HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2023 September 24.

Published in final edited form as:

Stat Methods Med Res. 2022 July ; 31(7): 1224–1241. doi:10.1177/09622802221085080.

A comparison of analytical strategies for cluster randomized trials with survival outcomes in the presence of competing risks

Fan Li[1,2,3], **Wenhan Lu**[1,*], **Yuxuan Wang**[1,*], **Zehua Pan**[1,*], **Erich J Greene**[1,2], **Guanqun Meng**[1], **Can Meng**[1,2], **Ondrej Blaha**[1,2], **Yize Zhao**[1,2], **Peter Peduzzi**[1,2], **Denise Esserman**[1,2]

[1]Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA

[2]Yale Center for Analytical Sciences, New Haven, CT, USA

[3]Center for Methods in Implementation and Prevention Science, Yale University School of Public Health, New Haven, CT, USA

Abstract

While statistical methods for analyzing cluster randomized trials with continuous and binary outcomes have been extensively studied and compared, little comparative evidence has been provided for analyzing cluster randomized trials with survival outcomes in the presence of competing risks. Motivated by the Strategies to Reduce Injuries and Develop Confidence in Elders trial, we carried out a simulation study to compare the operating characteristics of several existing population-averaged survival models, including the marginal Cox, marginal Fine and Gray, and marginal multi-state models. For each model, we found that adjusting for the intraclass correlations through the sandwich variance estimator effectively maintained the type I error rate when the number of clusters is large. With no more than 30 clusters, however, the sandwich variance estimator can exhibit notable negative bias, and a permutation test provides better control of type I error inflation. Under the alternative, the power for each model is differentially affected by two types of intraclass correlations — the within-individual and between-individual correlations. Furthermore, the marginal Fine and Gray model occasionally leads to higher power than the marginal Cox model or the marginal multi-state model, especially when the competing event rate is high. Finally, we provide an illustrative analysis of Strategies to Reduce Injuries and Develop Confidence in Elders trial using each analytical strategy considered.

Keywords

Multivariate survival analysis; competing risks; time-to-event outcomes; Fine and Gray model; sandwich variance estimator; permutation test

Corresponding author: Fan Li, Department of Biostatistics, Yale School of Public Health, New Haven, CT 065I I, USA. fan.f.li@yale.edu.

*The authors Wenhan Lu, Yuxuan Wang, and Zehua Pan contributed equally to this work.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Supplemental Material

Supplementary material for this article is available online.

1. Introduction

Cluster randomized trials (CRTs) are studies in which an intervention is delivered at the cluster level.[1] Common reasons for cluster-level randomization include administrative and logistical convenience, prevention of treatment contamination and ethical reasons. Statistical methods for the design and analysis of CRTs have been discussed extensively for decades.[2,3] As the popularity of adopting pragmatic clinical trials increases due to their ability to more closely resemble health care practice, there is a greater need for rigorous statistical methodology applicable to CRTs with more complex data structures. For example, these complexities could arise due to multiple levels of clustering, such as when participants are nested within physicians nested within practices. Further, participants can experience a terminal event such as death, which could compete with non-terminal events of interest (e.g. injuries, asthma attacks, stroke).[4] While clustered time-to-event outcomes present critical challenges due to censoring and competing risks, there are limited empirical evaluations of analytical strategies in the context of CRTs for such outcomes (an exception is Stedman et al.[5] but without competing risks). The complexities of clustered time-to-event data require careful consideration in order to ensure that models adequately reflect the data structure and that we avoid erroneous conclusions.

Motivated by the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) trial,[6–8] this article aims to provide an empirical comparison between several existing survival models for analyzing CRTs with time-to-event outcomes in the presence of a competing event. Briefly, the STRIDE study is a Patient-Centered Outcomes Research Institute and National Institute on Aging funded pragmatic, parallel CRT focusing on reducing serious fall injuries in community dwelling older adults at risk of falls. STRIDE tests the effectiveness of an evidence-based, multi-factorial, fallrelated injury prevention strategy compared to enhanced usual care. Between 2015 and 2017, the study enrolled 5451 adults aged 70 years and older from 86 primary care practices. These practices were randomized to either intervention or control. While the primary outcome was time to first serious fall-related injury,[8] participants could pass away before observing the primary outcome, leading to possible dependent censoring.

A recent systematic review indicated that CRTs with time-to-event or survival outcomes are not uncommon.[9] There is also a growing statistical literature in designing CRTs with such outcomes.[10–14] However, these articles have only addressed issues related to sample size and power calculations in the design stage and have not yet compared the operating characteristics of models used in the analysis stage. Besides, none of these articles have considered extensions to competing risks, and therefore the implications due to competing events in CRTs are not immediately clear. For the analysis of the STRIDE study, it was important to determine the best analytical approach to estimate the intervention effect while taking into account the competing event. This consideration motivated us to design a simulation study to identify an accessible and reliable analytical approach that maintains the nominal type I error rate and has the highest power to detect the treatment effect. Particularly, we vary the within-individual correlation (measuring the dependence between two latent event times of different causes for the same individual) and the between-individual correlation (measuring the dependence between two latent event times of the

same cause for two different individuals in the same cluster) and assess their impact on type I error rate and power. The relationship between study power and intraclass correlations (including both within-individual correlation and between-individual correlation) can also help characterize the magnitude of variance inflation due to clustering, which is not available in closed forms in this complex setting but is indispensable for accurate sample size determination and study monitoring purposes.

There are two mainstream regression frameworks to analyze CRTs with time-to-event outcomes: frailty models which account for clustering via random effects (cluster-specific models) and marginal models that adjust for clustering via the robust sandwich variance estimator (population-averaged models)[15–17] While each approach has its pros and cons, we focus on the marginal models for its straightforward population-averaged interpretation. The merits and limitations of the population-averaged approach have been discussed in detail by Preisser et al.[18] in parallel CRTs and Li et al.[19] in longitudinal CRTs. The marginal model separately specifies the marginal mean and working correlation structures, has a straightforward interpretation of the treatment effect parameter that does not depend on any unobserved variables, and has been shown to be robust to working correlation model misspecification.

The remainder of this article is organized as follows. In Section 2, we provide a brief overview of the survival models considered for the simulation study. Section 3 provides a description of the simulations and the methods to generate the complex clustered survival data. The simulation results are presented in Section 4. We provide an illustrative analysis of the STRIDE trial using these different analytical strategies in Section 5, and Section 6 concludes with a discussion.

2. Statistical methods

2.1 Overview

In this section, we briefly review survival models for analyzing time-to-event outcomes. In the motivating STRIDE study, death is technically a semi-competing event because it is terminal (while the outcome of interest, fall-related injury, is non-terminal). However, because we are interested in the time to first event, we treat death as a competing risk throughout this study. We consider four categories of models (see Table 1 for a summary). These include methods that: (1) do not account for clustering and censor the competing event (Section 2.2); (2) account for clustering but censor the competing event (Section 2.3); (3) account for the competing risk of death but do not account for clustering (Section 2.4); and (4) account for both clustering and the competing risk of death (Section 2.5). As we focus on applications to CRTs, we primarily focus on methods (2) and (4), and consider methods (1) and (3) as reference approaches. These approaches can be implemented using readily available packages in R (https://cran.r-project.org/). A summary of the packages (although not comprehensive) is also provided in Table 1 and sample R code for implementing the models is in Web Appendix A. In Section 2.6, we additionally discuss methods to carrying out permutation tests for these models to generate more robust small-sample inference in CRTs. R code for implementing the survival models and the permutation test is also available at the public GitHub Repository (https://github.com/kyleyxw/simCRTs).

2.2 Cox model that does not account for clustering and censors the competing event

For analyzing time-to-event outcomes, the Cox proportional hazards model is the most basic regression model in that it does not account for clustering and censors any competing event. Since this method ignores the clustering structure, we only consider it as a reference approach in the simulations to quantify the amount of variance inflation due to intraclass correlations in CRTs. For notation purposes, here we use a single subscript i to denote each individual and define N as the total number of individuals in the study. Writing T_i and C_i as the failure and censoring times, we define $X_i = \min\{T_i, C_i\}$ as the observed time to experience the event of interest and $\Delta_i = I(T_i \leq C_i)$ as the censoring indicator. In the competing risks context, this approach estimates the cause-specific hazard for the event of interest, and therefore the censoring time C_i is defined as the earliest time to experience the competing event of death, loss to follow-up or the end of study (administrative censoring).

The Cox model specifies the hazard for individual i as

$$\lambda_i(t \mid Z_i) = \lambda_0(t)\exp(\beta' Z_i) \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and Z_i is the $p \times 1$ design vector. Frequently in CRTs, the design vector Z_i includes only the treatment indicator ($p = 1$), and the exponentiated regression coefficient is interpreted as the population-averaged, cause-specific hazard ratio (HR_{CS}). HR_{CS} corresponds to the relative instantaneous hazard rate due to treatment among any individuals who survive all events up to any given time t. Although it is a valid measure of the *apparent* treatment effect, HR_{CS} does not necessarily translate into a measure of risk (defined by the cumulative incidence function), without assuming the independence between the competing events.[20] Alternatively, Z_i could include a list of pre-specified baseline covariates believed to affect the failure time, in which case the exponentiated coefficient corresponding to the treatment is interpreted as the adjusted HR_{CS}.

The estimation of β model (1) proceeds by maximizing the partial likelihood, or equivalently, by solving the partial score equations defined as

$$U\left(\beta\right) = \sum_{i=1}^{N} \Delta_i \left\{ Z_i - \frac{\sum_{s=1}^{N} Y_s(X_i)\exp(\beta' Z_s) Z_s}{\sum_{s=1}^{N} Y_s(X_i)\exp(\beta' Z_s)} \right\} = 0 \tag{2}$$

where $Y_i(t) = I(X_i \geq t)$ is the at-risk process. Because the implementation of this Cox model does not account for clustering, the variance estimator of $\hat{\beta}$ is simply obtained from inverting the information matrix, $\hat{A}\left(\beta\right) = -N^{-1}\partial U\left(\beta\right)/\partial\beta'$. In other words, we can estimate the standard error (SE) of $\hat{\beta}$ by taking the square root of the appropriate element in $\hat{A}^{-1}\left(\hat{\beta}\right)$ and define the Wald statistic as $\hat{\beta}/\mathrm{se}\left(\hat{\beta}\right)$. Under the null hypothesis of no treatment effect, the Wald statistic approximately follows a standard normal distribution. Whereas the above model-based variance estimator is consistent in the absence of clustering, it may underestimate the true variability of $\hat{\beta}$ in the presence of clustering.[17,16] On the one hand, this underestimation of the variability inflates the type I error rate and leads to incorrect conclusions about the treatment effect. On the other hand, the degree of variance

underestimation is directly associated with the degree of variance inflation in CRTs due to clustering, which is of interest for study planning purposes.

2.3 Marginal Cox model that accounts for clustering but censors the competing event

In CRTs, the marginal Cox model can be used to account for clustering through the use of the robust sandwich variance estimator.[16,17] Suppose we have $k = 1, \ldots, K$ clusters, each with m_k patients, then the total sample size is $N = \sum_{k=1}^{K} m_k$. We now write T_{kj} and C_{kj} as the failure and censoring times for individual j in cluster k (here the censoring time is defined the exact same way as in Section 2.2). We define $X_{kj} = \min\{T_{kj}, C_{kj}\}$ as the observed time to experience the event of interest and $\Delta_{kj} = I(T_{kj} \leq C_{kj})$ as the censoring indicator. The marginal Cox model specifies the hazard as

$$\lambda_{kj}(t \mid Z_{kj}) = \lambda_0(t)\exp(\beta' Z_{kj}), \tag{3}$$

where $\lambda_0(t)$ is an unspecified baseline hazard and Z_{kj} is a $p \times 1$ design vector for individual j in cluster k including the cluster-level treatment indicator. In fact, model (3) and model (1) are equivalent, except model (3) explicitly acknowledges the multilevel structure of the data with double subscripts. Importantly, the estimation of the marginal Cox model proceeds by solving the same partial score equation (2). Therefore, the point estimates are no different between model (3) and model (1) in CRTs and the interpretation of regression coefficients is the same between these two models (provided the same set of treatment and covariates are used).

Despite the identical point estimates, the variance of $\hat{\beta}$ from the marginal Cox model is estimated by a robust sandwich estimator. The sandwich variance estimator has been studied extensively in generalized linear models and GEE with noncensored outcomes[21] and was extended to the marginal Cox model.[16] The key idea is to regard the partial score equation (2) as an estimating equation with an independence working correlation matrix. Based on the theory of martingale estimating equations, Wei et al.[16] and later Spiekerman and Lin[22] proved that a valid variance estimator that properly accounts for clustering is $\hat{V}\left(\hat{\beta}\right) = \hat{A}^{-1}\left(\hat{\beta}\right)\hat{B}\left(\hat{\beta}\right)\hat{A}^{-1}\left(\hat{\beta}\right)$, where $\hat{B}\left(\hat{\beta}\right)$ is an empirical covariance estimator of the cluster-specific partial score. By adjusting for clustering through $\hat{B}\left(\hat{\beta}\right)$, the sandwich variance estimator $\hat{V}\left(\hat{\beta}\right)$ tends to be larger than the model-based variance estimator $\hat{A}^{-1}\left(\hat{\beta}\right)$ and reduces the tendency for the Wald test to incorrectly reject the null. In Web Appendix B, we provide the explicit forms of $\hat{A}\left(\hat{\beta}\right)$ and $\hat{B}\left(\hat{\beta}\right)$ and explain how the sandwich variance estimator leads to robust inference in CRTs. Furthermore, as alluded to in Section 2.2, the differences between $\hat{V}\left(\hat{\beta}\right)$ and $\hat{A}^{-1}\left(\hat{\beta}\right)$ can be quantified by the variance inflation factor, which is often of interest for design and monitoring purposes. As in Section 2.2, the marginal Cox model also equates the competing event with censoring and therefore the exponentiated coefficient is interpreted as HR_{CS}. This approach can create a violation of the fundamental assumption of independence between the time-to-event distribution and censoring distribution and run the risk of overestimating the cumulative incidence function.[23] Nevertheless, we include this approach in our simulations because it's a common method used in practice, and there

remains interest in assessing its robustness for analyzing CRTs, where, for instance, the competing event is possibly not dominant.

2.4 Models that do not account for clustering but address competing events

2.4.1 Fine and Gray model—The Fine and Gray model is a semi-parametric model that accounts for competing events by directly modeling the cumulative incidence function, or the so-called sub-distribution function.[4] Here, we still maintain T_i as the failure time, but additionally define $\varepsilon_i \in \{1, \ldots, l\}$ to be the causes of failure, where $l \geq 2$ causes are assumed to be observable. The right censoring time due to loss to follow-up or end of study is now defined by C_i. Furthermore, $X_i = \min\{T_i, C_i\}$ is still the observed survival time. Unlike the Cox model in Section 2.2, the Fine and Gray model does not censor the competing event but instead considers the sub-distribution hazard function for the event of interest (cause $\varepsilon = 1$)

$$\lambda^{\text{sub}}\left(t \mid Z\right) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\left\{t \leq T \leq t + \Delta t, \varepsilon = 1 \mid T \geq t \cup \left(T \leq t \cap \varepsilon \neq 1\right), Z\right\} = -d\log\left\{1 - F\left(t \mid Z\right)\right\}/dt, \tag{4}$$

where $F\left(t \mid Z\right) = P\left(T \leq t, \varepsilon = 1\right) = 1 - \exp\left\{\int_0^t \lambda^{\text{sub}}\left(s \mid Z\right) ds\right\}$ is the (sub-distribution) cumulative incidence function.

The Fine and Gray model is written as

$$\lambda_i^{\text{sub}}(t \mid Z_i) = \lambda_0^{\text{sub}}(t)\exp(\gamma' Z_i), \tag{5}$$

where $\lambda_0^{\text{sub}}(t)$ is an unspecified baseline sub-distribution hazard. Similar to Section 2.2, the design vector Z_i could include only the treatment indicator, in which case the exponentiated regression coefficient corresponds to the sub-distribution hazard ratio (HR_{sub}) due to the treatment. By the definition of (4), HR_{sub} accounts for the competing risk by actively maintaining the individuals experiencing the competing event in the risk set at time t. It is an effect measure that is due to both the treatment effect on the event of interest and the potentially differential impact of the competing event on the risk set in the population.[20] Due to the direct correspondence between sub-distribution hazard and cumulative incidence function, HR_{sub} directly translates into a measure of risk and describes the treatment effect on the cumulative incidence, which is considered as a major difference from HR_{cs}.

To estimate γ, Fine and Gray[4] modified the partial score equations of the Cox model. Let $N_i(t) = I(T_i \leq t, \varepsilon_i = 1)$ be the counting process for the event of interest and let $Y_i(t) = 1 - N_i(t-)$ be the at-risk process; $\hat{\gamma}$ is found by solving

$$U^{\text{sub}}\left(\gamma\right) = \sum_{i=1}^{N} I(\varepsilon_i = 1)\left\{Z_i - \frac{\sum_{s=1}^{N} w_s(X_i)Y_s(X_i)Z_s\exp(\gamma' Z_s)}{\sum_{s=1}^{N} w_s(X_i)Y_s(X_i)\exp(\gamma' Z_s)}\right\} w_i(X_i) = 0, \tag{6}$$

where $w_i\left(t\right) = I(C_i \geq \min\{T_i, t\})\hat{G}\left(t\right)/\hat{G}(\min\{X_i, t\})$ is the time-dependent inverse probability of censoring weight, and $\hat{G}\left(\cdot\right)$ is the Kaplan-Meier estimate of the survival function of the

censoring time. In addition, it is important to notice that $U^{\text{sub}}\left(\gamma\right)$ does not correspond to an actual likelihood, and therefore Fine and Gray provide a sandwich variance for inference and hypothesis testing. The sandwich variance estimator takes the form $\hat{\Omega}^{-1}\hat{\Sigma}\hat{\Omega}^{-1}$, where Ω^{-1} is the probability limit of the inverse of the partial derivative of $U^{\text{sub}}\left(\gamma\right)$ evaluated at the true parameter value γ_0 and Σ is the asymptotic variance of $N^{-1/2}U^{\text{sub}}(\gamma_0)$, $\hat{\Omega}^{-1}$ and $\hat{\Sigma}$ are consistent estimators based on sample averages.[4] The Wald normality-based test for $H_0: \gamma = 0$ could be defined analogous to Section 2.2 once the variance of $\hat{\gamma}$ is obtained from the sandwich variance estimator.

2.4.2 Multi-state model—The multi-state model is an alternative approach that addresses competing risks by formulating different transition intensities (hazards) to model the transitions between different states.[24] If we generically denote T as the time of reaching state h from state q, the hazard rate for the transition from state h to state q has the general form,

$$\lambda_{hq}\left(t\right) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{7}$$

Considering our context of the STRIDE trial, the multi-state model examined in this study becomes a unidirectional illness-death model (Figure 1),[23] with the fall as a *transient* state and death as the *absorbing* state. Define the transition probabilities in this model by $P_{hq}(r,t) = P$ (in state q at time t lin state h at time r) with $r < t$ and $h, q \in \{0,1,2\}$, and state 0,1 , and 2 represent healthy, fall, and death, respectively. These transition probabilities can be written in explicit terms as in Putter et al.,[23] which can be used for individual predictions. Assuming a Markov model with a clock-forward time scale, we consider the Cox proportional hazards model for each of the transition hazards separately. In STRIDE, we are primarily interested in the treatment effect on the transition hazard $\lambda_{01}(t)$, modeled by

$$\lambda_{01,i}(t \mid Z_i) = \lambda_{01,0}(t)\exp(\beta' Z_i), \tag{8}$$

where Z_i is the design vector defined previously. This specification leads to partial likelihood and score functions similar to equation (2) for the Cox model; the transition hazard ratio, $\exp(\beta)$, bears a similar interpretation to HR_{CS} as it is specific to the transition between the healthy and fall states. In the STRIDE trial, because the initial state for all participants is the healthy state, the transition-specific hazard ratio is indeed equal to HR_{CS}. Finally, similar to Section 2.2, we consider the Fine and Gray and multi-state model without clustering only as a reference approach in our simulations to quantify the amount of variance inflation due to intraclass correlations in CRTs. The contrast between methods with and without clustering helps clarify when clustering matters for valid inference and can demonstrate the amount of required sample size inflation in CRTs with complex survival outcomes under various combinations of parameters.

2.5 Models that account for clustering and address competing events

2.5.1 Marginal Fine and Gray model—The marginal Fine and Gray sub-distribution hazard model is an extension of the Fine and Gray model that properly reflects the multilevel structure of the data and accounts for clustering through a modified sandwich variance estimator.[25] Similar to Section 2.3, we assume there are K clusters, each with cluster size m_k. In addition to T_{kj}, C_{kj}, and Z_{kj}, we define $\varepsilon_{kj} \in \{1, \ldots, l\}$ to be the causes of failure, where $l \geq 2$ causes are assumed to be observable. The observed data then consist of $\{X_{kj} = \min\{T_{kj}, C_{kj}\}, I(T_{kj} \leq C_{kj})\varepsilon_{kj}, Z_{kj}\}$ and the marginal sub-distribution hazard model is

$$\lambda_{kj}^{\text{sub}}(t \mid Z_{kj}) = \lambda_0^{\text{sub}}(t)\exp(\gamma' Z_{kj}), \tag{9}$$

where $\lambda_0^{\text{sub}}(t)$ is the unspecified baseline sub-distribution hazard. When Z_{kj} only includes the intervention indicator, $\exp(\gamma)$ is interpreted as the population-averaged HR_{sub}. Zhou et al.[25] developed an estimating equations approach with an independence working assumption for estimating γ. In fact, the independence estimating equation reduces to equation (6) once we replace the two-level notations (subscripts , j) with the one-level notation (subscript i), and therefore the point estimate $\hat{\gamma}$ is equal to that obtained from the Fine and Gray model. To properly account for clustering, Zhou et al.[25] proposed a clustered sandwich variance estimator $\hat{\Omega}^{-1}\hat{\Lambda}\hat{\Omega}^{-1}$. While $\hat{\Omega}^{-1}$ is defined similarly as in Section 2.4.1, the matrix $\hat{\Lambda}$ is the empirical covariance of the cluster-specific contribution to the estimating equations (rather than the variance estimate $\hat{\Sigma}$ of the Fine and Gray model). With a sufficient number of clusters, the clustered sandwich variance estimator for $\hat{\gamma}$ is consistent even though the independence working assumption is used. The Wald test for $H_0: \gamma = 0$ can be obtained once $\hat{\gamma}$ and its sandwich SE are computed, and the corresponding marginal cumulative incidence function can be obtained based on a Breslow-type estimator for $\lambda_0^{\text{sub}}(t)$. Analogous to the comparison between Cox and marginal Cox models, the comparison between Fine and Gray and marginal Fine and Gray model can shed light on the degree of variance inflation in CRTs due to clustering, which is of interest for study planning purposes when the primary analysis considers cumulative incidence regression. Finally, while Zhou et al.[25] have performed simulations to examine the performance of the clustered sandwich variance estimator with at least 100 clusters, most CRTs have much fewer than 100 clusters and the model performance with a smaller number of clusters is currently unclear.

2.5.2 Marginal multi-state model—We also consider the marginal multi-state model[23], which is similar to the marginal Cox model presented in Section 2.3, that estimates the regression parameters using an independence working assumption adjusting for clustering via the clustered sandwich variance estimator. Since an independence working assumption is considered, the point estimate from the marginal multi-state model is same as that from the conventional multi-state model. The sandwich variance is obtained in a similar fashion as those described in Sections 2.3 and 2.5.1.

2.6 Permutation test

An important objective in analyzing CRTs is to make inference on the treatment effect parameter. While the normality-based Wald test in each of the above models may carry

the nominal size when there is a large number of clusters (say, 100), it may not guarantee adequate control of type I error rate when the number of clusters is small. Although the STRIDE study does not suffer from a limited number of clusters ($K = 86$), systematic reviews by Fiero et al.,[26] Murray et al.,[27] and Ivers et al.[28] suggest that more than half of the published CRTs included no more than 30 clusters, therefore, considerations on improving the small-sample performance of the Wald test are particularly relevant.

For testing the null hypothesis of no treatment effect, an alternative to the Wald-test is a permutation test. Because of its robust control of type I error rate in small samples, the permutation test has received considerable attention in the analyses of CRTs with non-censored outcomes (see, e.g. Gail et al.[29] and Li et al.[30,31] with continuous and binary outcomes). Cai and Shen[32] developed a non-parametric permutation test to compare the marginal survival functions; the test statistic was chosen to be the generalized linear rank and Renyi-type test statistics designed to detect the maximal deviation of survival functions across time. In addition, Wang and De Gruttola[33] developed a permutation test where the test statistic was the weighted average of treatment effect estimates between all pairs of clusters. While both approaches demonstrated adequate control of type I error rate and power in CRTs, they have not examined the use of the permutation test in analyzing CRTs with competing risks. To examine whether the permutation test could improve the performance of the Wald-test (especially when the number of clusters do not exceed 30), we developed permutation tests corresponding to each model as follows.

Define $\mathcal{D} = (\mathcal{X}, \mathcal{E}, \mathcal{Z})$ as the data matrix consisting of the collection of observed time-to-event outcomes $(\mathcal{X}, \mathcal{E})$ (where $\mathcal{X}$ is the set of all observed event times, and $\mathcal{E}$ denotes the set of all observed causes) and the treatment assignment $\mathcal{Z}$. Typically, because $K/2$ clusters are randomized to each arm, there are $S = \binom{K}{K/2}$ possible permutations of the treatment labels. Define $\mathcal{Z}^s$ as a permutation of $\mathcal{Z}$ in the randomization space and $\mathcal{Z}^*$ as the realized randomization scheme in the trial. We can write $\mathcal{D}(\mathcal{Z}^s) = (\mathcal{X}, \mathcal{E}, \mathcal{Z}^s)$ as the data matrix under permutation of treatment $\mathcal{Z}^s$. Under the strong null hypothesis of no treatment effect (treatment has no effect on the hazard or sub-distribution hazard of any individual), one can show that the realized data matrix $\mathcal{D}(\mathcal{Z}^*)$ can be regarded as a randomly selected element from the set $\mathbb{D} = \{\mathcal{D}(\mathcal{Z}^s): s = 1, \ldots, S\}$ consisting of S permuted data matrices. We then define a test statistic $\mathcal{T} = \mathcal{T}(\mathcal{D})$, and it holds that the observed test statistic $\mathcal{T}^* = \mathcal{T}(\mathcal{D}(\mathcal{Z}^*))$ is a random sample from the permutation distribution $\{\mathcal{T}(\mathcal{D}(\mathcal{Z}^s)): \mathcal{D}(\mathcal{Z}^s) \in \mathbb{D}\}$. As long as the number of permutations S is not unrealistically small, this test is guaranteed to maintain the nominal type I error rate under the strong null, even if there are a small number of clusters.[29,30,33]

To operationalize the permutation test for each model, we define two different test statistics. The permutation β-test uses the estimated treatment effect from a given model as the test statistic $\mathcal{T}$, while the permutation z-test statistic is chosen to be the z-score, or the Wald-test statistic based on the sandwich variance estimate. For example, the following steps are required to implement a permutation test under the marginal Fine and Gray model:

a. estimate the test statistic $\mathcal{T}^*$, either $\mathcal{T}^* = \hat{\gamma}$ or $\mathcal{T}^* = \hat{\gamma}/\sqrt{\text{var}(\hat{\gamma})}$, from the observed data;

b. obtain the permutation distribution of $\mathcal{T}^*$ by repeating step (a) with $\widetilde{S} \leq S$ randomly permuted, distinct data sets contained in $\mathbb{D}$; the value of $\widetilde{S}$ is sometimes chosen for computational considerations;

c. reject the null hypothesis if $\mathcal{T}^*$ lands in the rejection region of the discrete permutation distribution.

As an alternative to step (c), one could also estimate the permutation p-value as the proportion of the permuted test statistics that are equal or more extreme in absolute value than the observed test statistic, and compare with the nominal level. The permutation tests for the marginal Cox and multi-state models can be analogously implemented following the above three steps.

3 Simulation design

We conducted a series of simulations to evaluate the operating characteristics of the three types of methods with data generated to include a clustered survival outcome and competing event. Mimicking the structure of the STRIDE trial, we assumed a two-arm CRT with equal randomization. We considered three levels of total number of clusters $K \in \{10,30,100\}$. We used $K = 100$ to represent a sample size similar to STRIDE, and used $K = 30$ and $K = 10$ to represent a moderate and small numbers of clusters, conditions under which the robust sandwich variance estimator may inflate the type I error rate for testing the cluster-level intervention effect.[34] Finally, to resemble the STRIDE study, the cluster sizes m_k were sampled from the empirical distribution of the cluster sizes observed in the STRIDE study. The mean cluster size was 63 (ranges from 10 to 199) and the coefficient of variation was 0.53 .

3.1 Data generation

The marginal survival function of the event time $S_{1,kj}$ follows a Uniform (0, 1) distribution, from which we generated $S_{1,kj}$ for each individual j in cluster k. The Gumbel copula function[35] was then used to generate a second survival probability for the same individual, $S_{2,kj}$, representative of the competing event (death) with the copula association parameter controlling for the correlation between the event time of interest and the competing event time. Under the Gumbel copula, there is a one-to-one mapping between the association parameter and Kendall's tau (τ_w),[36] which we varied to represent different levels of the within-individual correlation between the two (latent) event times. To induce the between-individual correlation τ_b, we considered a frailty $G_k \sim \text{Gamma}(\text{shape} = a, \text{rate} = a)$ with $a = 0.5(1/\tau_b - 1)$, and generated the latent event times

$$T_{1,kj} = \frac{-\log(S_{1,kj})}{\lambda_1 \exp(\delta Z_k) G_k}, \quad T_{2,kj} = \frac{-\log(S_{2,kj})}{\lambda_2 G_k}, \tag{10}$$

where $T_{1,kj}$ is the latent time for the event of interest, $T_{2,kj}$ is the latent time for the competing event, λ_1 and λ_2 are the assumed constant baseline hazards (event rates) for these two

survival times, Z_k is the cluster-level treatment, and $\exp(\delta)$ is the latent hazard ratio for $T_{1,kj}$. Additional details on the data generating process are given in Web Appendix C. Web Appendix C also provides a derivation to show τ_b corresponds to the Kendall's tau, or the between-individual correlation for two event times of the same cause. In our data generating process, we require τ_b to be strictly between 0 and 1 to ensure the frailty distribution is well-defined.

We simulated the censoring time C_{kj} (time to censoring for reasons other than death) from a uniform distribution with the assumed censoring rate. The comparison of these three latent event times determined the final status of each individual. When the observed time $X_{kj} = \min\{T_{1,kj}, T_{2,kj}, C_{kj}\} = C_{kj}$, the individual was considered censored in the usual sense and $\varepsilon_{kj} = 0$. When $X_{kj} = \min\{T_{1,kj}, T_{2,kj}, C_{kj}\} = T_{2,kj}$, this individual was not censored in the usual sense, though the event time of interest was not observed since death happened earlier; in this case, $\varepsilon_{kj} = 2$. Finally, when $X_{kj} = \min\{T_{1,kj}, T_{2,kj}, C_{kj}\} = T_{1,kj}$, the event of interest was observed for this individual, and $\varepsilon_{kj} = 1$. For those individuals who experienced the event of interest, we assumed they could also be subject to the competing event or censoring even after time X_{kj}. For simplicity, we assumed that an individual had the same hazard for the competing event after experiencing the event of interest; that is, the survival function for the latent time $T_{2,kj}$ remained unaffected after conditioning on $\{T_{2,kj} \geq X_{kj} = T_{1,kj}\}$. Based on the simulated observed event time and event status, we formatted the data in two forms: a long format considering the competing risks after the event (used for the multi-state model) and a wide format considering only the first event (used for the Cox and Fine and Gray models).

3.2 Parameter configurations

We varied key parameters in the above data generating process to represent a range of scenarios that resemble the STRIDE study. We fixed the baseline hazard rate $\lambda_1 = 0.08$ and the dropout rate (used to specify the censoring distribution) at 0.03 . We varied the baseline hazard rate for death, $\lambda_2 \in \{0.02, 0.04, 0.08, 0.12\}$, representing scenarios in which a smaller to larger fraction of patients died before the event of interest could be observed. Because the clustering of event times plays an important role in the design and analysis of such trials, we varied the copula Kendall's tau, $\tau_w \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5\}$, and the frailty Kendall's tau, $\tau_b \in \{0.001, 0.01, 0.05, 0.1, 0.3\}$. In what follows, the copula Kendall's tau is referred to as the within-individual correlation, and the frailty Kendall's tau is referred to as the between-individual correlation. Under the null scenario, we set the latent hazard ratio $\exp(\delta) = 1$ (see equation (10) for definition of δ), and under the alternative, we used $\exp(\delta) \in \{0.5, 0.8, 2\}$. The value of 0.8 was the hypothesized effect size for designing the STRIDE study. We additionally included $\exp(\delta) = 0.5$ and 2 to evaluate whether the results are sensitive to the magnitude and direction of the effect size. In summary, our simulation used a factorial design consisting of four values of the latent hazard ratio, four levels of the baseline hazard rate for the competing event, five values of the between-individual correlation, seven values of the within-individual correlation, and three different number of clusters (sample sizes). In total, we have evaluated $4 \times 4 \times 5 \times 7 \times 3 = 1680$ scenarios.

For each scenario, we conducted 1000 simulations to investigate the performance of the three types of survival models (with clustered sandwich variance estimator) and the following performance metrics were considered: relative bias (RB) for estimating the treatment effect, calculated as the relative difference between the mean estimated coefficient and the true target parameter; SE ratio (SER), calculated as the ratio between the true Monte Carlo SE and the mean estimated SE; coverage probability (CP), calculated as the proportion of 95% confidence intervals for the treatment coefficient that contained the true value; empirical type I error rate, calculated as the proportion of tests that resulted in a rejection of the null hypothesis at the two-sided 5% level under the null; empirical power, calculated as the proportion of tests that resulted in a rejection of the null hypothesis at the two-sided 5% level under the alternative. In addition, we also report the degree of variance inflation in CRTs with different levels of within-individual and between-individual correlations, by comparing the SE estimates with (methods in Sections 2.3 and 2.5) and without (methods in Sections 2.2 and 2.4) accounting for clustering under each type of model. We fixed $K = 100$ for this evaluation to avoid potential finite-sample bias of the SE estimates from the clustered survival models.

For studying type I error rate and power, in addition to the Wald test, we also implemented the permutation test when the number of clusters was relatively small ($K = 30$ and $K = 10$). Those were the cases in which the Wald test with the sandwich variance estimator is suspected to be anti-conservative. When $K = 10$, we enumerated the $S = \binom{10}{5} = 252$ permutations from the randomization space, and when $K = 30$, we randomly simulated $\widetilde{S} = 500$ permutations from the randomization space of size $S = \binom{30}{15} > 155$ million with duplicates removed. This strategy was motivated by considerations of computational time.

Finally, whereas the evaluation of SER, type I error rate and power does not require the true treatment effect from each model, the evaluation of RB and CP requires such information. Under the null scenario ($\delta = 0$), the treatment has no effect on the distribution of both $T_{1,kj}$ and $T_{2,kj}$ conditional on the frailty, and therefore has no effect on the cause-specific hazard and cumulative incidence function for $T_{1,kj}$. This implies that the null holds for both the Cox model and the Fine and Gray sub-distribution hazard model, or equivalently, the true $\beta = \gamma = 0$. Under the alternative, the true values for δ, β, and γ are not necessarily equal[35]; this is because δ represents the latent hazard ratio conditional on the frailty (latent with respect to the unobserved distribution of $T_{1,kj}$ conditional on the frailty but marginalized over $T_{2,kj}$), whereas β and γ represent the population-averaged cause-specific and sub-distribution hazard ratio (population-averaged due to marginalizing over the cluster-specific frailty). To address this complexity, we defined the truth for β and γ as the probability limits under which the respective estimating equations have mean zero.[37] We numerically approximated the true values of β and γ under our data generating process using additional Monte Carlo simulations with a large number of clusters. Specifically, we independently carried out simulations with $K = 200$ clusters and obtained the averages of the $\hat{\beta}$ and $\hat{\gamma}$ across the 1000 simulations. The Monte Carlo SE of $\hat{\beta}$ and $\hat{\gamma}$ in this "truth simulation" is relatively minimal, confirming the accuracy of our approximation to the true treatment effect in each model.

4 Results from the simulation study

4.1 Relative bias

Web Figure 1 summarizes the relative bias (RB) for estimating the treatment effect with the three marginal models under $K = 100, \exp(\delta) = 0.8$, and different values of the correlation parameters, τ_w and τ_b. All models have negligible bias for estimating their respective true treatment effect (either HR_{cs} or HR_{sub}). The RB for each model is similar and does not exceed 3% in the majority of cases. There is also no obvious systematic pattern to how the RB changes with different values of τ_w and τ_b. We also observe small RB when the true effect in the data generating process is changed to $\exp(\delta) = 0.5$ (Web Figure 2) or $\exp(\delta) = 2$ (Web Figure 3). As the true effect moves further away from the null, the RB for the effect measure in each model decreases. Finally, when the number of clusters decreases to $K = 30$ and $K = 10$, the RB generally increases, with the largest RB to be around 10% when the number of clusters $K = 10$ (see Web Figures 4 to 6 for $K = 30$ and Web Figures 7 to 9 for $K = 10$).

4.2 Standard error

We first focus on the three marginal models, and compute the SER, which is obtained as the ratio between Monte Carlo SE and the mean of the estimated SE; the former can be regarded as the true variability of the estimator under our data generating process. When the SE of the treatment effect is well estimated, the value of SER should be close to one. A value of SER that deviates from unity signals the approach is either over- or under-estimating the uncertainty of the treatment effect, leading to incorrect statistical inference. Web Figure 10 summarizes the SER for estimating the treatment effect using the three marginal models coupled with their sandwich variance estimators under $K = 100$, $\exp(\delta) = 0.8$, and varying values of the correlation parameters. In general, the SER is close to 1 for all models across all values of τ_w and τ_b, confirming that the sandwich variance estimator is consistent with a sufficient number of clusters in CRTs. While results for $K = 30$ and alternative values of δ are largely similar (Web Figures 11 through 15), the results for $K = 10$ suggests important caveats for the sandwich variance estimators for all models. Specifically, Web Figures 16 through 18 indicate the mean estimated sandwich standard error has noticeable negative bias (SER ≈ 1.4, which suggests that the SE was over-estimated by 40%) when $K = 10$ and between-individual correlation $\tau_b \geq 0.05$. The negative bias also increases as the between-individual correlation increases but remains insensitive to τ_w.

Because the clustered robust sandwich variance estimates are accurate with $K = 100$, we further approximate the values of the design effect or variance inflation factor (VIF) by taking the ratio between the clustered robust sandwich variance and the variance ignoring clustering (reference approaches in Sections 2.2 and 2.4). Unlike in simpler cases with continuous and binary outcomes, a closed-form VIF is intractable with complex clustered survival outcomes, presenting a major challenge for designing CRTs in our setting. For this reason, we provide some intuitions on the magnitude of the VIF in our simulation study under $K = 100$ for each sets of models: marginal Cox versus Cox; marginal Fine and Gray with clustering versus Fine and Gray; marginal multi-state versus multi-state. The results are

presented in Table 2 under a fixed within-individual correlation ($\tau_w = 0.05$) but with varying between-individual correlations τ_b and competing event rates. Above all, Table 2 confirms that increasing the between-individual correlation has a pronounced effect on the VIF, and a larger competing event rate reduces the VIF only when the between-individual correlation is large (e.g. $\tau_b = 0.3$). Second, whereas in most cases the differences in VIF among the three types of models are small, the sub-distribution hazard model (Fine and Gray type model) has the smallest VIF when both the between-individual correlation and competing event rate become large. Third, the VIF in CRTs with complex survival outcomes is clearly no longer a linear function in the between-individual correlation τ_b, which discourages the direct application of the well-known VIF results from continuous and binary outcomes in designing survival CRTs. Finally, Web Tables 1 through 6 present the VIF values with smaller and larger within-individual correlations, τ_w. Interestingly, while the VIF remains insensitive to τ_w when $\tau_w \leq 0.1$, larger values of the within-individual correlation can also reduce the VIF when the competing event rate is large. Overall, these findings imply that VIF in CRTs with complex survival outcomes is driven by both the within-individual and between-individual correlations as well as the competing event rate in a non-linear fashion.

4.3 Type I error rate

Figure 2 summarizes the empirical type I error rate of the Wald z-tests when $K = 100$ with varying values of the correlations and competing event rates. Based on the margin of error with 1000 replicates from a binomial model and the nominal test size at 5%, we consider an empirical type I error rate between 3.6% and 6.4% to be acceptable, and indicate the acceptable bounds in the respective figures. For brevity, we refer to the Wald tests with the clustered sandwich variance estimators as clustered Wald tests. Figure 2 shows that with a sufficient number of clusters, all clustered Wald tests maintain close-to-nominal type I error rate under all combinations of correlation values and competing event rates; the empirical type I error rate only occasionally exceeds 6.4% (mostly in the extreme scenario when $\tau_b = 0.3$). In contrast, the non-clustered Wald tests consistently exhibit a substantially inflated in type I error as long as $\tau_b \geq 0.05$ (results not shown), confirming the necessity for the variance estimator to account for clustering in CRTs with complex survival outcomes.

Web Figures 19 and 20 present the simulation results for type I error rates with a smaller number of clusters ($K = 30$ and $= 10$). In contrast to Figure 2, the clustered Wald-test grows slightly anti-conservative when $K = 30$, but markedly so when $K = 10$. For example, the largest type I error rate of the clustered Wald-test for each model could be over 8% when $K = 30$ but can even reach 15% when $K = 10$, as a consequence of the negative bias in the sandwich variance estimator. This finding is consistent to previous findings with GEE analyses of non-censored outcomes in CRTs,[38] except that the type I error inflation with complex survival outcomes appears much more pronounced than that with non-censored outcomes.

To mitigate the concerns with the type I error rate inflation with a smaller number of clusters, we examined the two permutation tests introduced in Section 2.6 for each of the three marginal models. Web Figures 21 and 22 present the type I error rate for the permutation tests when $K = 30$ and $K = 10$. For permutation tests in Figure 3 and Web

Figure 22, we exclude the extreme scenario with the largest between-individual correlation parameter ($\tau_b = 0.3$) when there are only $K = 10$ clusters, because fitting the marginal survival models and sandwich variance estimators repeatedly over permuted treatments leads to frequent non-convergence. From the results, it is evident that the permutation tests had satisfactory control of the test size at the nominal level, regardless of the correlation values or the competing event rate. Additionally, there is a little difference between the permutation β-test and the permutation z-test in terms of type I error. To facilitate a direct comparison, Web Figure 23 and Figure 3 present the type I error rate for the permutation β-test and the clustered Wald test under each of the three models, when $K = 30$ and $K = 10$. No clear pattern of type I error is shown as the simulation parameters, such as correlations and competing event rates, vary for the permutation β-test. Under most circumstances, the type I error for the permutation β-test lies randomly around 0.05, or below 0.05. For the Wald test, the type I error raises to around 0.1 when $K = 30$ and to 0.15 when $K = 10$, which is a serious issue threatening the validity of inference.

4.4 Statistical power

In parallel to Section 4.3, we consider the statistical power of each test from the marginal survival models. Figure 4 summarizes the statistical power for each model when $K = 100$ and $\exp(\delta) = 0.8$ under varying values of the correlation parameters and competing event rates. As the between-individual correlation τ_b increases, the power for all models, as expected, markedly decreases. However, for a given τ_b, it appears that larger values of the within-individual correlation τ_w translate to slight increases in the power of each test. The impact of τ_w on power is more pronounced with a larger competing event rate (0.12), but becomes negligible when the competing event rate is small (0.02).

Web Figures 24 and 25 summarize the corresponding results with $K = 100$ but with larger absolute effect sizes, that is, $\exp(\delta) = 0.5$ and 2. When the true treatment effect further deviates from the null, we observe that the marginal Fine and Gray model tends to have higher power compared to the marginal Cox and marginal multi-state models. For example, when $\exp(\delta) = 2$ (Web Figure 25), the between-individual correlation and competing event rate are both large, the marginal Fine and Gray model could have over 5% greater power than the marginal Cox and marginal multi-state models.

Web Figures 26 through 31 present the power for testing the treatment effect under each of the models considered with $K = 30$ and $K = 10$ clusters, and three effect sizes ($\exp(\delta) = 0.5, 0.8, 2$). With a smaller number of clusters, the power of the three Wald tests becomes more similar. The power comparison for these Wald tests, however, may be cautioned with $K = 30$ and $K = 10$ since the clustered Wald tests often carry inflated type I error rates. In this small-sample scenario, we also compare the power of the permutation tests since they have been demonstrated to maintain the nominal type I error rate. We summarize the empirical power for the permutation β-test and the permutation z-test in Web Figures 32 and 33 when $K = 30$ and $K = 10$, $\exp(\delta) = 0.5$, event rate = 0.08 and dropout rate = 0.03. The two types of permutation tests have negligible differences in power, and they appear only slightly less powerful compared with the clustered Wald tests. This may

be because that the sandwich variance estimator tends to have a negative bias with a small number of clusters, and more frequently rejects the null. Overall, when using the permutation tests with $K = 30$ and $K = 10$, the power difference across the three models is small, although occasionally the permutation tests based on the marginal Fine and Gray model have a slightly higher power.

4.5 Coverage probability

Web Figure 34 summarizes the coverage probability of the 95% confidence interval (CI) estimator from each of the models when $\exp(\delta) = 0.8$ and $K = 100$. The CIs are all constructed based on the normality assumption and the clustered sandwich variance estimator (results based on the non-clustered variance estimator not shown). Overall, the CIs from all marginal survival models maintain nominal coverage across all scenarios. Results for larger effect sizes are qualitatively similar and are presented in Web Figures 35 and 36.

In Web Figures 37 through 42, we summarize the results for coverage probabilities with $K = 30$ and $K = 10$. While the comparison between CIs based on the non-clustered and clustered sandwich variances for $K = 30$ and $K = 10$ are qualitatively similar to that for $K = 100$, one notable difference is that the clustered CI from each model tends to have undercoverage with a smaller number of clusters.[39] As we alluded to in Section 4.3, this is because the sandwich variance tends to underestimate the true variance with a limited number of clusters.[34] A potential remedy is to invert the permutation test described in Section 2.6. However, the inversion necessitates an exhaustive search of null hypotheses which the permutation test does not reject and will require substantially more computational effort than implementing a permutation test. We do not pursue this work here and return to a discussion in Section 6.

5 Illustrative analysis of the STRIDE trial

As described in Section 1, the STRIDE study was a pragmatic, parallel CRT focusing on reducing serious fall injuries in community-dwelling older adults at risk of falls.[7,6] The study enrolled 5451 adults aged 70 years and older from 86 primary care practices. The primary outcome was the time from enrollment to first serious fall-related injury,[8] and participants could die before observing the primary outcome. Because we are interested in the time to first event, death is considered as a competing risk. The event rates (first serious fall-related injury) among the control and intervention primary care practices were observed to be 5.3 and 4.9 per 100 person-years of follow-up, respectively, while the observed competing event rate (death) was smaller, 3.3 per 100 person-years of follow-up in both intervention and control practices.[6] Participants withdrew consent at a rate of 3.6 per 100 person-years of follow-up,[40] and 4187 participants (76.8%) were administratively censored.

The actual STRIDE study was designed using covariate-constrained randomization[41,42,30,31] to assign practices to treatment arms, and the primary analysis was specified to adjust for baseline covariates that are balanced in the design stage,[7,6] For demonstration, however, and in keeping with the preceding simulations, here we provide an unadjusted analysis that includes only the intervention effect. We considered the Cox model, the Fine and Gray model, and the multi-state model, with standard errors estimated either by the information

matrix or the robust sandwich variance estimator. The comparison between the two types of variance estimators can help assess the implication of clustering in STRIDE. The results are summarized in Table 3. Although the intervention effect parameters in the Cox, multi-state Cox, and Fine and Gray models have different interpretations (e.g. HR_{cs} vs. HR_{sub}), the estimated intervention coefficients in the STRIDE example are similar across three types of models, as a result of low event rate and competing event rate. Furthermore, accounting for clustering in each model through the sandwich variance estimator has minimal impact on the standard error estimates. This signals that intraclass correlations (both the within-individual and between-individual correlations) may be minimal for the survival times in each primary care practice. Indeed, when the intraclass correlation is minimal and the competing event rate is small compared to the event rate of interest, our simulation results show that all the models produce similar results. Although the STRIDE trial recruited a large number of clusters which alleviates the small sample considerations as in our simulation study, we additionally implement the permutation β-test and the permutation z-test with 10,000 permutations, both of which produce similar p-values to the corresponding Wald tests. For illustrative purposes, we interpret statistical significance at the 0.05 level and find that the intervention does not have a statistically significant effect on the risk of fall-related injuries among the STRIDE trial population. We do acknowledge, however, that the interpretation of the study results should not only rely on a single dichotomy of a p-value at the 0.05 threshold.

Beyond the overall analysis, we additionally performed subgroup analyses based on two potential effect modifiers—age (70–79 years vs. ≥ 80 years) and fear of falling only (yes vs. no; the participant had a negative response to all the fall-related screening questions except the question about whether he or she had a fear of falling). Due to the low event rate and death rate, we also observe little difference among the three models in subgroup analyses (Web Tables 7 to 10). Although our analysis of the STRIDE trial shows little difference among the three survival models as well as between the non-clustered and sandwich variance estimates, this example may not always reflect the usual case when both the event rates and intra-class correlations are higher. In those cases, our preceding simulations indicate that the sandwich variance estimator is required to maintain valid inference and the three survival models can produce different results.

6 Discussion

Motivated by the STRIDE trial, we provided an empirical comparison among analytical methods for CRTs with time-to-event outcomes in the presence of competing risks. We focus on readily implementable population-averaged survival models, and through extensive simulations (although not exhaustive), we study their operating characteristics for estimating their respective treatment effect parameters. Our results demonstrate that (i) all methods show trivial bias under all combinations of parameters in CRTs; (ii) the clustered sandwich variance estimators for all models are accurate with $K = 100$ clusters, but have negative bias with $K = 30$ and $K = 10$ clusters, resulting in inflated type I error rates and undercoverage; (iii) the permutation test has more robust control of type I error rates with small samples; (iv) under the alternative, while a larger value of the between-individual correlation reduces

the empirical power, a larger value of the within-individual correlation could slightly improve the power; (v) the three types of models have similar power in most cases, likely due to the low baseline event rate we assumed in the data generating process (to resemble the STRIDE trial). However, the marginal Fine and Gray model can have higher power than the other two marginal models, especially when the effect size is large, between-individual correlation is large and the within-individual correlation is small.

Although our study is motivated by the analysis of CRTs, it also has potential to inform the design of CRTs, particularly in terms of sample size estimation. To demonstrate the implications on the design of CRTs due to both clustering of survival outcomes and competing risks, we provided values of the VIF in our simulations and summarized them in Table 2. We found that for a given total sample size, VIF can depend on the between-individual correlation, within-individual correlation as well as the competing event rate in a nonlinear fashion. Importantly, this indicates that the usual VIF obtained for non-censored continuous or binary outcomes does not apply anymore to censored survival outcomes with competing risks. For accurately designing CRTs with survival outcomes subject to competing risks, our simulation routine exemplifies a simulation-based power calculation procedure, for each of the models we considered. Of note, there is a body of literature advocating simulation-based power calculation as a powerful and flexible approach in complex scenarios where the closedform sample size or VIF is unavailable.[43] To facilitate the design of complex CRTs, we provide our simulation code on the GitHub Repository (https://github.com/kyleyxw/simCRTs) so that others could adapt our code as a tool for simulation-based power calculation in CRTs.

In our simulation design, we used the Kendall's tau as a rank correlation to represent the degree of clustering.[44] While the Kendall's tau is a common measure of association for analyzing clustered survival data, it remains less familiar to investigators working with CRTs. This may be partly because the concept of intraclass correlation coefficient (ICC) as a linear correlation has now become a standard measure of clustering in CRTs with non-survival outcomes,[2] and relatively few published CRTs focused on survival outcomes. In fact, despite previous attempts that formalize the definition of ICC with clustered survival data,[44,11,45,46,12] there has not yet been a consensus on which of these definitions should be recommended for best practice.[47] Additionally, these definitions were currently restricted to clustered survival outcomes without competing risks, and future research is needed to provide such an extension and to better elucidate the pros and cons of alternative correlation measures in CRTs.

While accounting for clustering through the sandwich variance estimator is recommended in CRTs for each of the models we considered, the validity of the sandwich variance estimator only holds with a large number of clusters. With a limited number of clusters such as $K = 30$ and $K = 10$, our simulations show that the sandwich variance estimator has negative bias, leading to inflated type I error rate as large as 15% in certain scenarios. This is somewhat expected from prior studies which indicate the inadequacy of the sandwich variance estimator in CRTs with non-censored outcomes.[38] Of note, an alternative approach to summarize the power results in our simulations with small samples is to derive the size-adjusted power for each test through the receiver operating characteristic curves.[48,49]

However, we do not pursue this approach here because our objective is to identify tests that already maintain the nominal size with standard software implementation. On the other hand, prior simulation studies in parallel, crossover and stepped-wedge CRTs have found that the bias-corrected sandwich variance estimators may improve the validity of inference in CRTs.[19,50,51] While Fay and Graubard[34] suggested a bias correction to the marginal Cox model, we are not aware of any existing software packages that implement this method, nor any extensions to the marginal Fine and Gray or multi-state models. As a solution, we considered the permutation test as a flexible alternative that adequately controls the type I error rates in small CRTs. Although we have not studied the permutation-based CI estimators due to computational challenges in inverting the test, it remains important future work to develop computationally efficient approaches to invert the permutation test accounting for competing risks.

A possible limitation of our study is that we focused on the marginal model carrying a population-averaged interpretation. Frailty models, as an alternative, were not considered in this work because of their non-convergence issues in small samples and the cluster-specific interpretation of the treatment effect parameter. However, frailty models may have an efficiency advantage because the estimation of the conditional treatment effect parameter naturally accounts for the intraclass correlations through marginalizing the frailty distributions. A second limitation is that we have mainly considered scenarios with a competing event rate similar to the STRIDE trial, and we have not considered unequal competing event rates by treatment groups. The simulation results, therefore, may not be generalizable to more extreme settings where the competing event dominates the event of interest, or the competing event is affected by treatment. A third limitation is that we have not considered the possible recurrence of our target event (fall-related injury) in the simulations, and have not addressed the death event as a semi-competing risk.[52] This may be one of the reasons why the multi-state models performs similarly to the Cox models, as we are merely interested in the time to first event. In more general settings, the multi-state model can be more suitable to complex survival data with transitions to more than two states.[23]

Finally, we have not addressed the challenges in estimating the intraclass correlation parameters in the current simulations, and assumed working independence following the standard implementations in common software packages. While the correlation parameters have been traditionally regarded as nuisance parameters, it is especially important to report such values in analyzing CRTs because they are likely to inform the sample size calculation of future CRTs with similar endpoints.[53] Therefore, it would be of great interest to develop and compare methods for estimating the correlation parameters, which we demonstrate to be key determinants of the VIF in CRTs with complex survival outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by CTSA Grant Number UL1 TR0001863 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH) and by the STRIDE study, which was funded by the Patient Centered Outcomes Research Institute (PCORI), with additional support from the National Institute on Aging at NIH (U01AG048270). The authors thank the handling editor, Professor Andrew Forbes, and two anonymous reviewers for providing constructive comments, which greatly improved our manuscript.

References

1. Murray DM. Design and Analysis of Group-Randomized Trials. New York, NY: Oxford University Press, 1998.
2. Turner EL, Li F, Gallis JA et al. Review of recent methodological developments in group-randomized trials: Part 1—design. Am J Public Health 2017; 107: 907–915. [PubMed: 28426295]
3. Turner EL, Prague M, Gallis JA et al. Review of recent methodological developments in group-randomized trials: Part 2-analysis. Am J Public Health 2017; 107: 1078–1086. [PubMed: 28520480]
4. Fine JP and Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999; 94: 496–509.
5. Stedman MR, Lew RA, Losina E et al. A comparison of statistical approaches for physician-randomized trials with survival outcomes. Contemp Clin Trials 2012; 33: 104–115. [PubMed: 21924382]
6. Bhasin S, Gill TM, Reuben DB et al. A randomized trial of a multifactorial fall injury prevention strategy. N Engl J Med 2020; 383: 129–140. [PubMed: 32640131]
7. Bhasin S, Gill TM, Reuben DB et al. Strategies to reduce injuries and develop confidence in elders (STRIDE): A cluster-randomized pragmatic trial of a multifactorial fall injury prevention strategy: Design and methods. The Journals of Gerontology: Series A 2018; 73: 1053–1061.
8. Ganz DA, Siu AL, Magaziner J et al. Protocol for serious fall injury adjudication in the strategies to reduce injuries and develop confidence in elders (STRIDE) study. Inj Epidemiol 2019; 6: 1–8. [PubMed: 30637568]
9. Caille A, Tavernier E, Taljaard M et al. Methodological review showed that time-to-event outcomes are often inadequately handled in cluster randomized trials. J Clin Epidemiol 2021; 134: 125–137. [PubMed: 33581243]
10. Manatunga AK and Chen S. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. Biometrics 2000; 56: 616–621. [PubMed: 10877325]
11. Xie T and Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. Stat Med 2003; 22: 2835–2846. [PubMed: 12953283]
12. Jahn-Eimermacher A, Ingel K and Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. Stat Med 2013; 32: 739–751. [PubMed: 22865817]
13. Zhong Y and Cook RJ. Sample size and robust marginal methods for cluster-randomized trials with censored event times. Stat Med 2015; 34: 901–923. [PubMed: 25522033]
14. Li J and Jung SH. Sample size calculation for cluster randomization trials with a time-to-event endpoint. Stat Med 2020; 39: 3608–3623. [PubMed: 33463748]
15. Vaida F and Xu R. Proportional hazards model with random effects. Stat Med 2000; 19: 3309–3324. [PubMed: 11122497]
16. Wei LJ, Lin DY and Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc 1989; 84: 1065–1073.
17. Lin DY and Wei LJ. The robust inference for the Cox proportional hazards model. J Am Stat Assoc 1989; 84: 1074–1078.
18. Preisser JS, Young ML, Zaccaro DJ et al. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med 2003; 22: 1235–1254. [PubMed: 12687653]
19. Li F, Turner EL and Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics 2018; 74: 1450–1458. [PubMed: 29921006]

20. Lau B, Cole SR and Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol 2009; 170: 244–256. [PubMed: 19494242]
21. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13–22.
22. Spiekerman CF and Lin D. Marginal regression models for multivariate failure time data. J Am Stat Assoc 1998; 93: 1164–1175.
23. Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. Stat Med 2007; 26: 2389–2430. [PubMed: 17031868]
24. Andersen PK and Keiding N. Multi-state models for event history analysis. Stat Methods Med Res 2002; 11: 91–115. [PubMed: 12040698]
25. Zhou B, Fine J, Latouche A et al. Competing risks regression for clustered data. Biostatistics 2012; 13: 371–383. [PubMed: 22045910]
26. Fiero MH, Huang S, Oren E et al. Statistical analysis and handling of missing data in cluster randomized trials: A systematic review. Trials 2016; 17: 72. [PubMed: 26862034]
27. Murray DM, Pals SL, Blitstein JL et al. Design and analysis of group-randomized trials in cancer: A review of current practices. J Natl Cancer Inst 2008; 100: 483–491. [PubMed: 18364501]
28. Ivers N, Taljaard M, Dixon S et al. Impact of consort extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000–8. Bmj 2011; 343: d5886. [PubMed: 21948873]
29. Gail MH, Mark SD, Carroll RJ et al. On design considerations and randomization-based inference for community intervention trials. Stat Med 1996; 15: 1069–1092. [PubMed: 8804140]
30. Li F, Lokhnygina Y, Murray DM et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med 2015; 35: 1565–1579. [PubMed: 26598212]
31. Li F, Turner EL, Heagerty PJ et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. Stat Med 2017; 36: 3791–3806. [PubMed: 28786223]
32. Cai J and Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. Stat Med 2000; 19: 2963–2973. [PubMed: 11042626]
33. Wang R and De Gruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. Stat Med 2017; 36: 2831–2843. [PubMed: 28464567]
34. Fay MP and Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 2001; 57: 1198–1206. [PubMed: 11764261]
35. Emura T, Shih JH, Ha ID et al. Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. Stat Methods Med Res 2019; 29: 2307–2327. [PubMed: 31868107]
36. Nelsen RB. An introduction to copulas. Springer, New York, NY: Springer Science & Business Media, 2007.
37. White H Maximum likelihood estimation of misspecified models. Econometrica: Journal of the Econometric Society 1982; 50: 1–25.
38. Li P and Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Stat Med 2015; 34: 281–296. [PubMed: 25345738]
39. Lu B, Preisser JS, Qaqish BF et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics 2007; 63: 935–941. [PubMed: 17825023]
40. Gill TM, McGloin JM, Shelton A et al. Optimizing retention in a pragmatic trial of community-living older persons: The STRIDE study. J Am Geriatr Soc 2020; 68: 2492–2499. [PubMed: 32949145]
41. Moulton LH. Covariate-based constrained randomization of group-randomized trials. Clinical Trials 2004; 1: 297–305. [PubMed: 16279255]
42. Greene EJ. A SAS macro for covariate-constrained randomization of general cluster-randomized and unstratified designs. J Stat Softw 2017; 77: 1–20.
43. Landau S and Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. Stat Methods Med Res 2013; 22: 324–345. [PubMed: 22491174]

44. Hougaard P. Analysis of multivariate survival data. Springer, New York, NY: Springer Science & Business Media, 2012.
45. Gangnon RE and Kosorok MR. Sample-size formula for clustered survival data using weighted log-rank statistics. Biometrika 2004; 91: 263–275.
46. Jung SH. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: A simulation method. J Biopharm Stat 2007; 17: 839–849. [PubMed: 17885869]
47. Kalia S, Klar N and Donner A. On the estimation of intracluster correlation for time-to-event outcomes in cluster randomized trials. Stat Med 2016; 35: 5551–5560. [PubMed: 27790737]
48. Lloyd CJ. Estimating test power adjusted for size. J Stat Comput Simul 2005; 75: 921–933.
49. Heritier S, Lloyd CJ and Lo SN. Accurate p-values for adaptive designs with binary endpoints. Stat Med 2017; 36: 2643–2655. [PubMed: 28470713]
50. Li F, Forbes AB, Turner EL et al. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. Stat Med 2019; 38: 636–649. [PubMed: 30298551]
51. Thompson J, Hemming K, Forbes A et al. Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. Stat Methods Med Res 2021; 30: 425–439. [PubMed: 32970526]
52. Zhao Y, Tian X, Cai J et al. (2022). Bayesian semi-parametric inference for clustered recurrent events with zero-inflation and a terminal event. arXiv preprint arXiv:2202.06636.
53. Campbell MK, Piaggio G, Elbourne DR et al. Consort 2010 statement: Extension to cluster randomised trials. BMJ 2012; 345: 1–21.

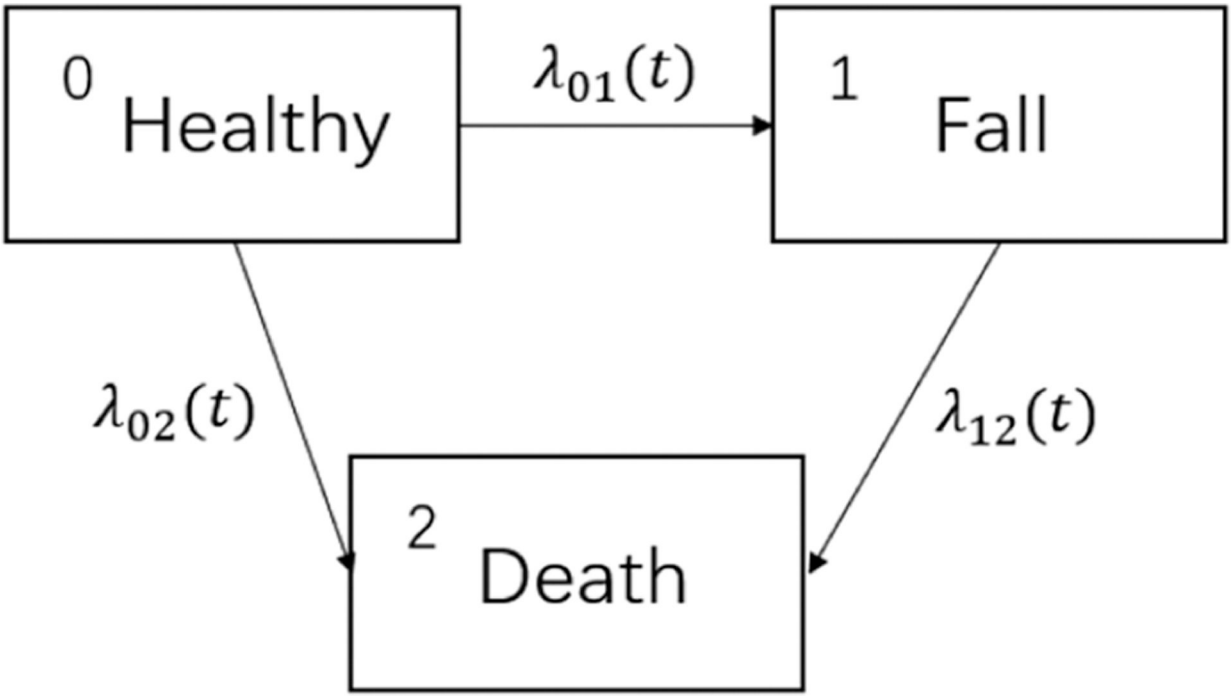

Figure I.
A schematic illustration of the unidirectional illness–death model in the context of the STRIDE study.

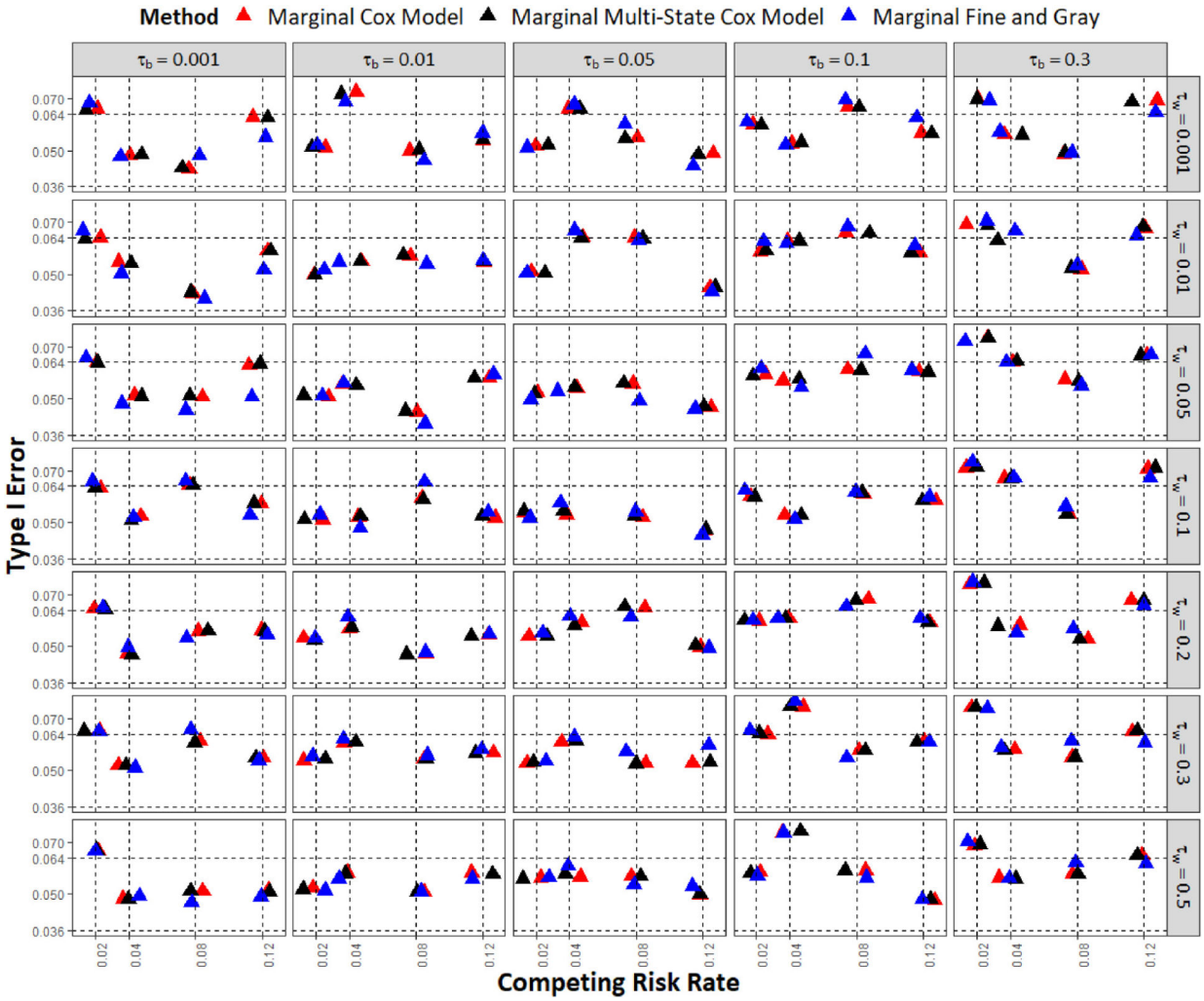

Figure 2.

Type I error rate with varying competing event rates (0.02–0.12) under different combinations of within-individual correlation (τ_w) and between-individual correlation (τ_b); the number of clusters $K = 100$ event rate = 0.08 and dropout rate = 0.03. Empirical type I error rates between 3.6% and 6.4% (indicated by horizontal dashed lines) are considered close to nominal based on a binomial model with 1000 replicates.

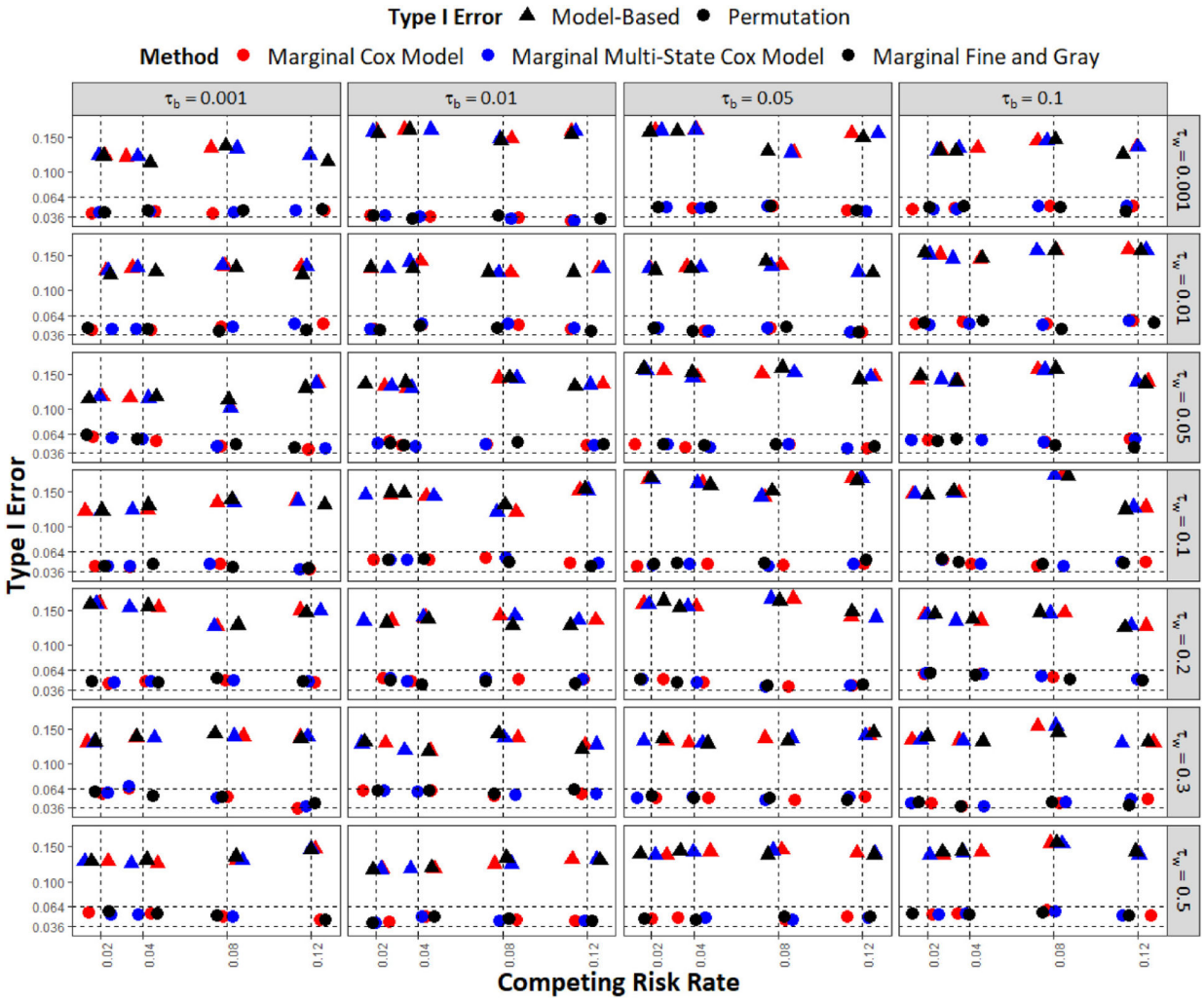

Figure 3.

Direct comparison of type I error rate between the permutation β-test and the sandwich variance based Wald test with varying death rates (0.02–0.12) under different combinations of within individual correlation (τ_w) and between individual correlation (τ_b); the number of clusters $K = 10$, event rate = 0.08 and dropout rate = 0.03. Empirical type I error rates between 3.6% and 6.4% (indicated by horizontal dashed lines) are considered close to nominal based on a binomial model with 1000 replicates. The scenario $\tau_b = 0.3$ is excluded due to non-convergence.

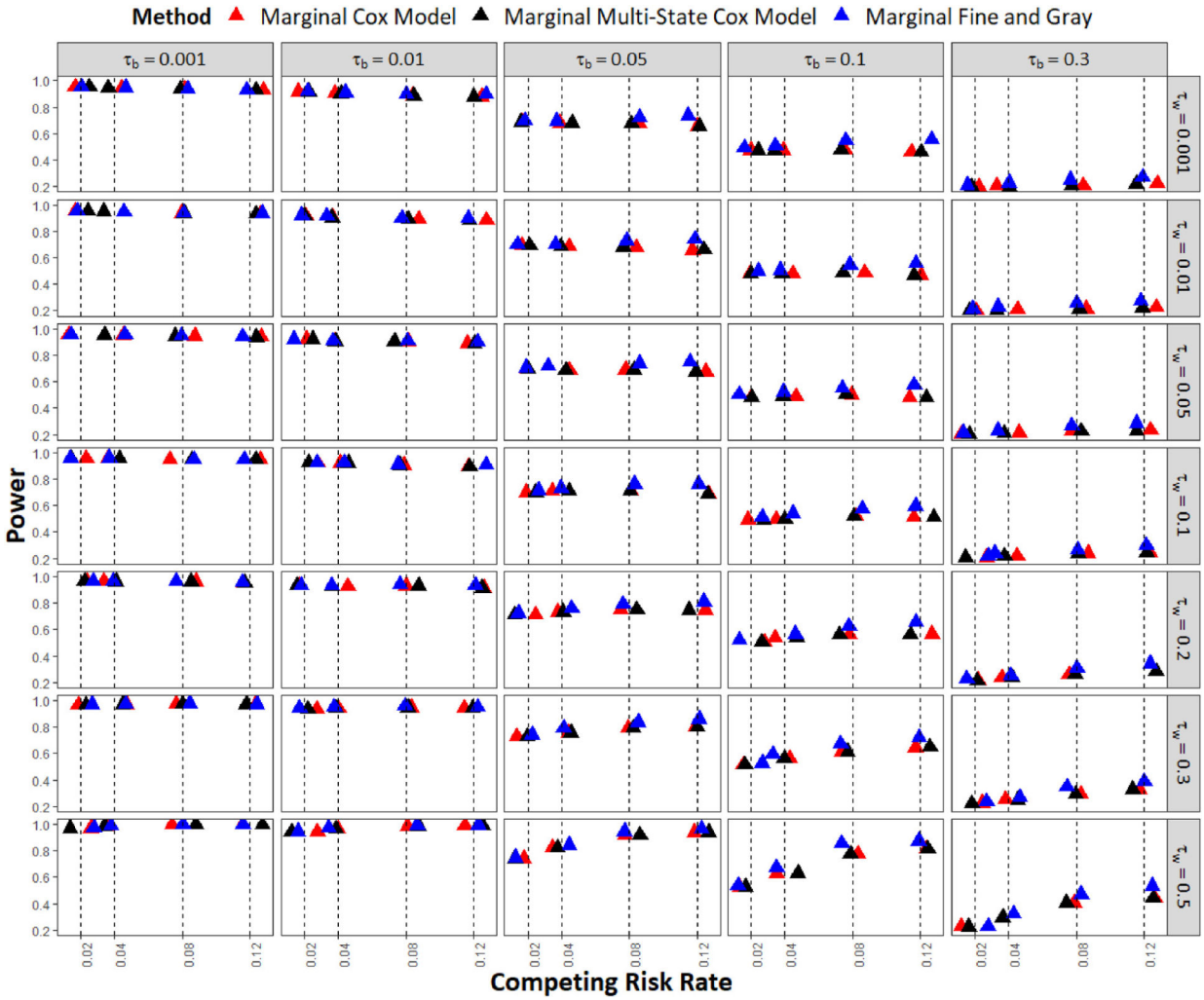

Figure 4.

Power with varying competing event rates (0.02–0.12) under different combinations of within-individual correlation (τ_w) and between-individual correlation (τ_b); the number of clusters $K = 100$, latent hazard ratio = 0.8, event rate = 0.08, and dropout rate = 0.03.

Table I.

Survival models under consideration and their implementations in existing R packages. Cox and marginal Cox models estimate the cause-specific hazard ratio; Fine and Gray and marginal Fine and Gray models estimate the sub-distribution hazard ratio; multi-state and marginal multi-state models estimate the transition-specific hazard ratio.

Censor competing event	Account for clustering	Model	Package	Function call
Yes	No	Cox	survival	coxph
	Yes	Marginal Cox	survival	*coxph* with **cluster** argument
No	No	Fine and Gray	comprsk	crr
		Multi-state Cox	survival	*coxph* with **as.factor(status)** specification and **id** argument
	Yes	Marginal Fine and Gray	crrSC	*crrc* with **cluster** argument
		Marginal multi-state Cox	survival	*coxph* with **as.factor(status)** specification, **id** and **cluster** arguments

Table 2.

Variance inflation with $K = 100$, varying competing event rates (from 0.02 to 0.12), and between-individual correlations (τ_b from 0.00I to 0.3) when the within-individual correlation is at a fixed level ($\tau_w = 0.05$). Variance inflation is calculated as the ratio of SER between the model accounting for clustering and the model that does not account for clustering. Cox: ratio of SER between marginal Cox and traditional Cox; Fine and Gray: ratio of SER between marginal Fine and Gray and traditional Fine and Gray; Multi-state: ratio of SER between marginal multi-state and traditional multi-state.

		Competing event rate			
Model	τ_b	**0.02**	**0.04**	**0.08**	**0.12**
Cox	0.001	1.013	1.012	1.004	1.028
	0.01	1.325	1.344	1.342	1.320
	0.05	2.175	2.135	2.045	1.962
	0.1	3.873	3.741	3.505	3.280
	0.3	10.045	9.409	8.369	7.647
Fine and Gray	0.001	1.008	1.004	0.999	1.013
	0.01	1.313	1.313	1.291	1.247
	0.05	2.102	1.998	1.801	1.643
	0.1	3.711	3.428	2.962	2.573
	0.3	9.559	8.517	6.910	5.841
Multi-state	0.001	1.013	1.012	1.005	1.028
	0.01	1.326	1.344	1.343	1.321
	0.05	2.176	2.136	2.046	1.963
	0.1	3.874	3.742	3.507	3.282
	0.3	10.050	9.414	8.374	7.651

Table 3.

Analyses of STRIDE trial using: Cox with (marginal Cox) and without (Cox) clustering; Fine and Gray with (marginal Fine and Gray) and without (Fine and Gray) clustering; Multi-state with (marginal multi-state) and without (multi-state) clustering. The HR column refers to the cause-specific HR for the Cox and multi-state models, while the HR column refers to the sub-distribution HR for the Fine and Gray models.

					p-value		
Model	**Intervention coefficient**	**Standard error**	**HR**	**95% Wald CI**	**Wald**	**perm β**	**perm z**
Cox	−0.0936	0.0822	0.9106	(0.7751, 1.0698)	0.255	–	–
Marginal Cox	−0.0936	0.0821	0.9106	(0.7753, 1.0696)	0.254	0.270	0.262
Fine and Gray	−0.0932	0.0822	0.9110	(0.7755, 1.0702)	0.257	–	–
Marginal Fine and Gray	−0.0932	0.0821	0.9110	(0.7757, 1.0700)	0.256	0.268	0.262
Multi-state Cox	−0.0909	0.0822	0.9131	(0.7773, 1.0727)	0.269	–	–
Marginal multi-state Cox	−0.0909	0.0826	0.9131	(0.7767, 1.0735)	0.271	0.287	0.276