Assessing the Performance of Different Models across Varying Proportions of Paired Recurrent Event Data

Assessing the Performance of Different Models across Varying Proportions of Paired Recurrent Event Data

Analysis of paired recurrent event data is important in medical studies where patients can have recurring events in paired anatomical features. Extensions of the Cox proportional hazards model have been proposed to handle clustered or recurring events. When the primary research question involves recurrent events *within* a pair, extra considerations must be given due to the nature of the data. As a motivating example tympanoplasties were considered, medical procedures used to repair perforated eardrums. In practice, each patient can be considered a cluster with paired ears as a sub-cluster, where it is possible for a patient to have sub-clusters of one or both ears. Potential graft failure rates may differ by sub-cluster, where some ears may never experience an event, while others may experience multiple. This study assesses the performance of several statistical methods for nested correlated time to failure data through simulation and application to a research study on time to tympanoplasty. The statistical methods compare a generalized linear mixed model, the Prentice-Williams-Peterson model, a gamma frailty model, nested maximum and partial penalized likelihood frailty models, and a hierarchical copula model. Model performance is explored across simulated scenarios with recommendations made for which method best models the data.

Keywords: hierarchical clustering, nested clustering, paired recurrent time to failure, simulation study, survival analysis

1. INTRODUCTION

In biomedical research, studies can include a hierarchical grouping of survival times. For example, a chronic granulomatous disease study, which investigated the effectiveness of a new treatment in reducing rates of infections, encompasses clustering at two hierarchical levels (Rondeau, Filleul, and Joly 2006; Schukken et al. 2010). The upper hierarchical level, denoted the cluster, is formed by patients from the same hospital and the lower hierarchical level, denoted the sub-clusters, by the patient's multiple events. In such data it is thought that heterogeneity between patients and heterogeneity between hospitals may both be present and important to account for in statistical modeling (Rondeau, Filleul, and Joly 2006).

Data can also be naturally clustered at two hierarchical levels within an individual or unit of observation. As the motivating example for this study, we consider data related to pediatric tympanoplasties, a medical procedure used to repair perforated eardrums in children with a graft. In practice, each patient can be considered a cluster with ears as the sub-cluster, where it is possible for a patient to have a single sub-cluster of just one ear with a perforation or two sub-clusters, one for each ear with perforations needing repair. When the primary research question involves recurrent events within a paired anatomical feature, extra considerations must be given due to the unique nested nature of the data. In this motivating context, there may be a large number of clusters, being the number of individuals in the study population, but only a small number of sub-clusters possible within each cluster, as only one or two ears may be included and these may vary by child (cluster). Additionally, the number of observations in each sub-cluster can vary based on how many recurrent events are experienced. At a minimum a sub-cluster could have no event experienced (i.e., a successful initial graft placement), or could have multiple graft failures that may be due to graft type, patient characteristics, or other, possibly unknown reasons.

In many cases, hierarchical nested survival data are inappropriately analyzed by ignoring one of the levels of correlation in the data (Rondeau, Filleul, and Joly 2006), for example, by using a generalized estimating equation, a random effects model, a standard frailty model, or a recurrent events model. Ignoring the hierarchical correlation may lead to regression parameter estimates that are both biased and have underestimated variances (Rondeau, Filleul, and Joly 2006; Schukken et al. 2010). Recently, advances have been made with nested frailty models and copula-based approaches for data with two hierarchical cluster levels, making them ideal for this type of hierarchically clustered data. Both types of models have been validated with moderate cluster sizes and sub-cluster sizes. For example, Rondeau et al. (2006) looked at fixed cluster/sub-cluster/number of recurrent events of 50/6/5 and 20/4/5, and found their proposed nested frailty model using maximum penalized likelihood estimation to be unbiased in estimating both fixed effects and random effects. Su et al. (2019) proposed a method which uses a hierarchical Kendall copula that was also validated with similar cluster, sub-cluster and observations sizes. What has not yet been elucidated is the performance of these models when the sub-cluster sizes are small and the number of recurrent events may also be infrequent, making optimal model selection challenging given the previously unexamined trade-offs in bias, variance, and power for the various possible choices.

Since the structure of paired recurrent time to event data can be applied to ears, eyes, lungs, and even anatomical features of animals such as cow udders (Schukken et al. 2010), the proportion of subjects with paired body parts may vary depending on the question of interest and the field of study. The purpose of this novel investigation is to evaluate the performance of several models with respect to bias, variance, power, and type-1 error rate, and identify the approach(es) for data with varying proportions of pairing in the data and potentially infrequent

events that maintain satisfactory statistical operating characteristics. Section 2 introduces and describes the methods considered, introduces the motivating dataset, and the simulation study set-up. Section 3 presents the results and Section 4 includes a discussion of our findings.

2. METHODS

2.1 Models

Many different approaches have been proposed that could be implemented to model hierarchical nested time to event data. We examine six different models to provide an evaluation of analyses that have either been applied in prior research studies or that represent more recently developed statistical methodology and may be more appropriate for the complex structure of the data, but heretofore were not often considered in practice for evaluation of clinical research studies. The six models considered are a logistic random effects model, the Prentice-Williams-Peterson model (PWP), a standard Gamma frailty model, a nested frailty model using maximum penalized likelihood estimation (MPL) through the frailtypack (2019) R package (Rondeau, Filleul, and Joly 2006), a nested frailty model using a penalized partial likelihood estimation (PPL) through the coxme (2020) R package (Therneau 2020a), and a hierarchical Kendall copula model. The logistic random effects model uses a binary outcome of event or no event within a given timeframe, whereas all other models use time to event as the outcome. In the following subsections we briefly introduce the methods and highlight some of their potential limitations or advantages for hierarchically clustered data.

2.1.1 Logistic Regression with Random Effects

Random effects or generalized estimating equations for logistic regression models are able to account for levels of correlation; however, they still inefficiently use the data by ignoring the information on the timing of events and treating the outcome as dichotomous (Amorim and Cai

2015). While this modeling approach is statistically suboptimal given the nested nature of the data, this study will still examine the performance of a logistic regression random effects model that includes a random intercept for the sub-cluster level since this treatment of time to event data as dichotomous has been applied in prior tympanoplasty studies (Kessler, Potsic, and Marsh 1994; Baklaci et al. 2018; Cass, Patten, and Cass 2019).

2.1.2 Prentice-Williams-Peterson

Common approaches for analyzing recurrent time to event data include marginal models such as the Andersen and Gill (AG) model using the counting process or the Prentice-Williams-Peterson (PWP) model using gap times (Prentice, Williams, and Peterson 1981; Andersen and Gill 1982). Both the AG and the PWP models consider recurrent events to be ordered and use robust variance estimates to account for interdependence due to the repeated events within a subject (Schukken et al. 2010). These robust variance estimates assume that observations across clusters are independent, but are not independent within a cluster. Gap time models are thought to be better suited when individuals are restored to a similar physical state, which reflects the motivating context of our real world pediatric tympanoplasty dataset, where a graft is placed to repair the eardrum (Cook and Lawless 2007). Therefore, the PWP model will be evaluated in this study due to its use of gap time as the time scale. The cluster variable in the PWP model will be set to the lower-level cluster level (i.e. ears) since that is the level at which recurrent events are experienced. However, this does represent a limitation for the PWP model in that it only accounts for the correlation of repeated events through the robust standard errors and does not account for other potential hierarchical levels of clustering, such as the paired ears within an individual.

2.1.3 Gamma Frailty

The frailty model is a random effect survival model which accounts for correlation due to clustering within the data. The frailty term is used to summarize the unmeasured heterogeneity in the hazard that cannot be explained by other covariates and is commonly modeled assuming a gamma or lognormal distribution. This model can be quite useful in biomedical research as it is thought individuals within the same cluster share similar unobserved factors that would affect their survival. The frailty term in our context is set at the upper-cluster level (i.e., person) and modeled with a parametric distribution. A gamma frailty model can be written as

$$\lambda(t_{ij}|Z_i) = Z_i\lambda_0(t_{ij})\exp(\boldsymbol{\beta}^t\boldsymbol{X}_{ij}) \quad (1)$$

$$i = 1, \dots, n \quad (2a)$$

$$j = 1, \dots, m \quad (2b)$$

$$Z_i \sim \Gamma(1/\rho, 1/\rho) \quad (2c)$$

Where $\lambda_0(t_{ij})$ is an unspecified baseline hazard function, Z_i is the frailty of group i with a mean of 1 and a variance of ρ, and j is the subject within cluster i. However, standard frailty models are also only able to account for one level of correlation, which is a limitation when it is expected that hierarchical nested correlation is possible and may have a meaningful effect.

2.1.4 Nested Frailty

Nested frailty models are an extension of the standard frailty model and include two nested random effects (Fairbairn 2016). Often the two frailty terms are assumed to follow a gamma distribution for reasons of mathematical convenience. An advantage of the nested frailty model is that it simultaneously considers the two levels of clustering which, in theory, should lead to more accurate estimates, and explicitly accounts for the hierarchical nature of the data. (Rondeau, Filleul, and Joly 2006). Though originally developed in 1997 (Sastry 1997), the nested frailty model has only recently become available within software programs due to the computational

challenges encountered in model estimation. Within our study we focus on two different approaches to implementing the nested frailty model: maximum penalized likelihood (MPL) and partial penalized likelihood (PPL).

We briefly define the nested frailty model estimated with MPL estimation. The model can be written as,

$$\lambda_{ijk}\left(t\middle|v_i, w_{ij}\right) = \lambda_0(t) v_i w_{ij} \exp\left(\boldsymbol{\beta}' \boldsymbol{X}_{ijk}\right) \tag{3}$$

$$i = 1, \ldots, n \tag{4a}$$

$$j = 1, \ldots, m \tag{4b}$$

$$k = 1, \ldots, q \tag{4c}$$

$$v_i \sim \Gamma(1/\rho, 1/\rho) \tag{4d}$$

$$w_{ij} \sim \Gamma(1/\eta, 1/\eta) \tag{4e}$$

where $\lambda_{ijk}\left(t\middle|v_i, w_{ij}\right)$ is the conditional hazard function for event k of sub-cluster j in cluster i, conditional on the two frailties v_i and w_{ij} which account for the cluster and sub-cluster effects, respectively. $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}$ is the vector of regression parameters, and $\boldsymbol{X}_{ijk} = (X_{1ijk}, \ldots, X_{hijk})'$ represents the covariate vector for the kth event, with h covariates. This model assumes the two frailty terms follow a gamma distribution.

The nested frailty model using PPL estimation can be written with the same notation as the nested MPL frailty model. However, the nested PPL model assumes a Gaussian distribution for the two random effects instead of the gamma distribution.

2.1.5 Hierarchical Kendall Copula

The application of copulas in the analysis of hierarchical clustered data has made substantial progress in recent years thanks to increases in computational power that more efficiently accommodate a hierarchical dependence structure (Su, Nešlehová, and Wang 2019). The

hierarchical Kendall copula (HKC) model leverages Sklar's Theorem which states that the joint survival function of any random vector can be expressed by its univariate marginal distribution functions $F_{X_1}, \dots, F_{X_d}$ and a copula C (Sklar 1951):

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C(F_{X_1}(x_1), \dots, F_{X_d}(x_d)) \tag{5}$$

This allows for the dependence structure to be modeled independently of the marginal distribution, which potentially gives additional flexibility (Meyer and Romeo 2015). The HKC approach uses marginal estimates of the parameters from a survival model that can account for the correlation at the upper level of clustering. The parameter estimates are then incorporated in a step-wise approach to estimate the variance of the cluster and sub-cluster terms (Spiekerman and Lin 1998; Su, Nešlehová, and Wang 2019). The lower-level cluster copula parameter is then estimated by maximizing the pseudo log likelihood assuming the lower-level clusters are independent. Subsequently, the upper-level cluster copula parameter is estimated by imputing event times for censored observations and then maximizing the pseudo log likelihood on the imputed full dataset (Su, Nešlehová, and Wang 2019).

2.2 Motivating Dataset

Tympanic membrane perforation is a widespread condition of all ages that often requires repair with tympanoplasty. If a tympanoplasty is deemed a failure, the patient will undergo a second surgery to repair the eardrum. In addition to the possibility of having more than one tympanoplasty procedure on a patient's ear over their lifetime, a single patient may have a tympanoplasty done in both ears. In this particular context we can consider the patients to be the clusters and the left and/or right ear to be the sub-clusters, with the observations being the potentially recurrent surgeries in the patient's ears. Those patients who experience the need for a tympanoplasty in both ears may also experience failures at different rates for each ear.

Tympanoplasty failure data was obtained on 600 children who underwent at least one ear graft surgery at the Children's Hospital Colorado (CHCO) from 1999 to 2019 (Colorado Multiple Institution Review Board protocol number 18-2517). Of the 600 children in this dataset, 655 ears had a tympanoplasty, resulting in about 10% of the patients having both ears treated. Over the 20-year study period a total of 784 tympanoplasties were done. Eligibility for the study was determined by having a tympanoplasty graft placed in an ear, with failure defined as having a repeat tympanoplasty surgery in the same ear. The number of known failed tympanoplasty surgeries an ear experienced ranged from zero to three. However, since there were only a few ears that had three failures, the data were truncated to include only the first two failures. While multiple types of graft are used in practice, for simplicity graft type was dichotomized into Fascia vs non-Fascia for evaluating the potential treatment effect (Patil, Baisakhiya, and Deshmukh 2014). The covariates modeled with the CHCO dataset were ear (side), treatment, and the interaction between the two. The six previously described models were fit to the CHCO dataset, with conclusions evaluated in light of the of the simulation study results for 10% sub-cluster pairs corresponding to the 10% of patients with both ears treated.

2.3 Simulation Studies

The performance of the six models was compared through simulation. Since the structure of the tympanoplasty data is inherently nested, the simulated datasets were generated from the nested MPL frailty Cox model. The model included cluster and sub-cluster random effects and the same three predictors as the CHCO dataset: side of sub-cluster (left or right ear), a treatment assignment (Fascia or non-Fascia), and the interaction between the two. Sub-cluster side was assumed to be constant over time, while treatment could vary by each ear and recurrent event. In order to investigate the effect of varying pair proportions (i.e., the proportion of clusters with two

sub-clusters vs one) we simulated across a range of percentages: 10%, 50%, 60%, 70%, 80% and 90%. A sample size of 600 individuals was simulated to reflect our motivating dataset. This means that when the percentage of individuals with pairs is 50%, there would be approximately 900 sub-clusters in total, and when the percentage was 90% there would be about 1,140 sub-clusters in total.

A total of 1,000 simulated datasets for each pair proportion was generated in R (version 3.6.2) using the technique described in Bender et al. (2005). For each simulation, the random variable for the cluster random effect was generated following model (3) with $\rho = 0.3$, while the random variable for the sub-cluster random effect was generated with $\eta = 0.15$. The magnitude of the cluster and subcluster term were chosen based on those examined in prior simulation studies for these methods (Rondeau, Filleul, and Joly 2006). The two fixed effects variables were randomly generated from a Bernoulli distribution with probability of 0.5, and the third variable being the interaction between the two. The time-to-event outcome in days to failure was generated based on a Weibull baseline hazard with scale and shape parameters of 0.038 and 1.2, respectively using a gap time scale. Beta coefficients were set to 0.3 for treatment, 0.4 for side and -0.25 for the interaction, resulting in hazard ratios of 1.35, 1.49 and 0.78, respectively. For each subject a right-censoring variable was independently generated from a uniform distribution on the interval of 1 to 130 days to achieve approximately 15% censoring. The maximum number of events a sub-cluster could experience was restricted to a maximum of two. Additionally, the simulations were repeated for a null case where all beta coefficients were set equal to zero.

Evaluations of each survival model were based on bias, variance, mean squared error (MSE), type-1 error rate, and power as estimated from the 1,000 simulated datasets for each simulation

scenario. The logistic regression model was assessed only on power due to the parameter estimates being on the odds ratio scale and ultimately not being directly comparable to the other models and data generating structure. The threshold for statistical significance was set at $\alpha = 0.05$ for all parameters. Sensitivity scenarios with varied considerations and assumptions were conducted and are included in the Supplementary Materials.

2.4 Implementation

All models were fit in R (version 3.6.2). The logistic regression model with a random intercept for the sub-cluster term was fit with the glmer function in the lme4 package (Bates et al. 2014). The PWP model and the gamma frailty model, with a frailty term for the sub-cluster term, both used the coxph function in the survival package (Therneau 2020b). The nested frailty MPL model used the frailtyPenal function from frailtypack package (Rondeau, Filleul, and Joly 2006). The nested frailty PPL model used the coxme function from the coxme package (Therneau 2020a). Lastly, the HKC model was implemented using R code found in the supplement document from Su, Nešlehová and Wang (Su, Nešlehová, and Wang 2019).

3. RESULTS

3.1.Model Evaluation

While we discuss the results across the range of pairs proportion, for brevity we only present tables for the bias and power from the simulation scenario with 90% pairs in Table 1 and from the simulation scenario with 50% pairs in Table 2. The results from all simulation scenarios, including variance, MSE and the null scenario, can be found in the supplementary material (Table 4-14). Line graphs of the bias, variance, MSE and power from all of the simulations are provided (Figures 1-4, Figures 5-8). AIC was also evaluated for all survival models where possible (Table 15).

3.1.1.Bias

Across the scenarios the parameter estimates remained fairly unbiased. For the parameter of treatment ($\beta_1 = 0.3$) for the scenario with 90% pairs, the PWP and the HKC regression models resulted in the most bias of -0.082, while the nested MPL frailty model resulted in the least amount of bias with 0.003. When the pair proportion was set to 50%, the gamma frailty model provided the least biased estimate with a bias of -0.004 and the HKC model provided the least accurate estimate with a bias of -0.076. From Figure 1 it can be seen that for the five time to event models the bias for treatment slightly varies as the pair proportions go from 90% to 10%, though remains relativity unbiased.

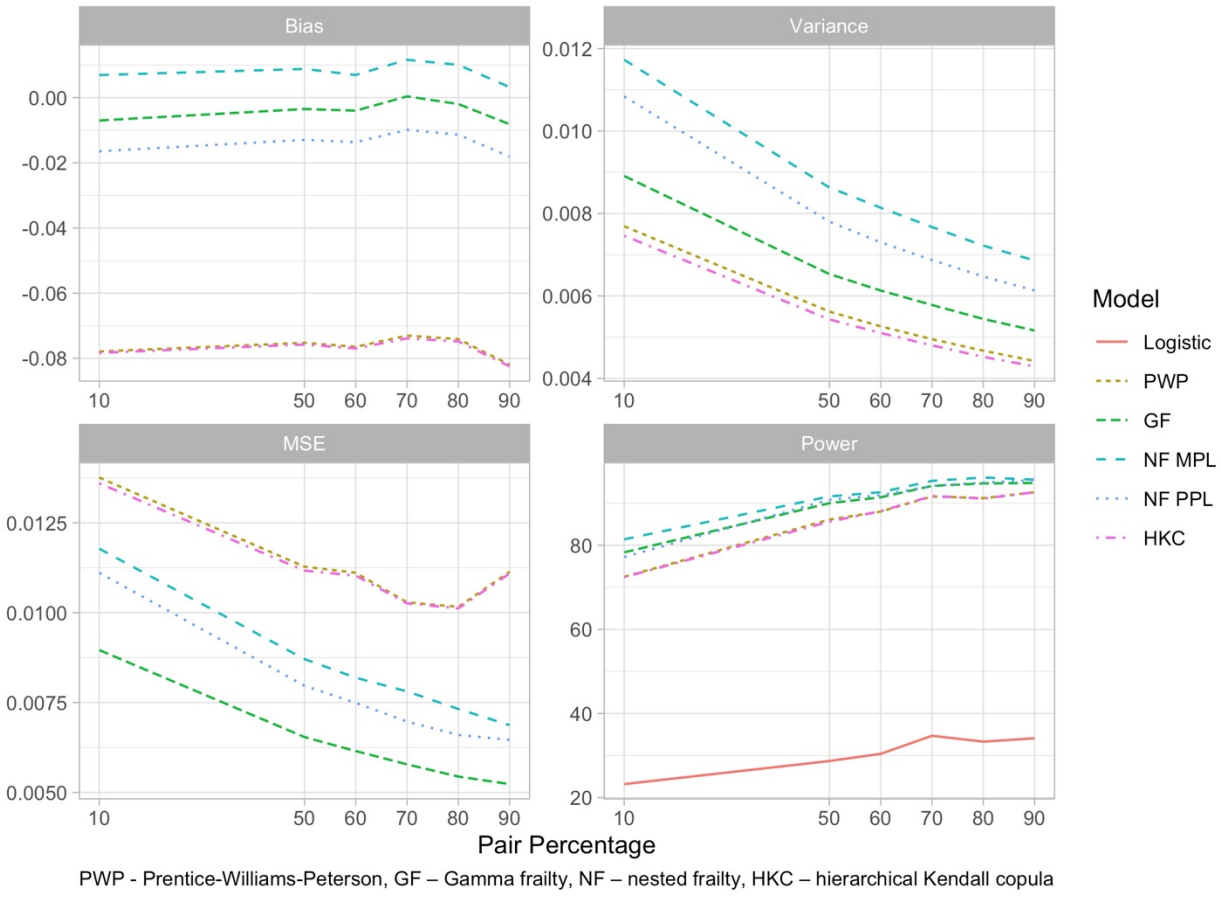

Figure 1. Line graphs for the bias, variance, MSE and power from all simulations across pair proportions for the treatment effect

For the parameter of ear ($\beta_2 = 0.4$), all models resulted in similar values of bias compared to what was observed for the estimate of treatment across all scenarios. However, it is worth highlighting that the nested MPL frailty model slightly overestimates the effect of ear, while the PWP, the HKC, nested PPL frailty and the gamma frailty model slightly underestimated the effect. The nested MPL frailty model had the lowest bias for estimating the parameter at 90%. For all scenarios, the PWP and the HKC resulted in the most biased estimates. From Figure 2 it can be seen that for the nested PPL frailty and the gamma frailty model the bias for ear effect decreases as the pair proportions go from 90% to 50%, and then slightly increase from 50% to 10%. The nested MPL frailty model shows a slight increase in the bias as the pair proportions move from 90% to 10%, while the HKC and PWP show a decrease in bias, though the change in bias remains minimal.

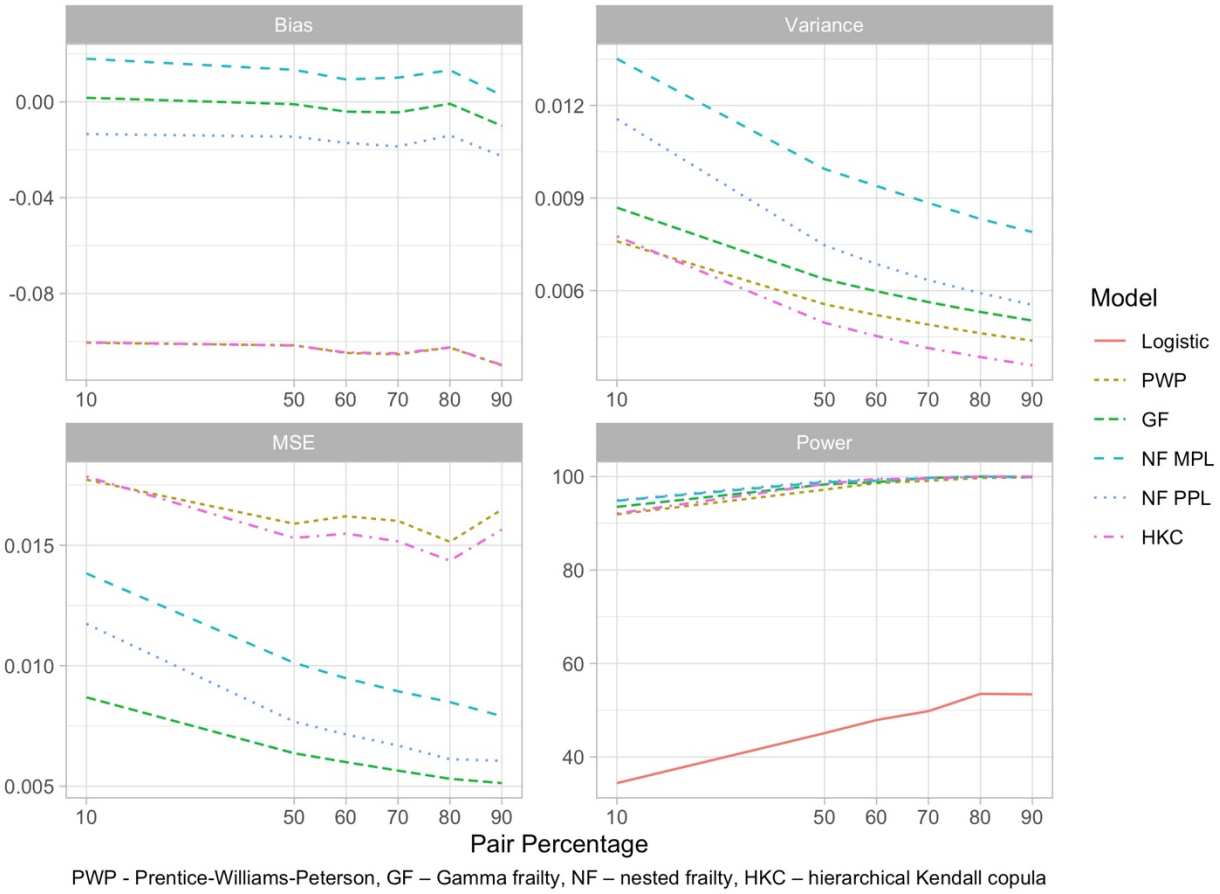

Figure 2. Line graphs for the bias, variance, MSE and power from all simulations across pair proportions for the ear effect.

Similar to the previous parameters, the interaction between treatment and ear ($\beta_3 = -0.25$) consistently had the largest bias when estimated by the PWP and the HKC model. The nested MPL frailty model resulted in the lowest bias across all five survival models. From Figure 3 it can be seen that the bias for the interaction between treatment and ear is fairly stable across pair proportions for the five survival models.

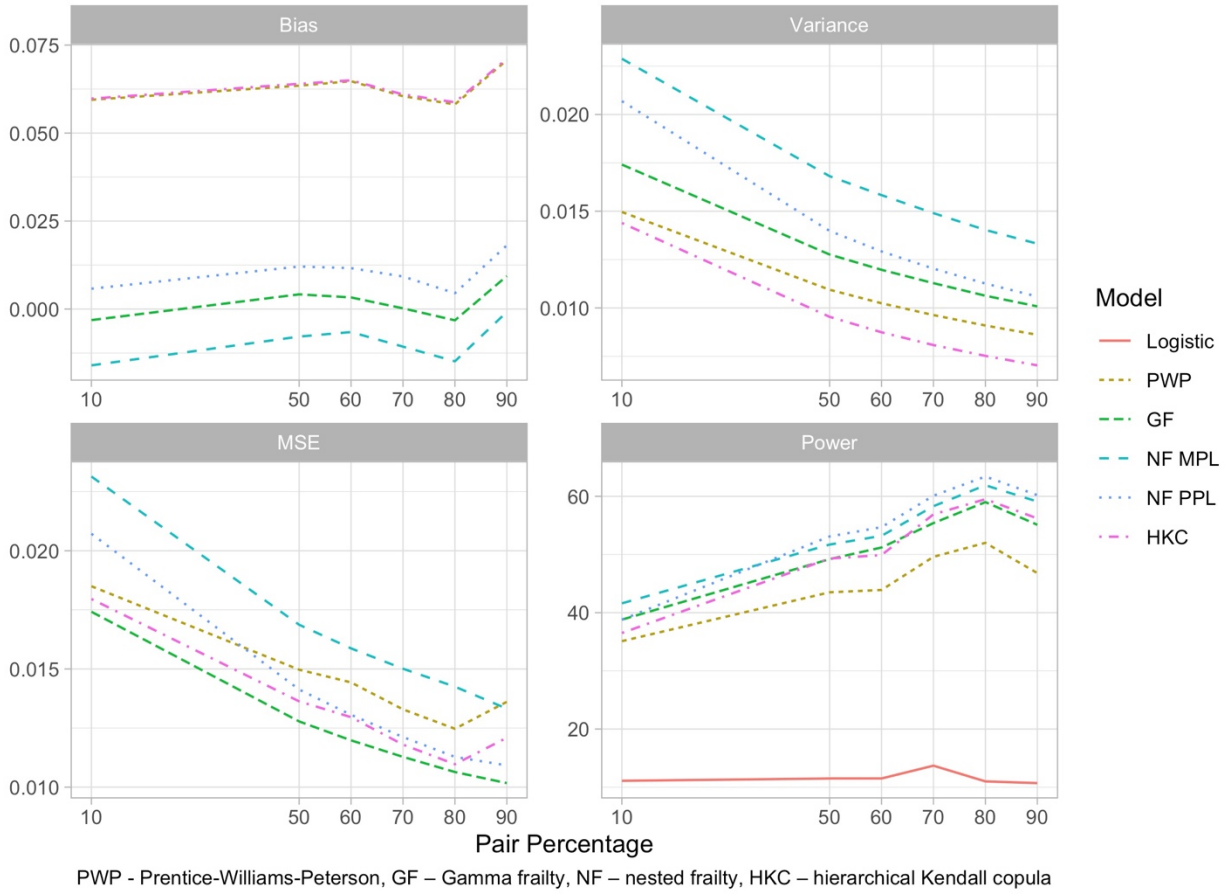

Figure 3. Line graphs for the bias, variance, MSE and power from all simulations across pair proportions for the interaction effect.

For the variance of the cluster parameter, the most accurate estimate was obtained from the HKC model and the nested PPL frailty model across all scenarios with the bias increasing as the pair proportion increased (Figure 4). The gamma frailty model and the nested MPL model stay contestant in bias across all the pair proportions. When looking at the bias for the variance of the sub-cluster term, the nested PPL frailty model most accurately estimates the variance. The HKC and the nested MPL frailty model result in greater bias across all pair proportions compared to the nested PPL frailty model. The nested MPL frailty model produced the same underestimate of the cluster variance throughout all pair proportions and consistently overestimated the sub-

cluster variance by approximately the amount of correlation in the clusters. All results were similar for the null simulations (Table 5).

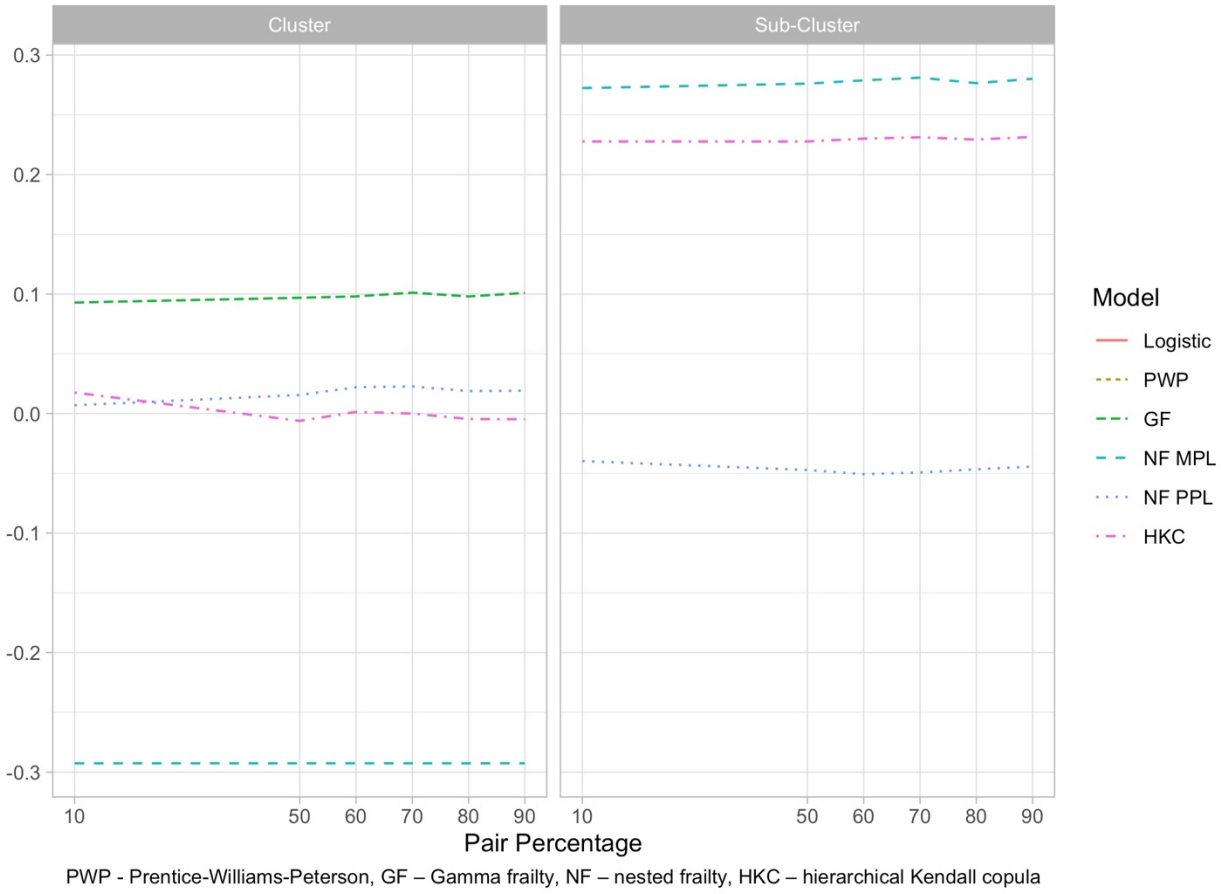

Figure 4. Line graphs for the bias from all simulations across pair proportions for the cluster terms.

3.1.2. Power

For the treatment variable (β_1) the nested MPL frailty model, PPL model and the gamma frailty model resulted in the highest power: 95.7%, 95.5% and 94.8%, respectively. The PWP and the HKC had slightly lower power than the three frailty models (Figure 1). However, the power obtained from the logistic model was much lower than the time to event models. All models saw a decrease in power as the pair proportions decreased from 90% to 10%, which is

expected since the overall information decreases with fewer pairs. The time to event models ranged from approximately 90-95% power with 90% pairs and decreased to 70-80% power with 10% pairs. The logistic regression model resulted in the lowest power ranging from 23% to 35% across proportions of pairs.

The parameter for ear (β_2) was estimated to have 100% power from 90% to 60% pairs for the five time to event models, with the power decreasing to the mid-nineties at 10% pairs. Similar to the treatment parameter, the power decreased for the ear parameter as the pair proportions decreased from 90% to 10%. For this parameter the logistic regression model had power slightly higher than the treatment parameter, ranging from 34% to 54%.

The interaction parameter (β_3) behaved differently from the treatment and ear parameter with respect to power. For all models the power increased as the pair proportion decreased from 90% to 80%. However, the power then decreased when the proportion of pairs fell from 80% to 10%. Overall, the two nested frailty models resulted in the highest power, followed by the HKC model, gamma frailty model, and the PWP model. The power for the interaction parameter was much lower than the other two parameters, ranging from 39% to 64%, which is not unexpected given the smaller effect size. Similar to the treatment parameter, the logistic regression model resulted in much lower power than the five time to event models, averaging around 10%.

3.1.3. Sensitivity scenarios

Additional simulations were conducted to examine (1) the cases of no cluster or subcluster variance, (2) no subcluster variance but the presence of cluster variance, and (3) a large subcluster variance and the same cluster variance at 90% pairs (Table 12 - 14). With respect to the treatment, ear, and interaction parameters, all five survival models estimated the effect similar to what was seen in the previously presented simulation study results.

When there is no sub-cluster variance the gamma frailty model underestimates the cluster correlation slightly, while the HKC model continues to overestimate the sub-cluster effect yet accurately estimates the cluster term. The nested MPL model resulted in a bias of 0.31 for the subcluster term, while resulting in a bias of -0.29 for the cluster term, appearing to account for the over estimation of the subcluster with the underestimation of the subcluster term. Lastly, the nested PPL model resulted in the closest estimation of the no variance sub-cluster term, with a bias of 0.008, while still accurately estimating the cluster term with a bias of 0.023.

When increasing the variance of the sub-cluster term there is a large amount of bias seen in the gamma frailty model, an amount similar to the subcluster variance. The nested MPL model underestimates the cluster term while accurately estimating the sub-cluster term. The nested PPL had the lowest bias for the cluster term, though more bias for the sub-cluster term compared to the nested MPL frailty model. Similarly, the HKC model produced bias for the cluster term and the most bias in estimating the sub-cluster term. In the sensitivity scenario with no cluster or subcluster variance, all models accurately estimated a zero variance for the sub-cluster and cluster term.

3.2. Application to motivating dataset

The six models were applied to the CHCO tympanoplasty dataset with all resulting in no significant estimates for treatment, ear, or their interaction. The estimates of the parameters for treatment, ear, interaction and the random effects are presented in Table 3. The estimates and standard errors were similar across the different models. The gamma frailty model and the HKC model produced the same estimates for treatment, ear, and the interaction, though the variance of the estimates differed by model. The variance of the cluster effect term from the gamma frailty model was reported to be 0.0000005, suggesting no cluster effect. However, from the simulation

studies we know that both the HKC and nested PPL model more accurately estimate the cluster term. The nested PPL frailty model reported a cluster variance of 0.080 which is larger than the reported cluster variance from the gamma frailty model, and the HKC model reported a variance of 0.082. The nested PPL frailty model reported a sub-cluster variance of 0.075, suggesting the potential for substantial effects at both the cluster and sub-cluster levels. This also demonstrates the importance of model selection in our context with a large number of clusters but small sub-cluster sizes with potentially repeated events, where the estimated beta coefficients may not have been significantly different but the correlation estimates were. Therefore, if the purpose was to evaluate potential correlation estimates, the chosen model could result in drastically different conclusions.

Table 3. Parameter estimates, Variance, P-value results the Children's Hospital of Colorado Time to Tympanoplasty Failure dataset.

Parameter	Model	Estimate	(95% CI)	Variance	P-Value
β_1 Treatment					
	Logistic	-0.140	(-0.664, 0.380)	0.266	0.599
	PWP	-0.132	(-0.619, 0.356)	0.241	0.597
	GF	-0.128	(-0.598, 0.342)	0.240	0.590
	NF MPL	-0.118	(-0.537, 0.301)	0.214	0.581
	NF PPL	-0.121	(-0.589, 0.347)	0.239	0.601
	HKC	-0.128	(-0.618, 0.362)	0.24	0.608
β_2 Ear					
	Logistic	-0.235	(-0.763, 0.286)	0.267	0.378
	PWP	-0.209	(-0.709, 0.290)	0.242	0.412
	GF	-0.207	(-0.681, 0.268)	0.242	0.390
	NF MPL	-0.206	(-0.626, 0.215)	0.215	0.388
	NF PPL	-0.208	(-0.719, 0.303)	0.261	0.359
	HKC	-0.207	(-0.707, 0.293)	0.242	0.418
β_3 Interaction					
	Logistic	-0.067	(-0.837, 0.702)	0.392	0.865
	PWP	-0.078	(-0.816, 0.660)	0.358	0.836
	GF	-0.081	(-0.783, 0.621)	0.358	0.820

	NF MPL	-0.095	(-0.772, 0.582)	0.345	0.783
	NF PPL	-0.085	(-0.778, 0.608)	0.354	0.799
	HKC	-0.081	(-0.818, 0.657)	0.358	0.830
ρ Cluster					
	GF	0.0000005			
	NF MPL	0.071			
	NF PPL	0.080			
	HKC	0.082			
η Sub-Cluster					
	NF MPL	0.184			
	NF PPL	0.075			
	HKC	0.139			

4. DISCUSSION

4.1.Model Evaluation

All of the survival models resulted in relatively tight confidence intervals for bias, and all included zero, which may be a bit unexpected across all models given our simulated data generating mechanism assumed a nested MPL frailty model. Given that none of the models had been previously explored with respect to large numbers of clusters with either one or two sub-clusters and a maximum of 2 observations per sub-cluster, it was not apparent *a priori* which model would best perform in this context.

The logistic regression model performed the worst out of all six models given the low power experienced over the pair proportions ranging from 50% - 10%. The type-1 error for the logistic regression model is considerably lower compared to other models, ranging from at most 5.6% to as low as 1.2%. The PWP model had the lowest power out of all the survival models evaluated for all parameters, which might be a reflection of the models' inability to account for the correlation structure. Out of all of the survival models considered, interestingly the gamma frailty model resulted in the least biased estimates of the parameters, surpassing that of the nested MPL frailty model from which the data were simulated. The bias of the cluster term for the gamma frailty model is consistently similar to the amount of correlation of the subcluster, which is not accounted for in the model. This suggests that the frailty term in this model is combining both the sub-cluster and cluster term in its estimation of the cluster frailty term.

Though the nested MPL frailty model did produce unbiased estimates with high power, the simulated datasets assumed a nested MPL frailty model, and therefore one would expect the nested MPL model to perform very well. The sub-cluster and cluster variances were, however,

still biased in our setting with small sub-cluster sizes. Simulating data from a different model than the nested MPL frailty may lead to different conclusions than the ones we observed.

A limitation of the nested MPL frailty model is that it did not always converge for our simulated datasets. At 90% pairs, 95.9% of the 1,000 models converged, with the percentage of convergence decreasing as the percentage of pairs in the dataset decreased to 10% pairs, where only 90.7% of the 1,000 models converged. In practice, the starting values for each nested MPL frailty model could be optimized to obtain convergence, but it was kept consistent throughout simulations to evaluate this potential model fitting challenge.

The nested PPL frailty model similarly estimates the model parameters compared to the nested MPL frailty model with generally low bias and high power relative to the other methods considered. The major difference is with respect to the bias in estimating the hierarchical correlation structure, the nested PPL frailty model resulted in the least amount of bias. This suggests that if the primary motivation is the estimation of the hierarchical correlation structure or both the hierarchical correlation structure and the fixed regression coefficients one could use this model without issues. One practical consideration for implementation is that a single iteration of the model ranged widely from a few minutes to over an hour in our simulation studies, so it is important to be aware that model convergence may take some time relative to the less complex methods described.

The hierarchical Kendall copula first fits a marginal model that accounts for one level of clustering similar to the PWP model, and therefore the resulting bias from the two models are very similar, which can be seen throughout all scenarios. The hierarchical Kendall copula had cluster variance estimate similar to the gamma frailty model and seemed to consistently estimate little to no sub-cluster variance. Given the increase in model complexity for our motivating

context, we would not recommend the hierarchical Kendall copula model relative to other modeling strategies.

4.2. Application to real world data

Based on the results from the simulation we expected the coefficients from the gamma frailty model to be the least biased when estimated by any of the survival models. We note very similar results from the CHCO time to tympanoplasty failure dataset across all models in our application. Although through simulation we expected the logistic regression model to result in either overestimation or underestimation, it had a similar effect size as well (albeit with respect to the odds ratio instead of the hazard ratio). When looking at the variance of the cluster parameters we observe that the nested MPL model resulted in a cluster variance term smaller than the nested PPL and HKC model, but the variance was reported to be nearly zero in the gamma frailty model. It can be seen in the CHCO dataset, the variance of the sub-cluster term is reported to be higher with the hierarchical Kendall copula model and the nested MPL model, where the sub-cluster variance is lower with the nested PPL model. Overall, the results across different models from the CHCO dataset agree with our simulation findings. We therefore would recommend the use of reporting the nested PPL frailty model for this dataset.

4.3.Limitations

There are a few important limitations to note with our simulation study. The first is that our simulations only consider a maximum of two possible events and two treatment options, but in practice the methods and assessment can be extended to more possible events and more treatment options. It may be important to consider further simulation studies that allow additional repeated events, although in our context this was uncommon. Second, other statistical models exist to evaluate the data, however we focused on methods that were either previously applied to

similar data or were more easily accessible to clinical research teams through software such as R or SAS.

4.4.Concluding remarks

Choosing the appropriate model in the presence of hierarchical nested correlation time to event data is challenging and only more recently have methods been developed that account for multiple levels of clustering. However, if clusters only have one or two sub-clusters with a maximum of two events, the challenge is to find a model that accurately reports the regression coefficient estimates along with the hierarchical structure of correlation between the cluster and sub-cluster. The approach of using the nested PPL frailty model to obtain fixed effect parameter estimates and the estimation of the hierarchical correlation were shown to be the most accurate representation of nested data based on our simulation results. If only estimation of the fixed effects terms is of interest, we note the gamma frailty model resulted in similar estimates to more complex models without convergence issues and may be considered in place of the nested PPL frailty model.

Table 1. Bias, Variance, Mean Squared Error (MSE) and Power results for the simulated scenario with 90% of clusters having two sub-clusters or pairs.

Parameter	Model	Bias	(95% CI)	Variance	MSE	Power
β_1 Treatment = 0.3						
	Logistic					34.1%
	PWP	-0.082	(-0.210, 0.046)	0.004	0.011	92.6%
	GF	-0.008	(-0.170, 0.154)	0.005	0.005	94.8%
	NF MPL[a]	0.003	(-0.161, 0.168)	0.007	0.007	95.6%
	NF PPL	-0.018	(-0.171, 0.135)	0.006	0.006	95.5%
	HKC	-0.082	(-0.210, 0.045)	0.004	0.011	92.6%
β_2 Ear = 0.4						
	Logistic					53.4%
	PWP	-0.110	(-0.226, 0.006)	0.004	0.016	99.8%
	GF	-0.010	(-0.164, 0.144)	0.005	0.005	99.9%
	NF MPL[a]	0.003	(-0.153, 0.159)	0.008	0.008	99.9%
	NF PPL	-0.023	(-0.169, 0.124)	0.006	0.006	100%
	HKC	-0.110	(-0.225, 0.006)	0.004	0.016	100%
β_3 Interaction = -0.25						
	Logistic					10.7%
	PWP	0.071	(-0.088, 0.229)	0.009	0.014	46.8%
	GF	0.009	(-0.193, 0.212)	0.010	0.010	55.1%
	NF MPL[a]	-0.001	(-0.206, 0.204)	0.013	0.013	59.1%
	NF PPL	0.018	(-0.177, 0.213)	0.011	0.011	60.2%
	HKC	0.071	(-0.087, 0.229)	0.007	0.012	56.2%
ρ Cluster = 0.3						
	GF	0.101	(-0.048, 0.250)			
	NF MPL[a]	-0.293	(-0.293, -0.293)			
	NF PPL	0.019	(-0.087, 0.125)			
	HKC	-0.005	(-0.108, 0.099)			
η Sub-Cluster = 0.15						
	NF MPL[a]	0.280	(0.136, 0.424)			
	NF PPL	-0.044	(-0.158, 0.070)			
	HKC	0.231	(0.111, 0.352)			

[a]NF MPL model includes 959 simulations; all other models include 1,000 simulations. GF = Gamma frailty, NF = nested frailty, MPL = maximum penalized likelihood estimation, PPL = penalized partial likelihood estimation, HKC = hierarchical Kendall copula

Table 2. Bias, Variance, Mean Squared Error (MSE) and Power results for the simulated scenario with 50% of clusters having two sub-clusters or pairs.

Parameter	Model	Bias	(95% CI)	Variance	MSE	Power
β_1 Treatment = 0.3						
	Logistic					28.7%
	PWP	-0.075	(-0.224, 0.074)	0.006	0.011	86.1%
	GF	-0.004	(-0.187, 0.180)	0.007	0.007	90.0%
	NF MPL[a]	0.009	(-0.176, 0.194)	0.009	0.009	91.6%
	NF PPL	-0.013	(-0.188, 0.163)	0.008	0.008	90.8%
	HKC	-0.076	(-0.225, 0.073)	0.005	0.011	85.6%
β_2 Ear = 0.4						
	Logistic					45.1%
	PWP	-0.102	(-0.244, 0.041)	0.006	0.016	97.2%
	GF	-0.001	(-0.189, 0.187)	0.006	0.006	98.3%
	NF MPL[a]	0.013	(-0.177, 0.203)	0.010	0.010	98.8%
	NF PPL	-0.015	(-0.191, 0.162)	0.007	0.008	99.0%
	HKC	-0.102	(-0.244, 0.040)	0.005	0.015	98.4%
β_3 Interaction = -0.25						
	Logistic					11.5%
	PWP	0.063	(-0.133, 0.260)	0.011	0.015	43.5%
	GF	0.004	(-0.241, 0.249)	0.013	0.013	49.2%
	NF MPL[a]	-0.008	(-0.255, 0.240)	0.017	0.017	51.7%
	NF PPL	0.012	(-0.223, 0.248)	0.014	0.015	53.1%
	HKC	0.064	(-0.132, 0.260)	0.010	0.014	49.3%
ρ Cluster = 0.3						
	GF	0.097	(-0.060, 0.254)			
	NF MPL[a]	-0.293	(-0.293, -0.293)			
	NF PPL	0.016	(-0.117, 0.148)			
	HKC	-0.006	(-0.140, 0.128)			
η Sub-Cluster = 0.15						
	NF MPL[a]	0.280	(0.136, 0.424)			
	NF PPL	-0.044	(-0.158, 0.070)			
	HKC	0.231	(0.111, 0.352)			

[a]NF MPL includes 937 simulations, all other models include 1,000 simulations. GF = Gamma frailty, NF = nested frailty, MPL = maximum penalized likelihood estimation, PPL = penalized partial likelihood estimation, HKC = hierarchical Kendall copula

The authors report there are no competing interests to declare.

REFERENCES

Amorim, L. D., & Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. International journal of epidemiology, 44(1), 324-333.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The annals of statistics, 1100-1120.

Baklaci, D., Guler, I., Kuzucu, I., Kum, R. O., & Ozcan, M. (2018). Type 1 tympanoplasty in pediatric patients: a review of 102 cases. BMC pediatrics, 18(1), 1-6.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. 'Fitting linear mixed-effects models using lme4', *arXiv preprint arXiv:1406.5823*.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, 24(11), 1713-1723.

Cass, N. D., Patten, L., & Cass, S. P. (2019). Collagen Allografts Compared With Autologous Tissue in Tympanoplasty. Otology & Neurotology, 40(6), 767-771.

Cook, R. J., & Lawless, J. (2007). The statistical analysis of recurrent events. Springer Science & Business Media.

Fairbairn, C. E. (2016). A nested frailty survival approach for analyzing small group behavioral observation data. Small group research, 47(3), 303-332.

Kessler, A., Potsic, W. P., & Marsh, R. R. (1994). Type 1 tympanoplasty in children. Archives of Otolaryngology–Head & Neck Surgery, 120(5), 487-490.

Meyer, R., & Romeo, J. S. (2015). Bayesian semiparametric analysis of recurrent failure time data using copulas. Biometrical Journal, 57(6), 982-1001.

Patil, K., Baisakhiya, N., & Deshmukh, P. T. (2014). Evaluation of different graft material in type 1 tympanoplasty. Indian Journal of Otology, 20(3), 106.

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. Biometrika, 68(2), 373-379.

Rondeau, V., Filleul, L., & Joly, P. (2006). Nested frailty models using maximum penalized likelihood estimation. Statistics in medicine, 25(23), 4036-4052.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. Journal of the American Statistical Association, 92(438), 426-435.

Schukken, Y. H., Bar, D., Hertl, J., & Gröhn, Y. T. (2010). Correlated time to event data: Modeling repeated clinical mastitis data from dairy cattle in New York State. Preventive veterinary medicine, 97(3-4), 150-156.

Sklar, A. (1959). N-dimensional distribution functions and their margins. Publications of the Statistical Institute of the University of Paris, 8, 229-231.

Spiekerman, C. F., & Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. Journal of the American Statistical Association, 93(443), 1164-1175.

Su, C. L., Nešlehová, J. G., & Wang, W. (2019). Modelling hierarchical clustered censored data with the hierarchical Kendall copula. Canadian Journal of Statistics, 47(2), 182-203.

The Coxme package (2019). The Comprehensive R Archive Network, https://cran.r-project.org/web/packages/coxme/(accessed August 2019)

The frailtypack package (2019). The Comprehensive R Archive Network, https://cran.r-project.org/web/packages/frailtypack/ (accessed December 2019)

Therneau, T. (2020a). Coxme and the Laplace approximation. In Mayo Clinic.

Therneau, T. (2020b). A package for survival analysis in R. R package version 3.1-12.