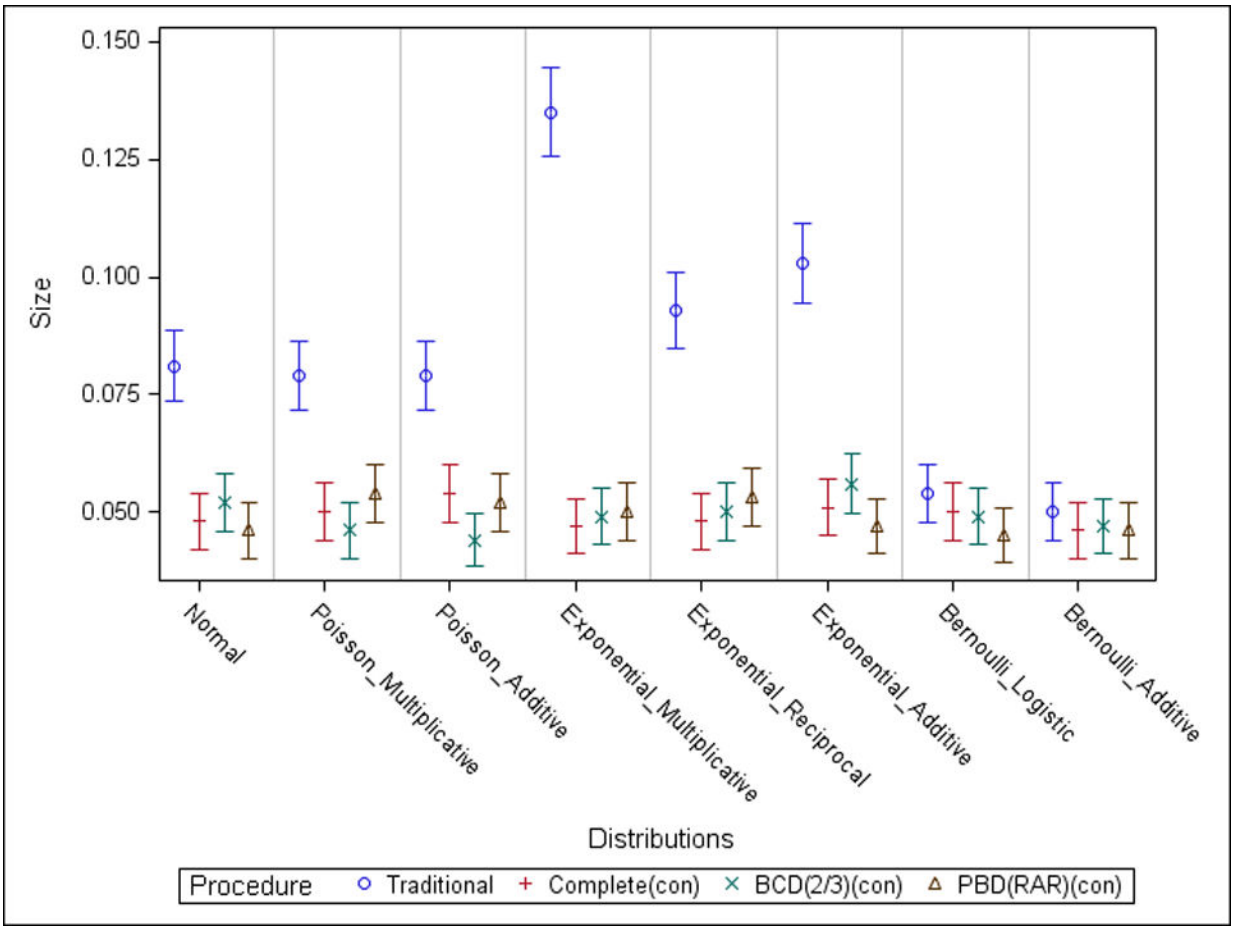

Figure 1.

Size (95% confidence interval) of the traditional test and conditional randomization tests at significance level of 0.05 under GLM for different types of distributions when X is omitted. The traditional procedure gives the size of the population test Z^* defined in Gail, Tan and Piantadosi [1]; conditional randomization tests include complete, BCD(2/3) and PBD(RAR) procedure respectively. These results are based on 500 replicates with the simulation setting of $n = 100$, $n_A = 46$, and $K = 2500$.

Table 1

Size and power of the traditional model-based population test and randomization tests for the PH model with omitted covariates under the PH model (n = 60, K = 2500, 5000 replications).

Test	Size	Power ($\alpha = 1$)	Power ($\alpha = 1.5$)
Model-based	0.0504	0.5788	0.9174
Randomization (Complete unconditional)	0.0418	0.4200	0.7812
Randomization (Complete conditional)	0.0446	0.5708	0.8596
Randomization (BCD(2/3) unconditional)	0.3080	0.4904	0.8522
Randomization (BCD(2/3) conditional)	0.0394	0.4558	0.8200

Table 2

Size and power of the traditional model-based population test and randomization tests for the PH model with lognormal contaminants under the PH model (n = 60, K = 2500, 5000 replications).

Test	Size	Power ($\alpha = 2$)	Power ($\alpha = 3$)
Model-based	0.1608	0.3018	0.5358
Randomization (Complete unconditional)	0.0403	0.5002	0.7948
Randomization (Complete conditional)	0.0396	0.5774	0.8418
Randomization (BCD(2/3) unconditional)	0.0388	0.5184	0.8012
Randomization (BCD(2/3) conditional)	0.0386	0.4542	0.7526

Table 3

Size and power of the traditional model-based population test and randomization tests for the PH model under the misspecified modeling assumption (n = 60, K = 2500, 5000 replications).

Test	Size	Power ($\alpha = 1.5$)	Power ($\alpha = 2$)
Model-based	0.0610	0.7352	0.9196
Randomization (Complete unconditional)	0.0438	0.7472	0.9306
Randomization (Complete conditional)	0.0455	0.7482	0.9330
Randomization (BCD(2/3) unconditional)	0.0413	0.7516	0.9366
Randomization (BCD(2/3) conditional)	0.0406	0.7480	0.9212

Table 4

Size and power of the traditional model-based population test and randomization tests under the correctly specified GLMM (n = 100, K = 2500, 5000 replications).

Test	Size	Power ($\beta_3 = 1$)	Power ($\beta_3 = 1.5$)
Model-based	0.0564	0.6968	0.9582
Randomization (complete)	0.0440	0.6878	0.9558
Randomization (conditional)	0.0426	0.6904	0.9546
Randomization (BCD(2/3) unconditional)	0.0445	0.6918	0.9572
Randomization (BCD(2/3) conditional)	0.0387	0.6824	0.9490

Table 5

Size and power of the traditional model-based population test and randomization tests under the GLMM with misspecified response distribution (n = 100, K = 2500, 5000 replications).

Test	Size	Power ($\beta_3 = 1$)	Power ($\beta_3 = 1.5$)
Model-based	0.0596	0.3738	0.6220
Randomization (complete)	0.0550	0.4438	0.7470
Randomization (conditional)	0.0390	0.4472	0.7476
Randomization (BCD(2/3) unconditional)	0.0444	0.4346	0.7600
Randomization (BCD(2/3) conditional)	0.0459	0.4264	0.7434

Table 6

Size and power of the traditional model-based population test and randomization tests under GLMM with misspecified distribution for random effects (n = 100, K = 2500, m = 5000).

Test	Size	Power ($\beta_3 = 1$)	Power ($\beta_3 = 1.5$)
Model-based	0.0288	0.2190	0.3792
Randomization (complete)	0.0451	0.5232	0.8488
Randomization (conditional)	0.0440	0.5246	0.8598
Randomization (BCD(2/3) unconditional)	0.0387	0.5460	0.8686
Randomization (BCD(2/3) conditional)	0.0465	0.5224	0.8588