Published in final edited form as:
Contemp Clin Trials. 2012 January ; 33(1): . doi:10.1016/j.cct.2011.08.008.

A comparison of statistical approaches for physician-randomized trials with survival outcomes

Margaret R. Stedman[a,b,*], **Robert A. Lew**[c,d], **Elena Losina**[a,b,c], **David R. Gagnon**[c,d], **Daniel H. Solomon**[e,b], and **M. Alan Brookhart**[f]

[a]Orthopedics and Arthritis Center for Outcomes Research, Department of Orthopedics, Brigham and Women's Hospital, Boston, MA, USA

[b]Harvard Medical School, Boston, MA

[c]Department of Biostatistics, Boston University School of Public Health, Boston, MA

[d]MAVERIC, VA Cooperative Studies, Boston V.A. Healthcare System, Boston, MA

[e]Division of Rheumatology, Immunology, Allergy, Department of Medicine, Brigham and Women's Hospital, Boston, MA

[f]Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC

Abstract

This study compares methods for analyzing correlated survival data from physician-randomized trials of health care quality improvement interventions. Several proposed methods adjust for correlated survival data however the most suitable method is unknown. Applying the characteristics of our study example, we performed three simulation studies to compare conditional, marginal, and non-parametric methods for analyzing clustered survival data. We simulated 1,000 datasets using a shared frailty model with (1) fixed cluster size, (2) variable cluster size, and (3) non-lognormal random effects. Methods of analyses included: the nonlinear mixed model (conditional), the marginal proportional hazards model with robust standard errors, the clustered logrank test, and the clustered permutation test (non-parametric). For each method considered we estimated Type I error, power, mean squared error, and the coverage probability of the treatment effect estimator. We observed underestimated Type I error for the clustered logrank test. The marginal proportional hazards method performed well even when model assumptions were violated. Nonlinear mixed models were only advantageous when the distribution was correctly specified.

Keywords

Cluster Randomized Trials; Survival Analysis; Physician-Randomized Trials; Permutation Test; Simulation Study; Shared Frailty Model

[*]Corresponding author. Address: Orthopedics and Arthritis Center for Outcomes Research, Department of Orthopedics, Brigham and Women's Hospital, 75 Francis Street, BC-4-4016, Boston, MA 02115, USA. Phone number: 617-525-7973, mstedman2@partners.org, mstedman2@gmail.com (Margaret R. Stedman).

1. Introduction

Physician-randomized trials are clinical trials where each physician is randomized to a treatment or control intervention. Although the intervention is delivered at the physician level, effectiveness of the intervention is often measured at the patient level, so that patient data are clustered by the randomized physician. This serves to reduce contamination among participants and allows physicians to provide consistent treatment to all their patients. However, the design induces correlations in the data which must be accounted for in the analysis. Performance of these methods depends upon: the number of clusters, the cluster size, and the correlation structure. Typically the correlation structure is unknown, the number of clusters exceeds one hundred, and the cluster sizes vary greatly [1].

Usual methods of analysis for survival data: the logrank test, the Cox proportional hazards model, and the accelerated failure time model, assume the data are independent. If these methods are applied to correlated data the estimation of effect size may be correct, but standard errors will not be adjusted for the correlations in the data and Type I error will exceed nominal size. Several methods have been proposed to correct the standard errors in these analyses. Jung and Jeong recently published a method for the clustered logrank test [2]. Lee, Wei, and Amato [3] popularized the marginal proportional hazards model with robust standard errors. Alternatively, one could model the frailty with a nonlinear mixed model as a counterpart to the parametric accelerated failure time approach [4].

Few studies have attempted to compare methods for analyzing clustered survival data. Glidden et al. [5] performed a simulation study of a clustered design where randomization occurred within cluster and found that conditional models performed better than marginal models. Loeys et al. [6] compared marginal and conditional methods for cluster randomized trials and found that the power was similar between the marginal and conditional methods. Cai et al. [7] compared the clustered permutation test to the usual logrank test and found that the clustered permutation test improved preservation of Type I error.

In this simulation study we compare marginal, conditional, and nonparametric methods for clustered survival analysis under the conditions of fixed cluster size, variable cluster size, and various random effects. We examine power, Type I error, bias, coverage, and mean squared error to determine the best method of analysis for physician-randomized trials. All data are simulated using the design and characteristics of two physician-randomized trials of an educational intervention for osteoporosis management. In section 2 we begin with a description of two example studies. In section 3 we discuss the application of the shared frailty model to the physician-randomized trial. In section 4 we review the various estimation methods that may be used to analyze the data. In sections 5 and 6 we perform a simulation study to evaluate the methods. Actual results from the example studies are presented for comparison in section 6.2. In section 7 we make recommendations on how best to analyze physician-randomized trials based on this research.

2. Example Studies

Osteoporosis is a disease of the elderly that makes bones prone to fracture. According to practice guidelines, all patients at moderate to high risk of osteoporosis should either receive a bone mineral density scan to rule out osteoporosis or be prescribed a preventive medication to treat the disease [8]. Despite these guidelines many patients do not receive adequate treatment [8]. Physician education or “academic detailing” programs have been designed to improve management and prevention of this disease, however the effectiveness of this education effort as an intervention is unknown. We attempted to evaluate the effectiveness of the education program on improving osteoporosis management in two trials, one occurring in Pennsylvania, the other in New Jersey [9, 10]. Data were analyzed by

survival analysis to determine if the physician education program significantly improved a patient's chances of being correctly managed to prevent disease. Both trials served as models in the design of the simulation studies.

The PACE study enrolled Pennsylvania Medicare beneficiaries who were eligible for a state-run pharmaceutical benefits plan. Patients at risk of osteoporosis were selected for the physician-randomized trial with a two-way factorial design. A total of 828 physicians with 13,455 patients participated in one of four groups: 1) physician education, 2) patient education, 3) patient education plus physician education, or 4) usual care [9]. For the purposes of this simulation we focused on comparing two groups: group 3, patient education plus physician education and group 4, the usual care group. The average age of this study population was 82. Patient claims data were used to measure the outcome of a bone mineral density scan or prescription for a preventative osteoporosis medication. Patients were followed using insurance claims data for 487 days or until they lost insurance coverage, were admitted to a nursing home, or died. From this, approximately 10% of the patients was censored. On average there were 16 patients per physician with a range of 2 to 65 patients per physician. See figure C.1 for distribution of the number of patients per physician.

The HORIZON study enrolled patients at-risk of osteoporosis from the Horizon New Jersey health insurance plan. This population was slightly younger than the PACE study with an average age of 68. Loss due to lack of coverage, nursing home admittance, or death was 15%. The trial randomized 434 physicians with a total of 1973 patients to a combination treatment of physician and patient education. Patients were followed using patient claims data to determine if they received osteoporosis management (bone mineral density scan or preventive osteoporosis medication). On average there were four patients per physician with a range of 1 to 148 patients per physician (see figure C.1) [10].

3. The Model

We used a shared frailty model to represent the example physician-randomized trials [11]. We defined $k = 1, ..., K$ physician clusters assigned at random to either treatment (X=1) or control (X=0) group. Each physician contributed $i = 1, ..., n_k$ patients to the study so that $N=\sum_{k=1}^{K} n_k$ was the total number of patients in the study. Let (T_{ki}; $k = 1, ..., K$, $i = 1, ..., n_k$) be the time until each patient received either a bone mineral density scan or preventive osteoporosis medication. Let (D_{ki}; $k = 1, ..., K$, $i = 1, ..., n_k$) be the censoring time for patient i of physician k. X_k; $k = 1, ...K$ is a binary indicator for treatment assignment at the physician level. We assumed that time until osteoporosis management and censoring were conditionally independent given the physician cluster and that the patients in each treatment group shared a common survival distribution, hazard, and cumulative hazard function conditioned on the physician visited. We observed the minimum follow-up time T_{ki}^{o} between T_{ki} and D_{ki} and indicated if censoring occurred by C_{ki}, where $C_{ki}=I\{T_{ki}^{o} \leq D_{ki}\}$ so that our data consisted of (T_{ki}^{o}, C_{ki}; $k = 1, ..., K$, $i = 1, ..., n_{k)}$.

A random effect w_k with density $f_\theta(.)$ is included in the model to measure physician to physician differences in treatment of patients at risk of osteoporosis. In the shared frailty model for the hazard,

$$\lambda_{ki}(t)=w_k\lambda_0(t)exp(\beta x_k)$$

$\lambda_{ki}(t)$; $k = 1, ..., K$, $i = 1, ..., n_k$, predicts the patient specific hazard at time t given X_k and w_k. $\lambda_0(t)$ is the baseline hazard and β is the coefficient for the treatment effect [11]. If we were to

combine w_k with the baseline hazard $\lambda_0(t)$, we could interpret the shared frailty model as a hazard model where the baseline hazard is shared among patients with the same physician.

4. Estimation Methods

We examined four different methods of estimation and hypothesis testing to evaluate the educational intervention. These included a parametric approach (nonlinear mixed effect models), a semi-parametric approach (marginal proportional hazards models with robust standard errors), and two nonparametric methods (the clustered permutation test and the clustered logrank test). Each of these methods uses a different means to adjust for the correlations in the data.

4.1. Parametric Method

Nonlinear mixed effect models use a parametric approach to fit the data. Users specify the distribution of the baseline hazard and random effect to fit the shared frailty model. Estimates are solved for by maximum likelihood estimation. If the distribution of the baseline hazard and random effect is correct, estimates should be asymptotically efficient and unbiased [12]. However, in applied situations these distributions are usually unknown. Common distributions selected for fitting the baseline hazard include the Weibull, log logistic, lognormal, and generalized gamma distribution [11]. The random effect is often fitted with the gamma, lognormal, positive stable, and inverse Gaussian distributions [11].

For a full description of the estimation method see Duchateau and Janssen [4]. Estimates are obtained from the likelihood, which is the product of the probability distribution function (p.d.f.) of the baseline hazard for the patients who receive treatment for osteoporosis, the survival function for the patients who are censored prior to receiving treatment, and the distribution of the random physician effect [4, 13]. The likelihood for cluster *k* is:

$$L(T_{ki}^{o}, C_{ki}, w_k; k=1,\ldots,K, i=1,\ldots,n_k) = \int_{-\infty}^{\infty} \prod_{i=1}^{n_k} [\lambda_{ki}(t_{ki})]^{C_{ki}} \exp\left[-\int_0^{t_{ki}} \lambda_{ki}(t)\,dt\right] f_\theta(w_k)\,dw_k$$

Parameter estimates may be obtained using the NLMixed procedure in SAS [14]. There is no closed form solution for the likelihood so integration of the likelihood is approximated by Gaussian quadrature [15].

4.2. Semi-parametric Method

Marginal proportional hazards with robust standard errors is a semi-parametric method for analyzing clustered survival data. For details on this method see Lee et al [3]. The marginal proportional hazards model does not assume a distribution for the baseline hazard or random effect but it does assume marginal proportional hazards. Only shared frailty models with a Weibull baseline hazard and a positive stable distribution for the random effect are known to satisfy the marginal proportional hazards assumption [11]. The positive stable distribution is highly skewed and may not be an appropriate assumption for all clustered data.

The marginal proportional hazards approach is one of the most popular methods of analysis for its ease of interpretation and application to survival data. Parameter estimates are derived by ignoring the clustering. This is also called an “independence working model” [3]:

$$\lambda_{ki}(t) = \lambda_0 exp(\beta X_k)$$

λ_0 is an average estimate of the baseline hazard for all patients (distinct from the conditional model where the baseline hazard was physician specific).

Estimation of the treatment effect, β, occurs as in usual Cox proportional hazards regression with a partial likelihood function. The partial likelihood is a ratio of the likelihoods for patients who receive osteoporosis management relative to patients who have not been managed for osteoporosis. For an ordered set of times $t_1 < t_2 < ... < t_J$ without ties the partial likelihood is based on the jth time [11]:

$$L(\beta)=\prod_{j=1}^{J}\frac{\exp\left[\beta x_{(j)}\right]}{\sum_{k\in Q_j}\sum_{i\in R_j} exp\left[\beta x_{ki}\right]}$$

where Q_j and R_j represent the respective set of doctors and patients at risk at time j. Since the baseline hazard is assumed to be marginally proportional across treatment groups it cancels out from the likelihood equation. Usual maximum likelihood estimation is applied to the ratio to solve for the parameter estimates. Lee et al. [3] proved that estimates from the working independence model are consistent and asymptotically Normal(0,V). A sandwich estimator is used to adjust the covariance matrix V for the correlation within physician post hoc [3].

4.3. Nonparametric Methods

Two nonparametric methods were tested on the simulated datasets: the clustered permutation test [16] and the clustered logrank test [2, 17]. Both tests do not estimate the hazard ratio however they may be used for significance testing. The advantage to these methods is that they make no assumptions about the structure of the variance, the distribution of the baseline hazard, or random effect.

The clustered logrank test adjusts the variance of the usual logrank test for the clustering in the data. See Jung and Jeong [2] for details on the method. In brief, the authors expanded the following formula for the usual logrank test statistic (R):

$$R=\sqrt{K}\int_{s=1}^{S} W(s)\{dH_1(t)-dH_2(s)\}$$
$$W=\frac{Y_1(s)Y_2(s)}{nY(s)}$$

to demonstrate that the within cluster statistic is asymptotically distributed as $N(0, \sigma^2)$. W is a weight specific to the logrank test which is non-preferential to early or late occurring events. Y is the total number of events occurring in each treatment group. S represents ordered intervals of time and H_1 and H_2 are the cumulative hazards for groups 1 and 2 estimated at time s. The within cluster mean squared error is summed across independent clusters to obtain the variance of the clustered logrank test. Since the distribution of the test statistic relies on asymptotic assumptions it may not perform as well for small samples. Under large sample conditions results should be comparable to results from the marginal proportional hazards method with robust standard errors [2].

The clustered permutation test is an exact method that can be applied to any test statistic. See Stedman et al. [16] for details on the method. The clustered permutation test permutes the treatment assignment while keeping the clusters intact to generate a distribution for the test statistic under the null hypothesis. (The null hypothesis assumes that the distribution of the outcome is the same for both treatment groups so that observations are equally likely to

be assigned to either treatment group.) The significance of the observed test statistic is measured with respect to the distribution of permuted test statistics.

5. Simulation Study

We performed four simulations to replicate the example data described in section 2. For all simulated datasets we assumed a shared frailty model (section 3) with a Weibull distribution for the baseline hazard. The Weibull distribution was selected for its flexibility because parameters have both proportional hazard and accelerated failure time forms. Observation time was censored at 296 days, the maximum follow-up time in the HORIZON study. Each simulation includes two treatment groups with 229 physician clusters per treatment group. Within physician correlation was simulated by including a physician specific random effect in the model.

For all simulations the baseline hazard, $\lambda_0(t)$, followed a Weibull distribution with a shape of 0.9. In the first simulation, "fixed cluster simulation", we generated 458 physician clusters with 8 patient observations per cluster. The conditional parameters simulated for the treatment effect were: 0.00, -0.04, -0.09, -0.18, -0.21, -0.45, and -0.72. The frailty was simulated from the lognormal distribution. For the second and third simulation "variable cluster simulations", we varied the cluster size to match the distributions of cluster sizes in the PACE and HORIZON datasets. See figure C.1 for the distribution of cluster size for these datasets. In the fourth simulation "random effect simulation" we fixed the cluster size at eight patients while varying the distribution of the frailty to include: the uniform(0,1), Gamma, Bernoulli, T, and Positive Stable [18] distributions. The Bernoulli and T random effects were exponentiated before being added to the model. The positive stable distribution ($\beta = 1$, $\alpha = 0.5$) was simulated from two independent uniform random variables as described by Chambers [18] and Shu [19]. The Bernoulli distribution was included to represent a scenario where there are two groups of doctors: those that respond well to treatment and those that respond poorly to treatment. In the random effect simulation we tested three levels for the treatment parameter: 0.00, -0.21, and -0.45. 1000 datasets were generated for each treatment effect. 6000 additional datasets were generated to evaluate Type I error. All data for this project were simulated with the IML procedure in SAS v9.1.

Each of the methods described in section 4 was applied to the simulated datasets. Nonlinear mixed effect models were fitted with a Weibull distribution for the baseline hazard and a lognormal distribution for the random effect [14]. The marginal proportional hazards model was implemented with the PHREG procedure in SAS assuming Breslow method for tied outcomes [20]. See Appendix A for examples of the SAS code for these methods. The clustered logrank test was applied using a SAS Macro [17]. The clustered permutation test [16] entailed 100 permutations of the Wald test statistic for each analyzed dataset.

Each estimation method was evaluated for Type I error, power, coverage probability, bias, and mean squared error (MSE) based on at least 1000 simulations. Power and Type I error were determined by the proportion of the results with a significant treatment effect. Coverage probability was estimated as the proportion of results where the 95% confidence interval included the simulated treatment effect (based on the marginal estimates). Bias was defined as the average difference between the estimated marginal treatment effect and the simulated marginal treatment effect. MSE was estimated from the average of the squared difference between the estimated marginal treatment effect and the simulated marginal treatment effect. Confidence intervals were based on normal approximations of the statistics mentioned. To test for differences in the methods we paired results by simulation and summarized our findings using the McNemar and paired t-tests.

5.1. Marginal versus Conditional Estimates

Parameters from nonlinear mixed effect models give conditional estimates of effect, meaning that the estimates depend on the physician visited. For example, by exponentiating $\hat{\beta}$ we obtain the hazard ratio for the risk of osteoporosis management in the treated group compared to the control group predicated on the primary care physician. Additionally, physician specific predicted hazards can be obtained from the model to describe the physician to physician variability in osteoporosis management with and without the intervention.

Marginal proportional hazards models compute marginal estimates of effect, meaning that the treatment effect of the educational intervention, β, is averaged across physicians. Population averaging results in "marginal estimates" that are shifted towards the null compared to conditional estimates [21]. Interpretation of marginal estimates is less complex than conditional models because there is a single estimate to describe the hazard ratio for the treatment effect. For example, by exponentiating $\hat{\beta}$ we obtain the hazard ratio for the risk of osteoporosis management in the treated group compared to the control group for the entire study population.

Choice of estimate depends upon the study question and the model assumptions. The estimates differ from both a computational and a conceptual perspective. If the goal is to examine physician to physician variability in response to treatment, this can only be evaluated with a conditional model. If the purpose of the study is to measure the average response to treatment, this is more easily estimated from a marginal model. A conditional estimate cannot be obtained from a marginal model and a marginal estimate cannot be easily obtained from a conditional model.

Since marginal proportional hazards models yield marginal estimates and the nonlinear mixed models yield conditional estimates we encounter some difficulties when attempting to make direct comparisons between the two. Conditional estimates from binary and survival outcomes tend to be more extreme and have greater variability than their marginal counterparts. Estimates from the marginal model tend to be biased towards the null. The problem is akin to the issue of collapsibility in nonlinear models [22–24]. In this case, the random effect, w_k from the shared frailty model is a conditional confounder. The marginal model averages over the omitted physician covariate which results in a biased estimate of the treatment effect. The degree of bias also depends on the amount of censoring and the distribution of the random effect [22, 25]. In order to compare like with like we have to convert all conditional estimates to marginal estimates. See Appendix B for the method used to convert conditional parameters from the simulated dataset and conditional estimates from the frailty model to marginal estimates.

6. Results

6.1. Results from Simulation

Table 1 displays results from the fixed cluster simulation where the random physician effect is lognormally distributed and the cluster size is fixed. The effect size listed in the first column is the conditional parameter for the treatment effect from the simulation. See Appendix C for a table of the hazard ratios for each effect size listed. The MSE and bias have been rescaled by a factor of 10. All results are based on at least 1000 iterations. We find that Type I error is below 0.05 for the clustered logrank test. The marginal proportional models with robust standard errors (MPHM) and nonlinear mixed models (NLMM) are slightly more powerful than the nonparametric methods. We found no significant difference in power between the MPHM and the NLMM (p=0.11). MSE is also similar between these methods (p=0.7163). Most intervals contains 95% coverage except for the NLMM result

with effect size −.72 where coverage exceeds 95% and the MPHM result with effect size 0 where coverage is below 95%. For both models the bias is consistently negative and shifts away from the null for all treatment effects. Also, the bias appears to decrease with larger effect sizes. NLMM yields less biased estimates than MPHM even after they are converted to population averaged estimates of effect (mean difference=−0.004, $p < 0.0001$).

Tables 2 and 3 present results from the variable cluster simulations where the cluster sizes varies according to the observed study distributions presented in Figure C.1. The results are similar to those reported for the fixed cluster simulation except that the NLMM appear to be more powerful than the MPHM. Again, we find an underestimate of Type I error for the clustered logrank test in the HORIZON pattern. Also, NLMM underestimates Type I error in the PACE pattern. All other intervals contain the nominal Type I error of .05. Across cluster size patterns most intervals contain a coverage probability of 95%. There were a few exceptions for the NLMM results with effect sizes −0.45 and −0.72 where coverage exceeded 95% and the MPHM null effect size, where coverage is below 95%. Overall, the NLMM had lower MSE than the MPHM (HORIZON mean difference= .0072, PACE mean difference=.0015, $p < 0.0001$ for both cluster size patterns). Again both methods tend to give negatively biased estimates. NLMM yield less biased estimates than the MPHM even after they are converted to population averaged estimates of effect (HORIZON mean difference = −0.005, PACE mean difference = −0.003, $p < 0.0001$ for both cluster patterns).

Table 4 displays results from the random effect simulation where the distribution of the random effect is varied and the cluster size is fixed. Only three treatment effects are simulated for each distribution of the random effect: 0, -.21 and -.45. Across all distributions for the random effect intervals for Type I error contained .05. The MPHM and the clustered logrank test are equally as powerful as the NLMM under non-lognormal random effects. Coverage is much more consistent with the MPHM than the NLMM. Many of the NLMM confidence intervals for the gamma, T, uniform, and positive stable distributions range slightly above 95%. Mean squared error is slightly less overall for the MPHM method than the NLMM method (mean difference=-92×10^{-5}, $p < .0001$). This difference is apparent in the gamma, T, and positive stable distributions. Bias is inconsistent across methods and distributions. The simulated uniform random effect results in the most biased estimates with bias close to 0.05. In previous simulations the bias was consistently negative. Here we find a positive bias for the gamma, T, and positive stable simulated distributions. Although generalizations should be made with caution, overall the MPHM have less biased estimates than the NLMM (mean difference=−0.008 $p < 0.0001$).

6.2. Results from Example Studies

Each of the four estimation methods was tested on actual data from the PACE and HORIZON studies (see section 2). The HORIZON study randomized 434 physicians to physician education or the control intervention. Treatment was targeted to the physician and the patients of the physician. Number of patients at risk per physician ranged from 1 to 148 patients. Patients were followed for osteoporosis management over a duration of 296 days [10]. In Table 5 we show that there is an improvement in receiving osteoporosis management of at least 21% greater in the physician education group than in the control group (MPHM HR=1.48, NLMM HR=1.21). All methods except NLMM found the improvement to be significant ($p < 0.05$).

The PACE study enrolled 828 physicians into a 2-way factorial design of education for osteoporosis management. For the purpose of this study we simplified the design to compare only the group receiving the combined treatment of patient and physician education to the group with usual care (414 PCPs). The number of patients at risk per physician was between 2 and 65. The maximum follow-up time was 487 days [9]. Irrespective of the method

selected, no difference in risk of osteoporosis management was detected between the education group and the control group (see Table 5).

7. Discussion

Our findings show that the marginal proportional hazards models with robust standard errors and nonlinear mixed models perform better than the nonparametric methods. The nonlinear mixed effect model has a slight advantage over the marginal proportional hazards method when the distribution of the random effect is lognormally distributed. When the distribution of the random effect is not lognormally distributed the marginal proportional hazards method performs better with respect to coverage, bias, and MSE. It is interesting that the positive stable distribution does not significantly favor the marginal proportional hazards model, although it is the only frailty distribution with marginal proportional hazards. Type I error is preserved across most methods and scenarios tested. It is somewhat surprising that the clustered logrank test does not perform better since it is asymptotically equivalent to the marginal proportional hazards model [2]. It is possible that it might perform better with larger cluster sizes.

Our results differ from some results reported in the literature. Glidden et al. [5] compared methods of analysis for trial randomization within cluster. They found that marginal proportional hazards underestimate coverage when there are under 5 clusters and relative efficiency declines for frailty models when cluster size is small. It is possible that the marginal method may perform worse in cluster randomized trials where there are fewer clusters, but this cannot be confirmed by our results. We did not experiment with fewer clusters because physician-randomized trials typically randomize many physicians. We did experiment with cluster size. Typically we randomize small clusters in physician randomized trials and our results show that the relative efficiency of the nonlinear mixed model declines when we imposed the cluster pattern of the PACE and HORIZON studies. (In these simulations many of the clusters only contained one or two patients.) However, we find that the decline in efficiency is worse for the marginal model than for nonlinear mixed model. van Breukelen et al. [26] compared equal and unequal cluster sizes and found that highly skewed cluster sizes can decrease efficiency by 10%. We find the decline in efficiency to also depends on the estimation method selected.

Both Glidden et al. [5] and Hsu et al. [27] found the gamma random effect model was robust to misspecification of the distribution. We found that the lognormal conditional model was not as robust as the marginal proportional hazards model under the various random effect distributions of the data simulated. A gamma distributed random effect offers more parameters than the lognormal random effect we tested. SAS methods exist to test a conditional model with a gamma random effect however convergence of the estimates may be less reliable [13].

In this study we found that most estimates converged under the conditions tested. The clustered logrank test underestimated Type I error. The clustered permutation test had reliable Type I error but it was not as powerful as the other methods considered. Based on these results, we recommend either the marginal proportional hazards with robust standard errors or nonlinear mixed effect models for analysis of physician-randomized trials with survival outcomes depending on if it is reasonable to assume a lognormal random effect. The nonlinear mixed effect model is the most powerful and is the least biased when it is safe to assume lognormal random effects. Results from the nonlinear mixed effect model additionally allow one to obtain physician specific estimates. This may be useful for characterizing individual clusters and developing a more targeted intervention.

We did not include the penalized quasi-likelihood (PQL) approach among the methods tested. In a previous analysis with binary outcomes we found PQL to perform poorly under the conditions of our physician randomized trials [1]. Others have reported estimation convergence problems with PQL and cluster randomized trials [6]. We also did not consider bootstrap methods for adjustment of standard errors. Others have reported reasonable results with this method [27]. In previous studies with a binary outcome we found the method may need further refinement for variable cluster sizes [1]. Since the nonparametric methods tested were not as powerful, we expect that the bootstrap method would not be as powerful as the nonlinear mixed models and the Cox method. We did not consider other methods for handling tied data with Cox proportional hazards. It is possible that performance may vary depending on the number of tied outcomes in the dataset. Our results are limited to the design structures that were tested. It is likely that these methods would perform differently for community intervention trials where there are few clusters of large cluster size. We did not consider different censoring patterns, left censoring, or informative censoring. More studies are needed to test the methods under various censoring conditions.

This study demonstrates performance of several methods for analyzing physician-randomized trials with survival outcomes. The marginal proportional hazard method has become popular for its ease of use and interpretability. This study confirms that marginal proportional hazards models perform as well as the other methods available for analyzing these data.

Acknowledgments

The authors would like to thank Youyi Shu for her SAS program for simulating positive stable data and Sin-Ho Jung for his Fortran program for the logrank test. This work was supported in part by grants from the National Institutes of Health to Drs Brookhart (AG-027400) and Stedman (NRSA/T32 AR055885).

References

1. Stedman MR, Gagnon DR, Lew RA, Solomon DH, Brookhart MA. An evaluation of statistical approaches for analyzing physician-randomized quality improvement interventions. Contemporary Clinical Trials. 2008; 29(5):687–95. [PubMed: 18571476]
2. Jung S, Jeong J. Rank tests for clustered survival data. Lifetime Data Analysis. 2003; 9:21–33. [PubMed: 12602772]
3. Lee, EW.; Wei, LJ.; Amato, DA. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein, J.; Goel, P., editors. Survival Analysis: State of the Art. 1992. p. 237-45.
4. Duchateau, L.; Janssen, P. The Frailty Model. New York, New York: Springer; 2008.
5. Glidden DV, Vittingho3 E. Modelling clustered survival data from multicentre clinical trials. Statistics in Medicine. 2004; 23:369–88. [PubMed: 14748034]
6. Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. Statistics in Medicine. 2001; 20:3753–67. [PubMed: 11782031]
7. Cai J, Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. Statistics in Medicine. 2000; 19:2963–73. [PubMed: 11042626]
8. Solomon DH, Brookhart MA, Gandhi TK, Karson A, Gharib S, Orav J, et al. Adherence with osteoporosis practice guidelines: A multilevel analysis of patient, physician, and practice setting characteristics. The American Journal of Medicine. 2004; 117:919–24. [PubMed: 15629730]
9. Solomon DH, Polinski JM, Stedman M, Truppo C, Breiner L, Egan C, et al. Improving care of patients at-risk for osteoporosis: A randomized controlled trial. Journal of General Internal Medicine. 2007; 22:362–7. [PubMed: 17356969]
10. Solomon DH, Katz JN, Finkelstein J, Polinski J, Stedman M, Brookhart MA, et al. Osteoporosis improvement: A large-scale randomized controlled trial of patient and primary care physician education. Journal of Bone and Mineral Research. 2007; 22:1808–15. [PubMed: 17645403]

11. Klein, JP.; Moeschberger, ML. Survival Analysis. 2. New York, New York: Springer-Verlag; 1997.
12. Diggle, PJ.; Liang, KY.; LZS. Analysis of Longitudinal Data. 1. New York: Oxford University Press Incorporated; 1994.
13. Liu L, Huang X. The use of gaussian quadrature for estimation in frailty proportional hazards models. Statistics in Medicine. 2008; 27:2665–83. [PubMed: 17910008]
14. Gharibvand L, Liu L. Analysis of survival data with clustered events. SAS Global Forum 2009. 2009; 237:1–11.
15. SAS Institute Inc. SAS OnlineDoc 9.2. Cary, North Carolina: 2007.
16. Stedman MR, Gagnon DR, Lew RA, Solomon DH, Losina E, Brookhart MA. A sas macro for a clustered permutation test. Computer Methods and Programs in Biomedicine. 2009; 95:89–94. [PubMed: 19321221]
17. Stedman, MR.; Gagnon, DR.; Lew, RA.; Jung, S.; Losina, E.; Brookhart, MA. A sas macro for a clustered logrank test. 2011. In press
18. Chambers JM, Mallows CL, Stuck BW. A method for simulating stable random variables. Journal of the American Statistical Association. 1976; 71(354):340–4.
19. Shu, Y.; Klein, JP. Master's thesis. Medical College of Wisconsin; 1997. A sas macro for the positive stable frailty model.
20. Allison, PD. Survival Analysis Using the SAS System. Cary, North Carolina: SAS Institute Inc; 1995.
21. McCulloch, CE.; Searle, SR. Generalized, Linear, and Mixed Models. New York, New York: John Wiley & Sons Inc; 2001.
22. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials. Controlled Clinical Trials. 1998; 19:249–56. [PubMed: 9620808]
23. Gail MH, Wieand S, Piantadosi S. Biased estimates of tratment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984; 71(3):431–44.
24. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology. 1996; 7(5):498–501. [PubMed: 8862980]
25. Henderson R, Oman P. Effect of frailty on marginal regression estimates in survival analysis. Journal of the Royal Statistical Society Series B (Statistical Methodology). 1999; 61(2):376–9.
26. van Breukelen GJP, Candel MJJM, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in Medicine. 2007; 26:2589–603. [PubMed: 17094074]
27. Hsu L, Gorfine M, Malone K. On robustness of marginal regression co-effcient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified. Statistics in Medicine. 2007; 26:4657–78. [PubMed: 17348081]

Appendix A. SAS code for analytic methods

The clustered permutation test and clustered logrank test are freely available SAS macros. See Stedman et al. [16, 17] for documentation of the software.

Examples of the code to implement nonlinear mixed models and the marginal proportional hazards models are copied below. For documentation of these methods see the SAS manual. [15]

```
*****Nonlinear mixed
models*****************************************************
proc nlmixed data=example tech=quanew update=bfgs;
bounds gamma > 0;
parms b0=7
b1=0
```

```
gamma=1.0
logsig=1;
linp = b0 + b1*treat + z;
alpha = exp(-linp);
G_t = exp(-(alpha*time1)**gamma);
g = gamma*alpha*((alpha*time1)**(gamma-1))*G_t;
ll = ((cens1=0)*log(G_t)) + ((cens1=1)*log(g));
model time1 ~ general(ll);
random z ~ normal(0,exp(2*logsig)) subject=drid;
estimate 'b1_ph' (-1) * b1 / gamma;
estimate 'b0_ph' (-1) * b0 / gamma;
run;


*******Marginal proportional hazards
models************************************
proc phreg data=example covs (aggregate);
model time1 * cens1 (0)= treat;
id drid;
run;
```

Appendix B. Converting Conditional Estimates to Marginal Estimates

The conditional hazard from the shared frailty model at time t is:

$$\lambda(t|x,w)=w\lambda_0(t)exp(\beta x)$$

Then the marginal hazard function at time t is:

$$\lambda(t|x)=E_w\left[exp\left\{w\lambda_0(t)exp\left(\beta x\right)\right\}\right]$$

In survival analysis bias in the marginal estimates depends on the form of the frailty and the degree of censoring in the data. For a positive stable frailty distribution the marginal model is of a proportional hazards form so that unbiased marginal estimates may be obtained from the conditional estimates [25]. For other distributions marginal proportional hazards are not maintained and the integral in intractable so that bias can be more substantial. We suggest the following Monte-Carlo simulation method to approximate the integral:

1. Using the parameters simulated (or estimated from the nonlinear mixed models) simulate a large dataset with one observation per cluster. We simulated 40,000 clusters with one observation per cluster (20,000 clusters per treatment group). A random effect was added to each observation according to the distribution simulated (or estimated).
2. Apply regular Cox proportional hazards regression to estimate marginal parameters from the large dataset in part 1.

Appendix C. Translating conditional estimates from frailty models to hazard ratios

In all simulated and applied models a Weibull distribution was specified for the baseline hazard. Estimates from the Weibull model may be converted to either a proportional hazards or accelerated failure time form by applying the following formula. To distinguish the parameters we will denote the proportional hazards parameter as β_{ph} and the accelerated failure time parameter as β_{af}. The scale, represented by σ_{af}, of the Weibull distribution is the inverse of the shape of the distribution.

$$\beta_{ph} = \frac{-\beta_{af}}{\sigma_{af}}$$
$$\text{Hazard Ratio} = \exp\beta_{ph}$$

To facilitate interpretation of the failure time estimates presented in tables 1 through 4 we have translated the treatment effects simulated into hazard ratios (see table 6).

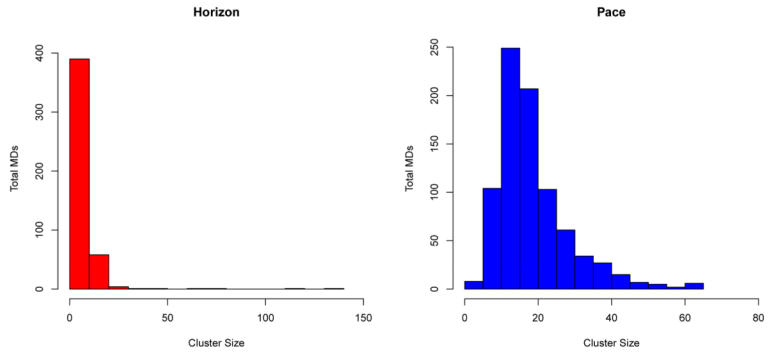

Figure C.1.
Distribution of cluster size

Table 1

Estimate and 95% CI for Power, Coverage, *MSE*[a], and *Bias*[a] by Method and Effect Size for Fixed Cluster Simulation

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
0.00[b]	Power	0.050 (0.044,0.055)	0.050 (0.045,0.055)	0.030 (0.026,0.034)	0.050 (0.045,0.055)
	Coverage	0.938 (0.932,0.943)	0.952 (0.947,0.957)		
	MSE	0.115 (0.111,0.118)	0.115 (0.112,0.119)		
	Bias	−0.245 (−0.270,− 0.221)	−0.232 (−0.256,− 0.207)		
−0.04	Power	0.052 (0.038,0.066)	0.053 (0.039,0.067)	0.050 (0.036,0.064)	0.051 (0.037,0.065)
	Coverage	0.948 (0.934,0.962)	0.959 (0.947,0.971)		
	MSE	0.106 (0.097,0.116)	0.107 (0.097,0.116)		
	Bias	−0.191 (−0.254,− 0.128)	−0.168 (−0.231,− 0.105)		
−0.09	Power	0.116 (0.096,0.136)	0.113 (0.093,0.133)	0.113 (0.093,0.133)	0.104 (0.085,0.123)
	Coverage	0.954 (0.941,0.967)	0.959 (0.947,0.971)		
	MSE	0.105 (0.096,0.115)	0.105 (0.096,0.115)		
	Bias	−0.199 (−0.261,− 0.136)	−0.174 (−0.236,− 0.111)		
−0.18	Power	0.334 (0.305,0.363)	0.340 (0.311,0.369)	0.331 (0.302,0.360)	0.310 (0.281,0.339)
	Coverage	0.949 (0.935,0.963)	0.957 (0.944,0.970)		
	MSE	0.104 (0.094,0.113)	0.102 (0.093,0.112)		
	Bias	−0.213 (−0.274,− 0.151)	−0.184 (−0.246,− 0.122)		
−0.21	Power	0.420 (0.389,0.451)	0.426 (0.395,0.457)	0.416 (0.385,0.447)	0.406 (0.376, 0.436)
	Coverage	0.955 (0.942,0.968)	0.960 (0.948,0.972)		
	MSE	0.100 (0.091,0.110)	0.100 (0.091,0.109)		
	Bias	−0.161 (−0.223,− 0.100)	−0.131 (−0.193,− 0.070)		
−0.45	Power	0.963 (0.951,0.975)	0.966 (0.955,0.977)	0.963 (0.951,0.975)	0.960 (0.948,0.972)
	Coverage	0.952 (0.939,0.965)	0.960 (0.948,0.972)		
	MSE	0.096 (0.087,0.105)	0.096 (0.087,0.105)		
	Bias	−0.120 (−0.180,− 0.060)	−0.057 (−0.118,0.004)		
−0.72	Power	1.000 (1.000,1.000)	1.000 (1.000,1.000)	1.000 (1.000,1.000)	1.000 (1.000,1.000)

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
	Coverage	0.952 (0.939,0.965)	0.975 (0.965,0.985)		
	MSE	0.091 (0.082,0.099)	0.090 (0.082,0.099)		
	Bias	−0.095 (−0.154,− 0.036)	−0.015 (−0.074,0.044)		

[a] Estimates have been rescaled by a factor of 10

[b] Estimates of Type I error are based on 7, 000 simulations

Table 2

Estimate and 95% CI for Power, Coverage, *MSE*[a], and *Bias*[a] by Method and Effect Size for HORIZON Cluster pattern (Variable Cluster Simulations)

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
0.00[b]	Power	0.057 (0.052,0.062)	0.048 (0.043,0.053)	0.039 (0.034,0.043)	0.047 (0.042,0.052)
	Coverage	0.923 (0.916,0.929)	0.956 (0.951,0.961)		
	MSE	0.061 (0.058,0.064)	0.126 (0.122,0.130)		
	Bias	−0.235 (−0.252,− 0.217)	−0.224 (−0.250,− 0.198)		
−0.04	Power	0.067 (0.051,0.083)	0.061 (0.046,0.076)	0.060 (0.045,0.075)	0.060 (0.045,0.075)
	Coverage	0.938 (0.923,0.953)	0.954 (0.941,0.967)		
	MSE	0.197 (0.181,0.214)	0.127 (0.116, 0.137)		
	Bias	−0.229 (−0.315,− 0.143)	−0.198 (−0.267,− 0.129)		
−0.09	Power	0.105 (0.086,0.124)	0.114 (0.094,0.134)	0.096 (0.078,0.114)	0.096 (0.078,0.114)
	Coverage	0.937 (0.922,0.952)	0.952 (0.939,0.965)		
	MSE	0.197 (0.181,0.214)	0.124 (0.114,0.134)		
	Bias	−0.241 (−0.326,− 0.155)	−0.202 (−0.270,− 0.134)		
−0.18	Power	0.240 (0.213,0.267)	0.298 (0.270,0.326)	0.222 (0.196,0.248)	0.214 (0.189,0.239)
	Coverage	0.941 (0.926,0.956)	0.952 (0.939,0.965)		
	Bias	0.195 (0.178,0.211)	0.121 (0.111,0.130)		
	MSE	−0.253 (−0.338,− 0.168)	−0.213 (−0.280,− 0.146)		
−0.21	Power	0.291 (0.263,0.319)	0.383 (0.353,0.413)	0.276 (0.248,0.304)	0.274 (0.246,0.302)
	Coverage	0.940 (0.925,0.955)	0.954 (0.941,0.967)		
	MSE	0.191 (0.175,0.207)	0.119 (0.109,0.128)		
	Bias	−0.198 (−0.283,− 0.113)	−0.157 (−0.224,− 0.090)		
−0.45	Power	0.812 (0.788,0.836)	0.947 (0.933,0.961)	0.818 (0.794,0.842)	0.780 (0.754,0.806)
	Coverage	0.941 (0.926,0.956)	0.965 (0.954,0.976)		
	MSE	0.184 (0.168,0.199)	0.112 (0.103,0.121)		
	Bias	−0.143 (−0.227,− 0.059)	−0.073 (−0.139,− 0.008)		
−0.72	Power	0.981 (0.973,0.989)	1.000 (1.000,1.000)	0.987 (0.980,0.994)	0.977 (0.968,0.986)

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
	Coverage	0.941 (0.926,0.956)	0.970 (0.959,0.981)		
	MSE	0.177 (0.162,0.192)	0.105 (0.097,0.114)		
	Bias	−0.120 (−0.202,− 0.038)	−0.036 (−0.100,0.028)		

[a] Estimates have been rescaled by a factor of 10

[b] Estimates of Type I error are based on 7,000 simulations

Contemp Clin Trials. Author manuscript; available in PMC 2013 December 04.

Table 3

Estimate and 95% CI for Power, Coverage, *MSE*[a], and *Bias*[a] by Method and Effect Size for PACE Cluster pattern (Variable Cluster Simulations)

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
0.00[b]	Power	0.049 (0.044,0.055)	0.044 (0.039,0.048)	0.046 (0.041,0.051)	0.048 (0.043,0.054)
	Coverage	0.921 (0.914,0.927)	0.955 (0.950,0.960)		
	MSE	0.045 (0.043,0.047)	0.126 (0.122,0.130)		
	Bias	−0.244 (−0.259,− 0.229)	−0.224 (−0.250,− 0.199)		
−0.04	Power	0.072 (0.056,0.088)	0.064 (0.049,0.079)	0.067 (0.051,0.082)	0.065 (0.050,0.080)
	Coverage	0.942 (0.927,0.957)	0.961 (0.949,0.973)		
	MSE	0.139 (0.127,0.151)	0.123 (0.113,0.134)		
	Bias	−0.202 (−0.274,− 0.129)	−0.191 (−0.259,− 0.123)		
−0.09	Power	0.117 (0.097,0.137)	0.114 (0.094,0.134)	0.112 (0.092,0.132)	0.111 (0.091,0.131)
	Coverage	0.944 (0.930,0.958)	0.956 (0.943,0.969)		
	MSE	0.139 (0.127,0.150)	0.122 (0.112,0.132)		
	Bias	−0.210 (−0.281,− 0.138)	−0.194 (−0.261,− 0.126)		
−0.18	Power	0.282 0.254,0.310)	0.297 (0.269,0.325)	0.278 (0.250,0.306)	0.265 (0.238,0.292)
	Coverage	0.949 (0.935,0.963)	0.950 (0.936,0.964)		
	MSE	0.135 (0.124,0.147)	0.120 (0.110,0.130)		
	Bias	−0.225 (−0.296,− 0.154)	−0.209 (−0.275,− 0.142)		
−0.21	Power	0.354 (0.324,0.384)	0.378 (0.348,0.408)	0.352 (0.322,0.382)	0.340 (0.311,0.369)
	Coverage	0.947 (0.933,0.961)	0.952 (0.939,0.965)		
	MSE	0.132 (0.121,0.143)	0.118 (0.108,0.128)		
	Bias	−0.171 (−0.242,− 0.100)	−0.153 (−0.220,− 0.086)		
−0.45	Power	0.934 (0.919,0.949)	0.952 (0.939,0.965)	0.934 (0.919,0.949)	0.916 (0.899,0.933)
	Coverage	0.945 (0.931,0.959)	0.964 (0.952,0.976)		
	MSE	0.124 (0.114,0.135)	0.111 (0.102,0.121)		
	Bias	−0.113 (−0.182,− 0.044)	−0.063 (−0.128,0.003)		
−0.72	Power	1.000 (1.000,1.000)	1.000 (1.000,1.000)	1.000 (1.000,1.000)	1.000 (1.000,1.000)

β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
	Coverage	0.947 (0.933,0.961)	0.969 (0.958,0.980)		
	MSE	0.118 (0.108,0.128)	0.105 (0.096,0.114)		
	Bias	−0.090 (−0.157,− 0.023)	−0.022 (−0.085,0.042)		

[a] Estimates have been rescaled by a factor of 10

[b] Estimates of Type I error are based on 7,000 simulations

Contemp Clin Trials. Author manuscript; available in PMC 2013 December 04.

Table 4

Estimate and 95% CI for Power, Coverage, *MSE*[a], and *Bias*[a] by Method, Effect Size and Distribution of the Random Effect (Random Effect Simulations)

Distribution	β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
Bernoulli(.3)	0.00[b]	Power	0.054 (0.048,0.059)	0.053 (0.048,0.058)	0.052 (0.047,0.057)	0.054 (0.049,0.059)
		Coverage	0.947 (0.942,0.952)	0.953 (0.948,0.958)		
		MSE	0.070 (0.067,0.072)	0.070 (0.068,0.072)		
		Bias	−0.085 (−0.105,− 0.066)	0.026 (0.007,0.046)		
Bernoulli(.3)	−0.21	Power	0.639 (0.609,0.669)	0.639 (0.609,0.669)	0.638 (0.608,0.668)	0.634 (0.604,0.664)
		Coverage	0.940 (0.925,0.955)	0.949 (0.935,0.963)		
		MSE	0.066 (0.060,0.071)	0.063 (0.058,0.069)		
		Bias	−0.093 (−0.143,− 0.043)	−0.028 (−0.078,0.021)		
Bernoulli(.3)	−0.45	Power	1.000 (1.000,1.000)	1.000 (1.000,1.000)	1.000 (1.000,1.000)	0.999 (.997,1.001)
		Coverage	0.943 (0.929,0.957)	0.957 (0.944,0.970)		
		MSE	0.062 (0.056,0.067)	0.061 (0.056,0.066)		
		Bias	−0.087 (−0.135,− 0.038)	−0.044 (−0.092,0.004)		
Gamma(.6,1)	0.00[b]	Power	0.052 (0.047,0.057)	0.051 (0.046,0.057)	0.051 (0.046,0.056)	0.051 (0.045,0.056)
		Coverage	0.947 (0.942,0.953)	0.969 (0.965,0.973)		
		MSE	0.206 (0.199,0.213)	0.222 (0.215,0.230)		
		Bias	−0.075 (−0.108,− 0.041)	0.083 (0.048,0.117)		
Gamma(.6,1)	−.21	Power	0.252 (0.225,0.279)	0.245 (0.218,0.272)	0.247 (0.220,0.274)	0.231 (0.205,0.257)
		Coverage	0.945 (0.931,0.959)	0.967 (0.956,0.978)		
		MSE	0.200 (0.180,0.219)	0.221 (0.199,0.243)		
		Bias	0.074 (−0.014,0.162)	0.233 (0.141,0.324)		
Gamma(.6,1)	−0.45	Power	0.810 (0.786,0.834)	0.798 (0.773,0.823)	0.805 (0.780,0.830)	0.801 (0.776,0.826)
		Coverage	0.946 (0.932,0.960)	0.972 (0.962,0.982)		
		MSE	0.191 (0.172,0.210)	0.215 (0.194,0.237)		
		Bias	0.108 (0.023,0.194)	0.263 (0.173,0.352)		

Distribution	β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
T_3	0.00[b]	Power	0.052 (0.047,0.057)	0.052 (0.047,0.058)	0.050 (0.045,0.055)	0.052 (0.047,0.057)
		Coverage	0.948 (0.943,0.953)	0.964 (0.960,0.968)		
		MSE	0.142 (0.137,0.146)	0.161 (0.156,0.167)		
		Bias	0.057 (0.029,0.085)	0.061 (0.031,0.090)		
T_3	−0.21	Power	0.257 (0.230,0.284)	0.246 0.219,0.273)	0.253 (0.226,0.280)	0.255 (0.228,0.282)
		Coverage	0.944 (0.930,0.958)	0.963 (0.951,0.975)		
		MSE	0.135 (0.123,0.147)	0.157 (0.143,0.170))		
		Bias	0.090 (0.018,0.162)	0.138 (0.061,0.215)		
T_3	−0.45	Power	0.822 (0.798,0.846)	0.793 (0.768,0.818)	0.820 (0.796,0.844)	0.812 (0.788,0.836)
		Coverage	0.947 (0.933,0.961)	0.966 (0.955,0.977)		
		MSE	0.128 (0.116,0.139)	0.151 (0.138,0.164)		
		Bias	0.012 (−0.058,0.083)	0.149 (0.073,0.224)		
Uniform	0.00[b]	Power	0.049 (0.044,0.054)	0.075 (0.068,0.081)	0.048 (0.043,0.053)	0.048 (0.043,0.053)
		Coverage	0.939 (0.933,0.944)	0.955 (0.950,0.960)		
		MSE	0.161 (0.155,0.166)	0.152 (0.147,0.157)		
		Bias	−0.404 (−0.432,− 0.376)	−0.333 (−0.361,− 0.305)		
Uniform	−0.21	Power	0.378 (0.348,0.408)	0.388 (0.358,0.418)	0.372 (0.342,0.402)	0.361 (0.331,0.391)
		Coverage	0.946 (0.932,0.960)	0.962 (0.950,0.974)		
		MSE	0.148 (0.135,0.161)	0.145 (0.133,0.158)		
		Bias	−0.419 (−0.490,− 0.348)	−0.385 (−0.456,− 0.314)		
Uniform	−0.45	Power	0.957 (0.944,0.970)	0.958 (0.946,0.970)	0.958 (0.946,0.970)	0.948 (0.934,0.962)
		Coverage	0.949 (0.935,0.963)	0.973 (0.963,0.983)		
		MSE	0.128 (0.116,0.139)	0.126 (0.114,0.137)		
		Bias	−0.233 (−0.301,− 0.164)	−0.194 (−0.263,− 0.126)		
Positive Stable(.5)	0.00[b]	Power	0.048 (0.043,0.053)	0.049 (0.044,0.054)	0.047 (0.042,0.052)	0.049 (0.044,0.054)
		Coverage	0.952 (0.947,0.957)	0.966 (0.962,0.970)		
		MSE	0.161 (0.155,0.166)	0.187 (0.181,0.193)		

Contemp Clin Trials. Author manuscript; available in PMC 2013 December 04.

Distribution	β_1		MPHM Robust SE	NLMM	Clustered Logrank	Clustered Permutation Test
		Bias	−0.060 (−0.089,0.030)	−0.057 (−0.089,−0.025)		
Positive Stable(.5)	−0.21	Power	0.145 (0.123,0.167)	0.148 (0.126,0.170)	0.145 (0.123,0.167)	0.141 (0.119,0.163)
		Coverage	0.950 (0.936,0.964)	0.965 (0.954,0.976)		
		MSE	0.153 (0.140,0.166)	0.173 (0.158,0.188)		
		Bias	−0.055 (−0.132,0.021)	0.059 (−0.023,0.141)		
Positive Stable(.5)	−0.45	Power	0.491 (0.460,0.522)	0.517 (0.486,0.548)	0.491 (0.460,0.522)	0.478 (0.447,0.509)
		Coverage	0.948 (0.934,0.962)	0.956 (0.943,0.969)		
		MSE	0.147 (0.135,0.160)	0.171 (0.156,0.186)		
		Bias	−0.047 (−0.122,0.028)	0.165 (0.085,0.246)		

[a] Estimates have been rescaled by a factor of 10

[b] Estimates of Type I error are based on 7,000 simulations

Contemp Clin Trials. Author manuscript; available in PMC 2013 December 04.

Table 5

Actual Results from PACE and HORIZON Studies by Study and Method

Description	PACE	HORIZON
Number of Clusters	414	434
Cluster Size	2–65	1–148
ICC	0.04	0.03
% Censored	10%	15%
mean follow-up time	381.45 days	263.53 days
MPHM	p=0.59, HR=1.06 95%CI:(0.87, 1.28)	p=.02, HR=1.48 95%CI:(1.05, 2.07)
NLMM	p=0.47, HR=1.11 95%CI:(0.84, 1.46)	p=.14, HR=1.21 95%CI:(.90, 1.51)
Clustered Logrank	p=0.59	p=.04
Permutation Test	p=0.52	p=.02

Table 6

Hazard ratio for conditional hazard ratio parameters

Conditional β_1	Conditional HR
0.00	1
−0.04	1.04
−0.09	1.09
−0.18	1.18
−0.21	1.21
−0.45	1.51
−0.72	1.92