Restricted mean survival time to estimate an intervention effect in a cluster randomized trial

Floriane Le Vilain-Abraham, Elsa Tavernier, Etienne Dantan, Solène Desmée, Agnès Caille

Restricted mean survival time to estimate an intervention effect in a cluster randomized trial

Floriane Le Vilain--Abraham[1], Elsa Tavernier[1,2], Etienne Dantan[3], Solène Desmée[1], Agnès Caille[1,2]

1 Univ Tours, Nantes Université, INSERM, MethodS in Patients-centered outcomes and HEalth ResEarch, SPHERE, Tours, France

2 INSERM CIC1415, CHRU de Tours, Tours, France

3 Nantes Université, Univ Tours, INSERM, MethodS in Patients-centered outcomes and HEalth ResEarch, SPHERE, Nantes, France

Abstract

For time-to-event outcomes, the difference in restricted mean survival time (RMST) is a measure of the intervention effect, an alternative to the hazard ratio, corresponding to the expected survival duration gain due to the intervention up to a predefined time t^*. We extended two existing approaches of RMST estimation for independent data to clustered data in the framework of cluster randomized trials (CRTs): one based on the direct integration of Kaplan-Meier curves and the other based on pseudo-values regression. Then, we conducted a simulation study to assess and compare the statistical performance of the proposed methods, varying the number and size of clusters, the degree of clustering and the magnitude of the intervention effect under proportional and non-proportional hazards assumption. We found that the extended methods well estimated the variance and controlled the type I error if there was a sufficient number of clusters ($\geq$50) under both proportional and non-proportional hazards assumption. For CRTs with a limited number of clusters (<50), a permutation test for pseudo-values regression was implemented and corrected the type I error. *We also provided a procedure to estimate permutation-based confidence intervals which produced*

adequate coverage. All the extended methods performed similarly, but the pseudo-values regression offered the possibility to adjust for covariates. Finally, we illustrated each considered method with a CRT evaluating the effectiveness of an asthma-control education program.

Keywords

Cluster randomized trial; time-to-event outcome; restricted mean survival time; pseudo-values; Kaplan-Meier estimate

1. Introduction

Cluster randomized trials (CRTs) are trials in which intact social units, such as hospitals, geographical areas or medical practices, are randomized between intervention and control groups.[1] This study design has rapidly spread in the field of health research. CRTs are well suited to evaluate an intervention delivered at the cluster level and to minimize the potential risk of between-group contamination.[2]

In CRTs, outcomes assessed on individuals within a given cluster tend to be more similar than outcomes on individuals between different clusters. Statistical analysis must account for the clustering induced by such correlated data to avoid an underestimation of the variance leading to increased risk of type I error when estimating the intervention effect.[3] Most of the methods developed to account for clustering in CRTs address continuous or binary outcomes, with less interest in time-to-event outcomes.[4] However, a recent review found that time-to-event outcomes are not uncommon in CRTs but concluded that appropriate analysis methods are infrequently used.[4]

In addition to clustering, specificities of time-to-event data such as right-censoring must be accounted for. In individually randomized controlled trials with time-to-event outcomes, the intervention effect is usually estimated by a hazard ratio (HR) obtained from classical survival models such as Cox models.[5] In the specific context of CRTs, the HR can be estimated by using a conditional approach accounting for clustering using frailty terms or a marginal approach with a sandwich variance estimator.[6,7]

Although the HR has a long history in survival analysis,[8-10] it may be difficult to interpret clinically, especially when the proportional hazards assumption is violated.[11-15] As mentioned in Uno et al.[14], the HR is not a probability measure or a relative risk but

rather a ratio of hazard rates. Therefore, the magnitude of the intervention benefit could be difficult to assess. In addition, its causal interpretation has been questioned.[16,17] The restricted mean survival time (RMST) up to time t^*, defined as the average survival time up to t^*, does not require a proportional hazards assumption[18] and has a causal effect interpretation.[19] Royston et al.[20] recently proposed the use of the difference in RMST to design and analyze clinical trials. The difference in RMST between the intervention and control groups could be an alternative measure of the intervention effect.

Several methods have been developed to estimate a difference in RMST for independent data such as the direct integration of the Kaplan-Meier survival curves[21], Andersen's pseudo-values regression method[22-24], Royston and Parmar's flexible parametric survival model[15,20] as well as methods based on the inverse probability of censoring weighting accounting for covariate-dependent censoring [25-27]. Recently, Chen et al.[28] adapted to clustered data a method initially proposed by Zhong and Schaubel[27] for independent data that estimates the RMST as a continuous function of the horizon time t^* allowing for covariate-dependent censoring. To our knowledge, no other RMST estimation approach has been extended to clustered data.

Our objective was to propose two original and easy-to-use approaches to estimate a difference in RMST for clustered data in CRTs. We extended two approaches developed for independent data to clustered data: direct integration of Kaplan-Meier curves and pseudo-values regression. Section 2 presents the RMST and details the Kaplan-Meier and pseudo-values regression methods for estimating a difference in RMST in case of independent data, then the extensions of these approaches to clustered time-to-event data. Section 3 reports the design and results of a simulation study to

assess and compare the statistical performance of the proposed methods. Section 4 illustrates the proposed methods with an application to a CRT evaluating an asthma-control education program. Section 5 concludes with a discussion.

2. Methods

2.1. Definition of RMST

Let T denote a random variable representing the time to an event and t^* a specific time horizon. The restricted mean survival time up to t^* is defined as the expectation of $\min(T, t^*)$ and corresponds to the area under the survival function, $S(\cdot)$, from 0 to t^*.

$$RMST(t^*) = E[\min(T, t^*)] = \int_0^{t^*} S(t)\, dt \qquad (1)$$

In a randomized trial with two parallel groups, the intervention effect could be summarized by a difference in RMST up to t^* ($\Delta RMST(t^*)$) between the intervention and control groups.

$$\Delta RMST(t^*) = RMST_1(t^*) - RMST_0(t^*) = \int_0^{t^*} S_1(t)\, dt - \int_0^{t^*} S_0(t)\, dt \qquad (2)$$

where $RMST_j(t^*)$ and $S_j(\cdot)$ are the RMST up to t^* and the survival function in the group j (j = 0 for the control group, 1 for the intervention group), respectively. The $\Delta RMST(t^*)$ is easily interpretable. For example, if the $\Delta RMST(t^*)$ equals one month for individuals followed up over 12 months, we may expect an increased life expectancy of one month over the first follow-up year for individuals in the intervention group versus the control group.

2.2. Methods of estimating $\Delta RMST(t^*)$ for independent data

2.2.1. Notations

Let X_l be the observed time defined as the time between the origin and the occurrence of the survival event or censoring (i.e., $X_l = \min(T_l, C_l)$, where T_l is the true time-to-event and C_l the right-censoring time for individual l). Thereafter, we assume that the event times are independent of the censoring times. The event indicator, δ_l, equals 1 if $T_l \leq C_l$ and 0 otherwise. Thus, for inference, we consider a sample of n independent and identically distributed individuals with data $\{X_l, \delta_l, Z_l; l = 1, \dots, n\}$, with the intervention indicator Z_l = 1 if the individual l received the intervention and 0 if the individual is in the control group.

2.2.2. Kaplan-Meier-based method for independent data (KM_{indep})

One commonly used method to estimate the RMST is the direct integration of Kaplan-Meier survival curves.[21] The survival function $S(\cdot)$ in equation (1) is replaced by the Kaplan-Meier estimate, denoted $\hat{S}(\cdot)$:

$$\widehat{RMST}(t^*) = \int_0^{t^*} \hat{S}(t)\, dt = \sum_{i=0}^{E} (t_{i+1} - t_i)\hat{S}(t_i) \qquad (3)$$

where $t_1, \dots, t_E$ are the E distinct event times before t^*, $t_0 = 0$ and $t_{E+1} = t^*$. The variance of $\widehat{RMST}(t^*)$ may be estimated by using the Greenwood formula.[29]

The $\Delta RMST(t^*)$ is estimated by the difference between the RMSTs estimated in each group. The variance of $\Delta\widehat{RMST}(t^*)$ is the sum of the variances of the estimated RMST in each group, and the $100(1-\alpha)$% confidence interval (CI) is estimated as:

$$\Delta\widehat{RMST}(t^*) \pm z_{1-\frac{\alpha}{2}}\sqrt{Var\left(\widehat{RMST}_1(t^*)\right) + Var\left(\widehat{RMST}_0(t^*)\right)} \qquad (4)$$

where z_α is the α-quantile of the standard normal distribution.[29]

2.2.3. Pseudo-values regression-based method for independent data (PV_{indep})

Andersen et al.[22,23] proposed a pseudo-values approach to estimate the $\Delta RMST(t^*)$. If the time-to-events were not censored, T_l $(l = 1, \dots, n)$ would be observed and $\min(T_l, t^*)$ would be available for each individual *l*. The following generalized linear regression model could be used to specify how $E\left[\min(T_l, t^*) | \widetilde{\boldsymbol{Z}}_{\boldsymbol{l}}\right]$ depends on $\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}}$:

$$g\left(E\left[\min(T_l, t^*) \mid \widetilde{\boldsymbol{Z}}_l\right]\right) = \boldsymbol{\beta}' \, \widetilde{\boldsymbol{Z}}_{\boldsymbol{l}} \quad (5)$$

where $\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}} = (1, Z_l)'$; $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, where β_0 is the intercept coefficient and β_1 is the intervention effect coefficient; and $g(.)$ is a link function. However, T_l is not observed for all individuals because of right-censoring and standard methods cannot be used.

The pseudo-values approach aims at replacing $\min(T_l, t^*)$ by a pseudo-value $\widehat{Y}_l$ for each individual *l*. Let $\widehat{RMST}(t^*)$ be the Kaplan-Meier-based estimator described in the previous section in equation (3), that is an approximately unbiased estimator for $RMST(t^*)$. The pseudo-value for the individual *l* is defined as

$$\widehat{Y}_l = n \cdot \widehat{RMST}(t^*) - (n-1)\, \widehat{RMST}^{-l}(t^*) = n \int_0^{t^*} \hat{S}(t)\, dt - (n-1) \int_0^{t^*} \hat{S}^{-l}(t)\, dt \quad (6)$$

where $\widehat{RMST}(t^*)$ is the Kaplan-Meier–based estimator for the entire sample and $\widehat{RMST}^{-l}(t^*)$ is the leave-one-out Kaplan-Meier–based estimator (i.e., the estimator for the dataset of size $n-1$ without the individual *l*). The pseudo-value $\widehat{Y}_l$ is then used as the response in the generalized linear regression model (equation (5)). Of note, in equation (5) we only consider intervention group as a cluster-level covariate, but we could adjust for any additional cluster or individual-level covariates by adding them into $\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}}$. Thereafter, we used an identity link function, usually used for RMST.[22,23]

The regression coefficients may be estimated by using the generalized estimating equation (GEE), and the variance for $\widehat{\boldsymbol{\beta}}$ could be estimated by using the sandwich estimator detailed in Andersen et al.[22] and provided in Web Appendix A.1. The $\Delta\widehat{RMST}(t^*)$ corresponds to $\widehat{\beta_1}$ and the associated 100(1-α)% CI is calculated as $\widehat{\beta_1} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\widehat{\beta_1})}$, where $\widehat{Var}(\widehat{\beta_1})$ is the robust sandwich variance of the regression coefficient associated with the intervention group.

2.3. Methods of estimating $\Delta RMST(t^*)$: extension to clustered data (CRT framework)

2.3.1. Notations

In this section, we extend the two previous methods for independent data to the clustered data framework. We now consider a sample of K clusters of size m_k ($k=1,\ldots,K$) with K_0 and K_1 clusters in the control and intervention groups, respectively. Then, we consider $n = \sum_{k=1}^{K} m_k$ individuals with the data $\{X_{kl}, \delta_{kl}, Z_{kl};, l = 1, \ldots, m_k, k = 1, \ldots, K\}$, where X_{kl} is the observed survival time, δ_{kl} the event indicator and Z_{kl} the intervention indicator for individual l from cluster k.

2.3.2. Kaplan-Meier-based method: Extension to clustered data (KM_{clust})

We extended the Kaplan-Meier-based method to the context of CRTs. The point estimate for RMST in the control and intervention groups was obtained with equation (3) because the Kaplan-Meier estimator is consistent for clustered data.[30] However, the variance needs to be adapted,[30] and thus we estimated the variance of $\Delta RMST(t^*)$ by using non-parametric bootstrap.[31] The bootstrap variance was obtained with 10 000 replications of one-stage resampling of K entire clusters with replacement, with K_0 and

K_1 clusters in the control and intervention groups, respectively. All individuals from a selected cluster were included without resampling because Ren et al.[32] showed that sampling only the highest level is the best sampling method for clustered data. The 100(1-α)% CI was estimated by using non-parametric percentiles bootstrap.[31]

2.3.3. Pseudo-values regression-based method: Extension to clustered data with an exchangeable working correlation matrix (PV_{ECM})

To extend the pseudo-values regression-based method to clustered time-to-event data in the context of CRTs, we computed the pseudo-values $\hat{Y}_{kl}$ for individual *l* in cluster *k* by using equation (6). The same generalized linear regression model specified in equation (5) was considered. The regression coefficients were still estimated by using GEE with a robust sandwich variance estimator but with a covariance matrix reflecting the correlation of the outcomes of individuals within a cluster.[33]

Let denote $\widehat{\boldsymbol{Y}}_k = (\hat{Y}_{k1}, \dots, \hat{Y}_{km_k})$ and $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{km_k})$, where $\mu_{kl} = g^{-1}(\boldsymbol{\beta}'\widetilde{\boldsymbol{Z}}_{kl})$. The regression coefficients were estimated from the following GEE:[33,34]

$$\boldsymbol{U}(\beta) = \sum_{k=1}^{K} \boldsymbol{U}_k(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}\right)' \boldsymbol{V}_k^{-1}\left(\widehat{\boldsymbol{Y}_k} - \boldsymbol{\mu}_k\right) = 0 \qquad (7)$$

where $\boldsymbol{V}_k$ is an $m_k \times m_k$ working covariance matrix for cluster *k*. It is defined as $\boldsymbol{V}_k = \phi \boldsymbol{A}_k^{\frac{1}{2}} \boldsymbol{R}(\rho) \boldsymbol{A}_k^{\frac{1}{2}}$, where ϕ is a dispersion parameter, $\boldsymbol{R}(\rho)$ is an $m_k \times m_k$ working correlation matrix and $\boldsymbol{A}_k$ an $m_k \times m_k$ diagonal matrix with $Var(\mu_{kl})$ as diagonal elements.[33,34] In CRTs, an exchangeable working correlation matrix is usually selected, assuming that the correlation for all pairs of outcomes within a cluster are identical and

common across all clusters: $\boldsymbol{R}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$ where, $\rho = Corr\left(\hat{Y}_{kl}, \hat{Y}_{ki}\right) \forall k, i \neq l.$[2] The estimations of ρ and ϕ is detailed in Appendix A.2. The variance of the regression coefficient estimates was estimated with the following sandwich estimator:[33]

$$\boldsymbol{\Sigma} = \boldsymbol{I}\left(\widehat{\boldsymbol{\beta}}\right)^{-1} \widehat{var}\left(U\left(\widehat{\boldsymbol{\beta}}\right)\right) \boldsymbol{I}\left(\widehat{\boldsymbol{\beta}}\right)^{-1} \quad (8)$$

where

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}\right)' \boldsymbol{V}_k^{-1} \left(\frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}\right), \quad (9)$$

$$\widehat{var}\left(\boldsymbol{U}\left(\widehat{\boldsymbol{\beta}}\right)\right) = \sum_{k=1}^{K} \boldsymbol{U}_k\left(\widehat{\boldsymbol{\beta}}\right) \boldsymbol{U}_k\left(\widehat{\boldsymbol{\beta}}\right)' \quad (10)$$

By accounting for clustering in equation (8), the variance estimator obtained with this method tends to be larger than that obtained for independent data with the PV_{Indep} method. This will avoid the underestimation of the variance when ignoring the clustering, which results in too-narrow CIs and thus inflated type I error rate.

2.3.4. Pseudo-values regression-based method: Extension to clustered data (CRT framework) with an independent working correlation matrix (PV_{ICM})

Although an exchangeable working correlation structure is usually assumed in CRTs, it can produce non-convergence when the number of clusters is limited.[35,36] Thus, we explored an alternative independent working correlation structure because GEE estimates have been shown to be consistent even when the correlation structure is misspecified.[34] Moreover, some previous work obtained satisfactory results with

pseudo-values regression methods for median survival time and cumulative incidence function estimations for clustered data with an independent correlation structure.[37,38] With PV_{ICM}, the correlation matrix, $\boldsymbol{R}(\alpha)$, is the identity matrix. The GEE and sandwich variance estimators are the same as those defined in equations (7) and (8) substituting the independent correlation matrix for the exchangeable correlation matrix.

2.4. Permutation test for pseudo-values regression

The difference in RMST obtained from the pseudo-values regression relies on a GEE approach, which can lead to inflated type I error rate when the number of clusters is small and the normality-based Wald test is used for inference.[39] Several methodological reviews indicated that CRTs often randomized fewer than 40 clusters,[40-42] so a permutation test could be relevant to control the type I error rate when estimating the $\Delta RMST(t^*)$ with pseudo-values regression. Permutation tests have been studied in CRTs for binary and continuous outcomes[43-46] and for time-to-event outcomes.[47,48]

In practice, we first performed the pseudo-values regression on the observed dataset and estimated the Wald-test statistic, equal to $\widehat{\beta_1}/\widehat{Var}(\widehat{\beta_1})$. Second, we randomly permuted the allocation of the clusters to the control and intervention groups. We estimated a Wald-test statistic for each permutation. The p-value for the null hypothesis of no intervention effect was estimated as the proportion of permutations with a test statistic equal to or greater than the test statistic from the observed dataset.

Complementarily, we also proposed the CIs estimation by inverting the permutation test. The 100(1-α)% permutation-based CI was obtained by searching the set of values $\beta^* \neq 0$ such as the null hypothesis $H_0: \beta_1 = \beta^*$ is not rejected. This approach is computationally challenging, since it requires to carry out a large number of

permutation tests. To reduce the computation time, we used the efficient method proposed by Rabideau and Wang[49] and Garthwaite[50], detailed in Appendix A.3.

2.5. Implementation

All analyses were performed with R v4.0.0.[51] We used the *rmst2* function from the R package "survRM2" to estimate the $\Delta RMST(t^*)$ using the direct integration of the Kaplan-Meier survival curves. We computed the pseudo-values with the *pseudomean* function from the R package "pseudo" and fitted the generalized linear regression model with the *gee* function from the R package "gee".[52]

3. Simulation study

3.1. Simulation study design

3.1.1. Data generation

We simulated a 2-group CRT with a time-to-event outcome expressed in days. We considered a total number of K clusters with exactly the same number of clusters in each group, $K_0 = K_1 = K/2$ and the variable cluster sizes m_k (k = 1, …, K) drawn from a negative binomial distribution with mean m and variance v.

We considered several scenarios depending on the degree of clustering and the magnitude of the intervention effect simulated under a proportional hazard or non-proportional hazard assumption. We generated clustered time-to-events with a Weibull gamma frailty regression model, with shape and scale parameters ρ and λ, respectively. Under the proportional hazards assumption, the hazard function for individual l from cluster k at time t was defined as $h_{kl}(t) = u_k \lambda \rho t^{\rho-1} \exp(Z_{kl}\beta_G)$, where β_G is the intervention effect and u_k the frailty term common to all the individuals of cluster k,

introducing a correlation between their survival times.[53] The frailty term for each cluster was drawn from a gamma distribution $\Gamma\left(\frac{1}{\theta},\frac{1}{\theta}\right)$, with a mean of one and variance θ. We quantified the degree of clustering by using the Kendall's tau τ, with $\tau = \frac{\theta}{\theta+2}$. Under the non-proportional hazards assumption, we assumed a delayed intervention effect with no intervention effect before time T_{delay} and a benefit after: $h_{kl}(t) = u_k\lambda\rho t^{\rho-1}\exp\left(Z_{kl}\beta_G\mathbb{1}\{t > T_{delay}\}\right)$. [54] With both proportional and non-proportional hazards assumptions, no additional covariate than intervention group was considered. We generated time-to-events by inverting the cumulative incidence function of the Weibull gamma frailty regression model. We simulated random censoring times independent of event times. We assumed that the censoring times of individuals from a same cluster were independent. More details about the simulation of the censoring times are available in Web Appendix C.1.

3.1.2. Scenarios

The simulation design was conducted to mimic real-world CRTs, choosing the parameters of our scenarios according to the distribution observed in a previous review of published CRTs.[4] We simulated CRTs with different numbers of clusters: $K = \{10, 20, 40, 50, 100\}$. Because GEE methods with a small number of clusters could lead to an inflation of the type I error rate for estimating the intervention effect,[39] we simulated scenarios with $K = \{10, 20, 40\}$ to assess the minimal number of clusters giving acceptable type I error rates. We considered a cluster size distribution of mean m = 80 and variance v=48² or m = 25 and variance v = 15², corresponding to a coefficient

of variation of the cluster size of 0.6 that seems realistic.[55] We did not consider the scenarios of 10, 20 and 40 clusters of size 25, which does not seem plausible.

We varied the degree of clustering by using the Kendall's tau $\tau \in \{0.001, 0.01, 0.05, 0.1, 0.2\}$.

The parameters (ρ, λ) of the Weibull distribution were set to (2, 0.000016), to obtain a survival rate in the control group of about 0.2 at the end of follow-up, fixed at 365 days (Web Figure C2). In the proportional hazards case, the HR (i.e., $\exp(\beta_G)$), was set to 0.5 and 0.8 to simulate a high and medium intervention effect, respectively. For non-proportional hazards assumption, we fixed the HR to 0.5 after the change point T_{delay}, set to 90 days. To assess the null hypothesis (i.e., absence of intervention effect), we set $\beta_G = 0$ corresponding to HR$= 1$.

We fixed the censoring rate ψ, reflecting the loss to follow-up, at 20% in all scenarios.

Finally, we explored 140 scenarios depending on number of clusters, size of cluster, degree of clustering, intervention effect and proportional hazards assumption or not (Table 1). We simulated 1000 datasets for each scenario.

3.1.3. Statistical analysis

We fixed the time horizon t^* at 365 days. When HR = 1, the true $\Delta RMST(t^*) = 0$. When HR $\neq 1$, the true $\Delta RMST(t^*)$ depends on both the Kendall's tau and the HR; for a fixed HR, it decreases with increasing Kendall's tau. The true $\Delta RMST(t^*)$ varied from 46.70 to 55.15 days for an HR of 0.5 and from 15.87 to 18.72 days for an HR of 0.8 under the proportional hazards assumption. We detail the procedure to compute the true

$\Delta RMST(t^*)$ up to 365 days in Web Appendix C.3 and provide the true value of $\Delta RMST(t^*)$ for each scenario in Web Table C3.

We set the significance level α to 5%. In each simulated dataset $d = \{1, \ldots, 1000\}$, we estimated the $\Delta RMST(t^*)$ between the intervention and control groups, the variance and the 95% CI at time horizon $t^* = 365$ days by using the five methods described above: KM_{indep}, PV_{indep}, KM_{clust}, PV_{ECM} and PV_{ICM}. In addition, we studied the permutation test for PV_{ECM} and PV_{ICM} when K = 10, 20, or 40. Because of exactly the same number of clusters in each group ($K_0 = K_1 = K/2$), the number of possible permutations of the intervention allocation is $C = \binom{K}{K/2}$. Therefore, for K = 10, C = 252 and all permutations were used. For K = 20 and 40 clusters, the number of permutations exceeded 100 000, so to obtain a reasonable computation time, we used 1000 randomly selected permutations from the C possible intervention allocations for the permutation test. First, the performance of methods in estimating the true $\Delta RMST(t^*)$ was assessed by the type I error rate corresponding to the proportion of 95% CIs for $\Delta RMST(t^*)$ that do not contain 0 when the true intervention effect equals 0 (i.e., when β_G and $\Delta RMST(t^*)$ equal 0). According to a binomial model with 1000 simulations and test size 5%, we considered a type I error rate of 3.6% to 6.4% acceptable. Second, for the methods with acceptable type I error rate, we studied the following performance criteria: relative bias, relative error of the estimated asymptotic standard error, and empirical coverage rate of the nominal 95% CIs. Performance criteria are detailed in Web Appendix B.4.

The code for simulation study is available on https://github.com/Le-Vilain-Abraham/Simulation-study-with-R-software.git.

3.2. Results

3.2.1. Convergence

Non-convergence occurred only for the pseudo-values regression with an exchangeable correlation matrix (PV_{ECM}), with a small degree of clustering and/or a limited number of clusters (Web Table D1). The percentage of convergence problems was low when $K \geq 40$, between 0.1% and 1.6%, but increased strongly with decreasing number of clusters, reaching up to 15.4% when $K = 10$ and $\tau = 0.001$. The performance measures of the PV_{ECM} method were computed on the basis of the simulation iterations which converged. The performance measures of the other methods were estimated with the 1000 simulated datasets.

In light of these results, we did not evaluate the PV_{ECM} method for the estimation of the permutation-based CI for $K < 50$. Indeed, this requires carrying out permutation tests of null hypothesis $H_0: \beta_1 = \beta^*$, for many different values of β^*, and therefore fitting multiple times model (9) of the Web Appendix B, that might not converge with an exchangeable working correlation matrix.

3.2.2. Type I error rate

Table 2 presents the type I error rate with the five methods and the permutation test for the pseudo-values regression for mean cluster size m = 80 and varying values of K and τ. Results for m = 25 (Web Table E1) are qualitatively similar.

As expected, methods that do not take into account the clustering (KM_{indep} and PV_{indep}) produced an uncontrolled type I error rate under all combinations of K and τ, except for the smallest τ value = 0.001. For the small degree of clustering τ = 0.01, the type I error

rate was about 20%. Of note, for those methods, the type I error rate increased with increasing τ and reached 70% when τ = 0.2.

In contrast, extended methods for clustered data (KM_{clust}, PV_{ECM} and PV_{ICM}) produced acceptable type I error rates if there was a sufficient number of clusters (≥50). There is no apparent pattern of the type I error rate according to Kendall's tau with these three methods. The type I error rate was slightly smaller with the PV_{ECM} method than KM_{clust} and PV_{ICM} methods. With $K < 50$, the type I error rate for the three methods KM_{clust}, PV_{ECM} and PV_{ICM} was largely above 6.4%.

We studied the permutation test to assess whether this method could be an acceptable alternative for CRTs with a limited number of clusters. The type I error rate for the permutation test for PV_{ECM} and PV_{ICM} are presented when K = 10, 20, or 40. For both methods, the permutation test produced an appropriate and similar type I error rate for all combinations of K and τ.

Thereafter, we excluded the methods that did not account for clustering, KM_{indep} and PV_{indep}, because they are not appropriate for CRTs.

3.2.3. Simulation results under proportional hazards assumption

The results for the performance criteria under the proportional hazards assumption for the three methods KM_{clust}, PV_{ECM} and PV_{ICM} are summarized in Table 3 for m=80. Results for m=25 (Web Table F1) are qualitatively similar.

All three methods lead to negligible bias for estimating their respective true $\Delta RMST(t^*)$. The relative bias, in absolute value, did not exceed 10% across all scenarios and was qualitatively similar for the three extended methods. The values of K, m, $\exp(\beta_G)$ and τ did not seem to affect the relative bias.

With $K \geq 50$, the relative error ranged from -6.1% to 2.1% for the methods accounting for clustering. The relative error was close to 0, regardless of the Kendall's tau. Therefore, with a sufficient number of clusters, the variance of the $\Delta RMST(t^*)$ was well estimated with the three methods. The relative error was slightly closer to 0 with the PV_{ECM} method than the other two methods, especially for high values of τ, which explains the slightly smaller type I error for the corresponding scenarios.

Overall, the coverage was close to 95% across all combinations, when $K \geq 50$. We found no obvious pattern when the simulation parameters varied.

With a limited number of clusters ($K < 50$), the relative error is negative, indicating that the three methods underestimate the variance. The negative bias in estimating the variance increased with decreasing total number of clusters. This negative bias explained the inflated type I error rates obtained in 3.2.2. as well as the coverage rate under the 95% nominal rate.

We also studied the coverage rate of the permutation-based CIs for PV_{ICM} when K = 10, 20, or 40 (Table 3). The coverage rate was close to 95% for all combinations of K, τ and $\exp(\beta_G)$, confirming that the permutation test is an appropriate alternative for CRTs with a limited number of clusters.

3.2.4. Simulation results under non-proportional hazards assumption

We provide the simulation results for the three methods accounting for clustering under non-proportional hazards assumption in Table 4 for m=80. Results for m=25 (Web Table F1) are qualitatively similar. Overall, the results are qualitatively similar as compared to those obtained under the proportional hazards assumption. The three methods have negligible bias in estimating the true $\Delta RMST(t^*)$ across all scenario

settings. When $K \geq 50$, the relative error ranged from -5.6% to 1.1% for the extended methods under all settings combinations. The coverage rates were close to 95% for the three clustered data methods across all values of K, m and τ. When K = 10, 20, or 40, the three methods had negative relative error and under-coverage.

The coverage rates for the permutation-based CIs for PV_{ICM} when $K < 50$ are presented in Table 4. The permutation test provided coverage rate close to 95% across all scenario settings.

4. Illustrative example

We illustrated the proposed methods using the education program for south Asians with asthma and their primary and secondary care clinicians (OEDIPUS) trial.[56] This two parallel-group CRT aimed at evaluating the effectiveness of an asthma-control education program dedicated to South Asian ethnic minority people in two London districts (Newham and Tower Hamlets). Clusters were general practices randomized to receive the intervention or usual care. In total, 84 general practices were randomized: 44 and 40 to the intervention and usual care groups, respectively. The study enrolled 375 patients: 183 and 192 in the intervention and usual care groups. Patients in both groups were comparable regarding the inclusion characteristics, such as sex, age and asthma severity (Table 2 in Griffiths et al.[56]).

In this study, two time-to-event outcomes were studied over the first year: 1) the time to the first unscheduled contact with an asthma exacerbation (one of the primary outcomes), and 2) the time to the first asthma review in primary care (one of the secondary outcomes). Because the outcomes of this trial were routinely collected, there was no loss to follow-up. At one year, the censoring rate was 30% for the primary

outcome and 36% for the secondary outcome. Kaplan-Meier curves for both outcomes are in Figure 1. The proportional hazards assumption, checked graphically by plotting the $\log(-\log(S(t)))$ curves, was met for both outcomes (Web Figure G1). The Kendall's tau was 0.02 and 0.12 for the primary and the secondary outcomes, respectively.

Using classic statistical analysis (marginal Cox model fitted with GEE), we found no significant effect of the asthma control education program on time to the first unscheduled contact with an asthma exacerbation (HR=1.09, 95% CI 0.85 to 1.41) but a significant intervention effect on time to the first asthma review in primary care (HR=2.73, 95% CI 1.54 to 2.78). The asthma-control education program may significantly contribute to reduce the time to the first asthma review in primary care.

We estimated the $\Delta RMST(t^*)$ for both outcomes by using the KM_{clust}, PV_{ECM} and PV_{ICM} methods. Although this is not necessary because the number of clusters is higher than 50, we estimated permutation-based CIs for comparison. The permutation-based CIs were calculated for both PV_{ICM} and PV_{ECM} methods, since the large number of clusters in the study prevents for non-convergence. The R code is available in Web Appendix G.2. We fixed t^* at 365 days, the maximum follow-up for the study because the data were collected over the first year. The results are in Table 5. We found similar conclusions as obtained with the HR but with complementary interpretations. For the three methods accounting for clustering, we estimated a 4-day non-significant difference between the intervention and control groups on time to the first unscheduled contact with an asthma exacerbation. In contrast, for the three methods, the asthma-control education program contributed to shortening the time to the first asthma review in primary care, by a significant decrease of nearly three months. Results obtained with

the three extended methods were very close. The permutation-based and the model-based CIs were similar and led to the same conclusions for both outcomes.

5. Discussion

The aim of this work was to propose and compare two easy-to-use approaches to estimate the $\Delta RMST(t^*)$ for CRTs with time-to-event outcomes. With a large simulation study, we demonstrated that i) all proposed methods accounting for clustering are unbiased, ii) they have good statistical performance in terms of type I error rate and variance estimation for CRTs with at least 50 clusters, and iii) the permutation test for pseudo-values regression allows for controlling for type I error rates when there are fewer than 50 clusters.

As expected, methods ignoring the clustered data structure produced an uncontrolled type I error rate, even with a small degree of clustering, and must be avoided. This has also been assessed in the presence of competing risks.[48] Not accounting for clustering in the statistical analysis could lead to incorrect conclusions, and appropriate methods must be applied. In contrast, the three proposed methods accounting for clustering, KM_{clust}, PV_{ECM} and PV_{ICM}, gave satisfactory estimates of the $\Delta RMST(t^*)$ and its variance, with controlled type I error rates if there was a sufficient number of clusters ($\geq$50). With a limited number of clusters, we observed inflated type I error rates. For the extended Kaplan-Meier-based method, the bootstrap percentile CIs tended to be too narrow for samples with a small number of units,[57] which explains the uncontrolled type I error rates. To our knowledge, no solution exists to control for type I error rate when using bootstrap CIs. For the pseudo-values regression methods, we found that the type I error can be controlled using a permutation test. An alternative approach could be the

use of bias-corrected sandwich variance estimators.[58-61] These correction methods were studied in CRTs for time-to-event outcomes and gave satisfactory results.[62-64] We will explore these bias-corrected estimators for pseudo-values regression in our future work.

Chen et al.[25] recently proposed an extended method for estimating $\Delta RMST(t^*)$ for clustered data based on a different approach directly modeling the RMST as a continuous function of the horizon time t^* and accounting for covariate-dependent censoring with inverse probability weights. Our method is suitable when there is a clear time horizon, but the Chen et al. approach is useful in an exploratory analysis to guide the selection of the most appropriate time point of interest t^*. Their method has the advantage of handling covariate-dependent censoring in CRTs. In our work, we assumed completely at random censoring, an assumption that may be violated in practice. It requires future research to assess the potential impact of a deviation from this assumption on the properties of the pseudo-values regression. Other censoring assumptions could be studied, such as covariate-dependent censoring, already investigated by Binder et al.[65] for independent data in the presence of competing risks, or clustered censoring times.

The three extended methods, KM_{clust}, PV_{ECM} and PV_{ICM} had similar good statistical performance to estimate $\Delta RMST(t^*)$ in a CRT with both proportional and non-proportional hazards assumptions and a reasonable number of clusters (≥ 50). In addition to the possibility of correcting type I error by using a permutation test when the number of clusters is small, another advantage of the pseudo-values regression method over the Kaplan-Meier-based methods is the possibility to adjust for other covariates than intervention group by including additional covariates in the generalized linear regression model described in equation (5). Adjustment on covariates is useful in CRTs

because the randomization of a smaller number of units does not always permit balance in cluster and individual level variables.[66,67] One other advantage of the extended pseudo-values regression methods is that they are already available in software such as R,[52] SAS[52] and Stata,[68] whereas the bootstrap variance estimates must be implemented. Consequently, we recommend using one of the pseudo values-based methods. When $K \geq 50$, the PV_{ECM} method gave slightly lower type I error rates than the other methods, especially with a high degree of clustering, so this method should be preferred if there is no convergence issue. As illustrated in the simulation study, for CRTs with less than 50 clusters, the PV_{ICM} method combined with a permutation test could be chosen to obtain both a p-value and a confidence interval.

One might be discouraged from estimating a confidence interval based on permutation test since it could be time-consuming. In our simulation study, the time to run a complete scenario depended on the number of clusters and varied from about two hours and twenty minutes, for 10 clusters, to four hours and 40 minutes, for 40 clusters. Therefore, on average, the computational time to estimate a permutation-based confidence interval for a single dataset was between 8.4 and 16.8 seconds. It is important to note that we benefited from a computing centre, that allowed us to parallelize the calculations with 30 cores. In the OEDIPUS trial, situation closer to practical use since the calculations were not parallelized, we estimated the confidence interval in a maximum time of 2 and a half minutes. As a reminder, the OEDIPUS trial randomized 84 clusters, so the computation time should be even shorter when the permutation-based confidence interval is used when $K < 50$. Therefore, the computation time seems reasonable to use permutation-based confidence interval in practice.

Our study has several limitations. First, we considered only the intervention group as a covariate in both the simulation of the data and the statistical analysis to allow for comparison of Kaplan-Meier and pseudo-values-based methods.[69] Additional research is needed to assess the performance of the pseudo-values regression with a larger number of covariates and clustered data. Second, we did not provide an analytical demonstration of the asymptotic properties of the pseudo-values regression in CRTs. Our work reported only empirical results on the performance of the proposed approaches based on the simulation study. Therefore, our results can only be reliably applied to configurations similar to those we specifically evaluated and should not be extrapolated to more general situations. Although the proposed sandwich variance estimation performed well in the simulation study for a sufficient number of clusters, it does not take into account the variability induced by the estimation of the pseudo-values with jackknife. For independent data, Jacobsen and Martinussen[70] and Overgaard et al.[71] showed that the sandwich variance estimate for pseudo-values regression is in general not consistent because of an omitted term. We could expect the same result for clustered data. It involves a complex theory, but one could possibly derive an appropriate variance estimator for pseudo-values regression in CRTs using von Mises expansion as in Jacobsen and Martinussen,[70] Overgaard et al.[71] and Zeng et al.[72] Third, the use of $\Delta RMST(t^*)$ as the main intervention effect estimate in a CRT will require a sample-size formula based on $\Delta RMST(t^*)$. Because to our knowledge, no sample-size formula based on $\Delta RMST(t^*)$ has been developed in the context of CRTs, future work should consider the development of such a formula.

The $\Delta RMST(t^*)$ has several advantages over the classical HR. The difference in RMST is easily interpretable as the expected survival duration gained due to the intervention

for patients followed up to t^*. The unit of measure is temporal and the magnitude of the intervention effect is quantified in a more understandable way than the HR. In addition, the difference in RMST does not rely on the proportional hazards assumption, and our simulation study confirmed the good performance of the $\Delta RMST(t^*)$ under the proportional hazards assumption as well as the non–proportional. The only complexity when using $\Delta RMST(t^*)$ is the choice of an appropriate time horizon of interest. The time horizon t^* must be a priori specified at the trial design stage and should be a clinically meaningful time based on the study objective.[13,29] If this complex choice of the time horizon of interest could lead to use a HR to summarize the intervention effect, one must be aware that the HR also relies on a time window. Not reporting the HR with a time window is similar to considering that the HR is constant forever.[73] Then the HR and the $\Delta RMST(t^*)$ both require the choice of a time horizon, and the choice of t^* should not be a criterion to select the HR instead of the $\Delta RMST(t^*)$.

In conclusion, the $\Delta RMST(t^*)$ is an interesting alternative measure of the intervention effect for time-to-event outcomes. We demonstrated that it can be accurately and simply estimated in CRTs by using one of our proposed methods under proportional and non-proportional hazards assumptions and a sufficient number of clusters ($\geq$50) and that we can perform a permutation test in case of a small number of clusters (<50).

Acknowledgments

The authors thank Bruno Giraudeau, Monica Taljaard and Jennifer Thompson for helpful discussions. The authors thank the Centre for Primary Care and Public Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, United Kingdom for sharing the data of the OEDIPUS trial. The authors also

thank Laura Smales for English editing. The authors thank the editor, and two anonymous reviewers for their valuable comments that improved the paper. The authors benefitted from the use of the cluster at the Centre de Calcul Scientifique en région Centre-Val de Loire.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the French National Research Agency [ANR-19-CE36-0002 –QUARTET].

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Supplemental Material

Web Appendix for this article is available online.

References

1. Donner A and Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.

2. Eldridge S and Kerry SM. *A practical guide to cluster randomised trials in health services research*. Chichester, West Sussex: John Wiley & Son, 2011.

3. Murray DM, Varnell SP and Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health* 2004; 94: 423-432.

4. 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: 125-37.

5. Cox DR. Regressions models and life-tables (with discussion). *J R Stat Soc Series B Stat Methodol* 1972; 34: 187-220.

6. Duchateau L and Janssen P. *The frailty model*. New York: Springer Verlag, 2008.

7. 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: 1065-73.

8. Therneau TM and Grambsch PM. *Modeling Survival Data*. Berlin: Springer, 2000.

9. Kalbfleich JD and Prentice RL. *The Statistical Analysis of Failure Time Data.* New York: Wiley, 1980.

10. Lawless JF. Statistical Models and Methods for Lifetime Data. New York: Wiley, 1982.

11. Messori A, Bartoli L and Trippoli S. The restricted mean survival time as a replacement for the hazard ratio and the number needed to treat in long-term studies. *ESC Heart Fail* 2021; 8: 2345-2348.

12. Trinquart L, Jacot J, Conner SC, et al. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. *J Clin Oncol* 2016; 34: 1813-9.

13. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the

between-group difference in survival analysis. *J Clin Oncol* 2014; 32: 2380-5.

14. Uno H, Wittes J, Fu H, et al. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. *Ann Intern Med* 2015; 163: 127-34.

15. Royston P and Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med* 2011; 30:2409-21.

16. Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010; 2:13-5.

17. Stensrud MJ, Aalen JM, Aalen OO, et al. Limitations of hazard ratios in clinical trials. *Eur* Heart J. 2019; 40: 1378-1383.

18. Irwin J. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg (Lond)* 1949; 47:188.

19. Chen PY and Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* 2001; 57: 1030-8.

20. Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; 13: 152.

21. Kaplan EL and Meier P. Nonparametric Estimation From Incomplete Observations. *J Am Stat Assoc* 1958; 53: 457-481.

22. Andersen PK, Hansen MG and Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Anal* 2004; 10: 335-50.

23. Andersen PK and Perme MP. Pseudo-observations in survival analysis. *Stat*

Methods Med Res 2010; 19: 71-99.

24. Ambrogi F, Iacobelli S and Andersen PK. Analyzing differences between restricted mean survival time curves using pseudo-values. *BMC Med Res Methodol*; 22:71.

25. Tian L, Zhao L and Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 2014; 15: 222-33.

26. Wang X and Schaubel DE. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime Data Anal* 2018; 24: 176-199.

27. Zhong Y and Schaubel DE. Restricted mean survival time as a function of restriction time. *Biometrics* 2022; 78: 192-201.

28. Chen X, Harhay MO and Li F. Clustered restricted mean survival time regression. *Biom J* 2022.

29. Hasegawa T, Misawa S, Nakagawa S, et al. Restricted mean survival time as a summary measure of time-to-event outcome. *Pharm Stat* 2020; 19: 436-53.

30. Ying Z and Wei LJ. The Kaplan-Meier estimate for dependent failure time observations. *J Multivar Anal* 1994; 50: 17-29.

31. Efron B and Tibshirani RJ. *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC, 1994.

32. Ren S, Lai H, Tong W, et al. Nonparametric bootstrapping for hierarchical data. *J Appl Stat* 2010; 37: 1487-1498.

33. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.

34. Zeger SL and Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42: 124-130.

35. Sharples K and Breslow N. Regression analysis of correlated binary data: some small samples results for the estimating equation approach. *J Stat Comput Simul* 1992; 42: 1- 20.

36. Lumley T. Generalized Estimating Equations for Ordinal Data: A Note on Working Correlation Structures. *Biometrics* 1996; 52: 354-361.

37. Ahn KW and Mendolia F. Pseudo-value approach for comparing survival medians for dependent data. *Stat Med* 2014; 33: 1531-1538.

38. Logan BR, Zhang M-J and Klein JP. Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics* 2011; 67: 1-7.

39. Leyrat C, Morgan KE, Leurent B, et al. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol* 2018; 47: 321-331.

40. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004; 1: 80-90.

41. Ivers NM, 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.

42. Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 2016; 17: 438.

43. 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.

44. Gail MH, Mark SD, Carroll RJ, et al. On design considerations and randomization-based inference for community intervention trials. *Stat Med* 1996; 15: 1069-92.

45. Li F, Lokhnygina Y, Murray DM, Heagerty PJ, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Stat Med* 2016; 35: 1565-79.

46. 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.

47. Cai J and Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. *Stat Med* 2000; 19: 2963–2973.

48. Li F, Lu W, Wang Y, et al. A comparison of analytical strategies for cluster randomized trials with survival outcomes in the presence of competing risks. *Stat Methods Med Res* 2022; 31: 1224-1241.

49. Rabideau DJ and Wang R. Randomization-based confidence intervals for cluster randomized trials. *Biostatistics* 2021; 22: 913-927.

50. Garthwaite, PH. Confidence Intervals from Randomization Tests. *Biometrics* 1996; 52: 1387–1393

51. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: 2010.

52. Klein JP, Gerster M, Andersen PK, et al. SAS and R functions to compute pseudo-values for censored data regression. *Comput Methods Programs Biomed* 2008; 89: 289-300.

53. Crowther MJ and Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013; 32: 4118-34.

54. Jachno K, Heritier S and Wolfe R. Impact of a non-constant baseline hazard on

detection of time-dependent treatment effects: a simulation study. *BMC Med Res Methodol* 2021; 21: 177.

55. Eldridge SM, Ashby D and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006; 35: 1292-300.

56. Griffiths C, Bremner S, Islam K, et al. Effect of an Education Programme for South Asians with Asthma and Their Clinicians: A Cluster Randomised Controlled Trial (OEDIPUS). *PLoS One* 2016; 11: e0158783.

57. Hesterberg TC. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *Am Stat* 2015; 69: 371-386.

58. Mancl LA and DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; 57: 126-34.

59. Kauermann G and Carroll RJ. (2001). A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc 2001; 96: 1387–1396.

60. Fay MP and Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001;57: 1198–1206.

61. Morel J, Bokossa M, and Neerchal N. Small sample correction for the variance of GEE estimators. *Biom J* 2003; 45: 395–409.

62. Blaha O, Esserman D, and Li F. Design and analysis of cluster randomized trials with time-to-event outcomes under the additive hazards mixed model. *Stat Med* 2022; 41: 4860-4885.

63. Wang X, Turner EL, and Li F. Improving sandwich variance estimation for marginal Cox analysis of cluster randomized trials. *Biom J* 2022: e2200113.

64. Chen X and Li F. Finite-sample adjustments in variance estimators for clustered competing risks regression. *Stat Med* 2022; 41: 2645-2664.

65. Binder N, Gerds TA, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Anal* 2014; 20: 303-15.

66. Wright N, Ivers N, Eldridge S, et al. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015; 68: 603-609.

67. Dron L, Taljaard M, Cheung YB, et al. The role and challenges of cluster randomised trials for global health. *Lancet Glob Health* 202; 9: e701-e710.

68. Parner ET and Andersen PK. Regression analysis of censored data using pseudo-observations. *Stata Journal* 2010; **10**: 408-422.

69. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; 38: 2074-2102.

70. Jacobsen M and Torben M. A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-Observations. *Scandinavian Journal of Statistics* 2016; 43:845–862.

71. Overgaard M, Parner E and Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. The Annals of Statistics 2017; 45: 1988-2015.

72. Zheng S, Li F, Hu L et al. Propensity Score Weighting Analysis of Survival Outcomes Using Pseudo-observations. Statistica Sinica 2021.

73. Uno H and Tian L. Is the Log-Rank and Hazard Ratio Test/Estimation the Best Approach for Primary Analysis for All Trials? *J Clin Oncol* 2020; 38: 2000-2001.

List of Tables and Figures

Figure 1: Kaplan-Meier curves for the outcomes of OEDIPUS trial: A) probability of no first unscheduled contact with asthma exacerbation at time t; B) probability of no asthma review in primary care at time t

Note: The median time to first unscheduled contact with an asthma exacerbation in the intervention and the control groups is 171 and 189 days, respectively. The median time to first asthma review in primary care in the intervention and the control groups is 71 and 324 days, respectively.

Table 1: Simulation settings for the number of clusters, mean cluster size, Kendall's tau, true hazard ratio (HR) and proportional hazards assumption or not.

Total number of clusters K	Mean cluster size m	Kendall's tau τ	True HR $\exp(\beta_G)$
Null hypothesis			
10, 20, 40	80	0.001, 0.01, 0.05, 0.1, 0.2	1
50, 100	25, 80	0.001, 0.01, 0.05, 0.1, 0.2	1
Proportional hazards			
10, 20, 40	80	0.001, 0.01, 0.05, 0.1, 0.2	0.5, 0.8
50, 100	25, 80	0.001, 0.01, 0.05, 0.1, 0.2	0.5, 0.8
Non-proportional hazards			
10, 20, 40	80	0.001, 0.01, 0.05, 0.1, 0.2	0.5 (after 90 days)
50, 100	25, 80	0.001, 0.01, 0.05, 0.1, 0.2	0.5 (after 90 days)

Table 2: Type I error rate (%) for different combinations of number of clusters (K = 10, 20, 40, 50, 100) and Kendall's tau (τ = 0.001, 0.01, 0.05, 0.1, 0.2) with mean cluster size m = 80. Type I error between 3.6% and 6.4%, indicated in bold, was considered acceptable based on a binomial model with 1000 simulated datasets.

Number of clusters K	Kendall's τ	Methods for independent data		Extended methods for clustered data			Extended methods and permutation test	
		KM_{indep}	PV_{indep}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ECM}	PV_{ICM}
10	0.001	**6.4**	6.5	12.5	18.0	16.0	**4.9**	**5.1**
	0.01	20.9	20.9	12.2	13.0	13.5	**5.6**	**4.5**
	0.05	48.3	48.3	12.7	13.2	16.6	**4.5**	**4.2**
	0.1	58.9	58.9	12.4	11.0	15.6	**4.7**	**4.8**
	0.2	68.6	68.6	13.4	13.3	16.7	**4.6**	**4.6**
20	0.001	**5.9**	**5.9**	7.2	9.6	8.5	**4.6**	**5.3**
	0.01	19.9	19.9	8.7	9.8	9.6	**5.2**	**4.9**
	0.05	48.3	48.3	7.6	7.4	9.1	**4.0**	**4.0**
	0.1	60.2	60.2	9.5	9.4	10.9	**5.3**	**5.8**
	0.2	72.3	72.3	7.2	6.6	9.5	**3.9**	**4.1**
40	0.001	6.8	6.8	6.9	7.6	7.1	**5.6**	**5.7**
	0.01	20.3	20.3	6.8	6.5	6.9	**4.9**	**5.1**
	0.05	47.7	47.7	7.0	7.2	7.6	**5.3**	**4.8**
	0.1	60.7	60.7	7.0	**6.2**	7.8	**5.2**	**5.3**
	0.2	70.1	70.1	6.5	**5.5**	7.1	**4.1**	**4.9**
50	0.001	**6.2**	**6.2**	6.7	6.9	6.6		
	0.01	18.3	18.3	**6.2**	**5.4**	**6.4**		
	0.05	46.3	46.3	7.1	**6.4**	7.4		
	0.1	59.9	59.9	**6.1**	**5.9**	6.6		
	0.2	71.7	71.7	**5.9**	**5.8**	**6.3**		
100	0.001	7.4	7.4	**6.2**	**6.2**	**6.1**		
	0.01	17.8	17.8	**4.9**	**4.9**	**5.3**		
	0.05	46.4	46.4	**5.6**	**5.1**	**6.0**		
	0.1	60.3	60.3	**5.7**	**5.4**	**6.0**		
	0.2	70.6	70.5	6.7	**6.1**	7.0		

Table 3: Relative bias (%), relative error (%) and coverage rate (%) under proportional hazards assumptions for different combinations of number of clusters (K = 10, 20, 40, 50, 100), Kendall's tau (τ = 0.001, 0.01, 0.05, 0.1, 0.2) and true HR (HR = 0.8, 0.5) with mean cluster size m = 80.

True HR	K	τ	Relative bias			Relative error			Coverage rate			
			KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ICM} and Permutation
0.8	10	0.001	0.05	-0.59	0.05	-7.5	-14.1	-14.2	87.9	84.0	85.7	95.6
		0.01	-0.63	-0.81	-0.66	-14.1	-18.2	-18.8	87.2	85.2	84.2	95.3
		0.05	7.24	7.07	7.17	-17.2	-14.4	-20.5	86.3	85.3	83.2	95.1
		0.1	-6.00	-6.39	-6.04	-13.6	-8.6	-17.1	88.1	88.4	84.7	95.4
		0.2	-2.22	-3.25	-2.42	-18.6	-14.2	-21.4	87.2	88.0	82.2	96.1
	20	0.001	0.51	0.52	0.50	-6.9	-10.3	-9.2	90.5	89.1	90.0	94.8
		0.01	-2.48	-2.45	-2.49	-3.6	-4.7	-5.1	92.4	91.7	91.8	95.7
		0.05	1.92	1.68	1.89	-4.7	-3.4	-5.5	92.5	91.7	91.0	95.4
		0.1	2.23	2.01	2.15	-5.8	-7.1	-6.5	91.6	91.4	90.1	95.2
		0.2	1.10	1.96	0.94	-9.1	-7.4	-9.7	90.8	92.2	89.7	95.2
	40	0.001	1.58	1.45	1.56	-1.1	-3.9	-2.0	95.1	94.1	94.8	96.4
		0.01	-0.64	-0.64	-0.66	-8.4	-8.0	-8.8	92.1	92.0	91.7	93.8
		0.05	0.39	-0.12	0.34	-8.1	-5.1	-8.0	92.1	93.2	91.1	94.8
		0.1	2.80	4.79	2.78	-6.8	-5.7	-6.6	91.8	92.8	90.9	94.2
		0.2	0.62	-0.51	0.55	-5.5	-4.4	-5.4	93.2	93.4	92.4	95.5
	50	0.001	0.44	0.44	0.42	-2.7	-4.8	-3.4	93.3	92.2	92.6	
		0.01	1.62	1.53	1.61	-1.1	-1.2	-1.3	94.4	93.8	94.0	
		0.05	-2.82	-3.30	-2.83	1.9	-0.9	2.1	94.0	93.0	93.7	
		0.1	0.19	0.97	0.15	-3.4	-0.5	-3.2	94.6	95.1	93.8	
		0.2	-6.26	-7.24	-6.32	-4.8	-4.2	-4.6	92.9	93.6	92.1	
	100	0.001	0.98	0.96	0.96	0.3	-0.6	0.1	94.7	94.6	94.5	
		0.01	1.12	0.89	1.11	-3.3	-3.4	-3.3	94.1	93.9	94.2	
		0.05	0.46	-0.13	0.44	-6.1	-5.2	-5.9	93.6	93.7	93.3	

		0.1	0.11	-0.48	0.08	-1.6	0.7	-1.5	94.1	94.7	93.9	
		0.2	-5.77	-5.97	-5.80	-4.2	-2.8	-4.1	94.1	94.4	93.5	
0.5	10	0.001	-2.12	-2.34	-2.25	-17.3	-13.5	-20.8	85.9	87.0	81.5	94.3
		0.01	0.08	-0.35	-0.08	-11.2	-18.4	-17.1	86.5	82.5	85.6	95.7
		0.05	-0.49	-0.59	-0.61	-13.1	-15.3	-17.8	86.6	84.2	82.7	96.1
		0.1	-2.37	-1.51	-2.49	-14.6	-12.3	-17.9	87.5	87.9	84.6	95.1
		0.2	-0.23	-0.79	-0.27	-12.7	-12.4	-16.7	87.2	85.3	84.1	96.4
	20	0.001	-0.68	-0.71	-0.82	-3.9	-7.5	-6.1	92.4	90.5	91.8	94.5
		0.01	-1.05	-1.03	-1.14	-8.7	-10.5	-10.1	90.4	89.4	90.2	94.1
		0.05	-0.51	-0.38	-0.63	-10.2	-6.2	-11.4	91.7	91.8	90.2	95.8
		0.1	0.70	0.79	0.51	-9.5	-5.9	-10.4	91.1	92.1	89.5	95.1
		0.2	0.33	0.83	0.23	-6.7	-5.3	-7.9	91.8	91.3	89.7	94.8
	40	0.001	-0.79	-0.84	-0.92	-2.7	-4.3	-3.7	92.6	91.4	92.2	94.2
		0.01	-1.11	-1.22	-1.24	-6.1	-5.1	-6.6	92.9	92.1	92.0	94.4
		0.05	-0.89	-1.26	-1.01	-3.7	-2.1	-4.1	92.5	93.6	92.4	95.1
		0.1	-1.90	-1.67	-2.02	-4.0	-3.0	-4.2	92.9	94.2	92.6	95.2
		0.2	-4.15	-3.25	-4.26	-7.9	-6.6	-8.2	91.2	92.0	90.2	93.6
	50	0.001	-0.61	-0.76	-0.73	0.9	-1.1	0.4	93.5	93.3	93.7	
		0.01	-0.32	-0.43	-0.44	-1.2	-1.4	-1.5	94.6	94.4	94.6	
		0.05	-0.83	-1.21	-0.94	-2.8	-2.8	-3.1	93.8	94.1	92.9	
		0.1	-2.58	-2.13	-2.68	-4.1	-5.4	-4.4	92.9	91.9	92.6	
		0.2	-1.77	-1.33	-1.88	-5.7	-4.5	-6.0	93.2	92.7	92.3	
	100	0.001	-0.69	-0.79	-0.81	-0.8	-1.1	-0.7	94.5	94.7	94.7	
		0.01	-0.44	-0.59	-0.57	-0.6	-1.4	-0.8	94.6	94.3	94.4	
		0.05	-1.03	-0.97	-1.15	0.8	0.5	0.7	95.0	95.8	94.7	
		0.1	-1.67	-1.45	-1.78	-4.8	-5.2	-5.0	93.5	93.2	93.2	
		0.2	-1.93	-1.83	-2.03	-5.0	-4.8	-5.2	93.5	92.8	93.2	

Table 4: Relative bias (%), relative error (%) and coverage rate (%) under non-proportional hazards assumptions for different combinations of number of clusters (K = 10, 20, 40, 50, 100). Kendall's tau (τ = 0.001, 0.01, 0.05, 0.1, 0.2) with true HR = 0.5 after 90 days and mean cluster size m = 80.

K	τ	Relative bias			Relative error			Coverage rate			
		KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ICM} and Permutation
10	0.001	2.33	2.29	2.26	-4.3	-13.5	-10.7	88.2	85.4	87.1	96.0
	0.01	1.78	1.67	1.75	-10.0	-12.8	-14.8	87.6	86.0	84.9	95.5
	0.05	1.11	1.62	0.99	-7.7	-6.1	-11.4	89.1	88.4	86.9	96.7
	0.1	-2.51	-1.17	-2.58	-14.3	-10.3	-17.9	87.8	88.0	84.6	94.8
	0.2	1.19	1.67	0.95	-12.4	-9.2	-15.5	88.6	88.5	84.9	95.9
20	0.001	2.49	2.54	2.37	-7.2	-11.5	-9.2	90.9	88.3	89.5	94.8
	0.01	2.70	2.73	2.61	-7.5	-7.7	-8.8	91.0	90.9	90.3	94.7
	0.05	2.90	2.68	2.83	-10.3	-7.9	-11.2	90.5	90.5	89.1	94.4
	0.1	1.83	1.88	1.69	-9.0	-8.0	-9.8	90.9	91.6	89.3	94.9
	0.2	6.47	6.33	6.31	-10.4	-7.2	-11.2	89.7	89.7	87.7	93.4
40	0.001	2.62	2.68	2.52	0.2	-1.8	-0.6	92.8	91.9	92.5	94.8
	0.01	2.47	2.49	2.36	-3.7	-3.1	-3.9	92.7	93.7	92.6	95.2
	0.05	1.96	1.90	1.87	-4.5	-3.7	-4.4	92.9	93.8	92.5	95.2
	0.1	2.59	1.67	2.47	-1.7	2.3	-1.6	93.1	94.0	92.2	95.7
	0.2	2.79	1.71	2.68	-3.7	-1.1	-3.7	93.2	92.9	92.3	95.2
50	0.001	2.80	2.75	2.71	-3.3	-4.2	-3.7	92.5	91.9	92.6	
	0.01	2.78	2.63	2.67	-5.1	-5.2	-5.2	91.8	92.2	92.3	
	0.05	3.17	2.49	3.08	-0.3	0.5	-0.2	94.6	94.2	94.3	
	0.1	0.20	0.92	0.10	-4.5	-0.8	-4.4	93.5	93.7	93.5	
	0.2	-0.80	-0.61	-0.90	-6.0	-1.1	-6.0	94.1	95.0	93.4	
100	0.001	2.52	2.45	2.43	-2.1	-2.5	-2.2	91.7	91.9	92.2	
	0.01	2.60	2.41	2.50	0.5	-0.2	0.7	94.4	93.7	94.2	
	0.05	1.96	2.22	1.88	-3.0	-0.1	-2.8	94.0	94.4	94.1	
	0.1	2.28	1.62	2.20	-0.6	0.9	-0.4	94.9	94.8	94.6	
	0.2	2.54	2.44	2.47	-3.7	-4.3	-3.7	93.6	93.8	93.4	

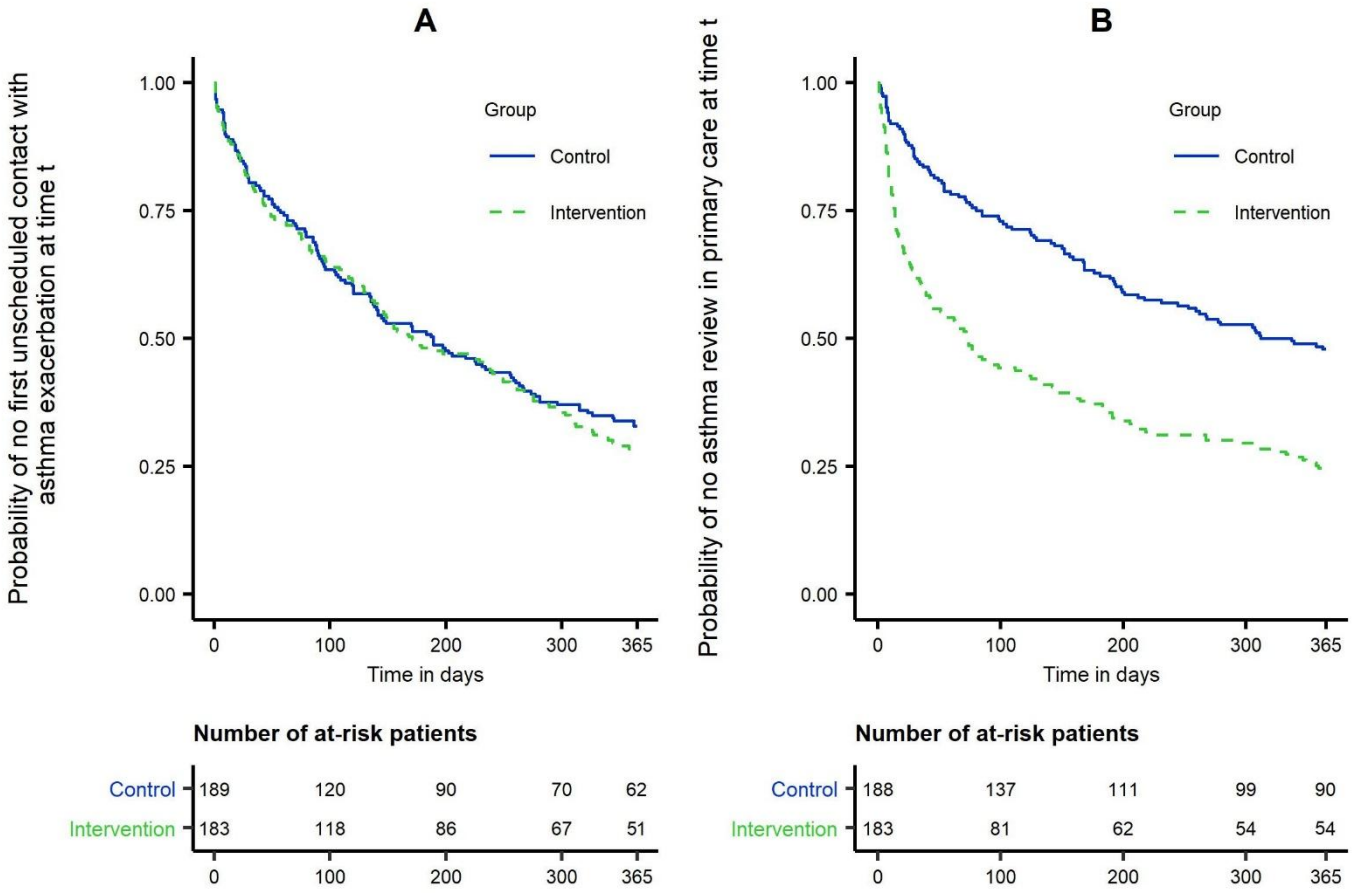
Table 5: Estimations of the $\Delta RMST(t^*)$ at 365 days and the 95% confidence interval (95% CI) for the two outcomes of the OEDIPUS trial: 1) time to first unscheduled contact with an asthma exacerbation (n = 372) and 2) time to first asthma review in primary care (n = 371).

Method	$\Delta RMST(365)$	95% CI	
		Model-based	Permutation-based
Time to first unscheduled contact with an asthma exacerbation			
KM_{clust}	-4.056	[-36.181;27.254]	-
PV_{ECM}	-4.718	[-36.128;26.691]	[-38.009;27.613]*
PV_{ICM}	-4.056	[-35.088;26.977]	[-35.907;27.788]*
Time to first asthma review in primary care			
KM_{clust}	-89.524	[-122.063;-53.317]	-
PV_{ECM}	-84.104	[-118.567;-49.642]	[-118.040;-47.220]**
PV_{ICM}	-89.524	[-123.717;-55.331]	[-126.179;-53.566]**

**The computation times for the PV_{ECM} and PV_{ICM} are 2 minutes and 30 secondes and 1 minutes and 48 secondes respectively.*

*** The computation times for the PV_{ECM} and PV_{ICM} are 2 minutes and 12 secondes and 1 minute and 48 secondes, respectivel*

Figure 1: Kaplan-Meier curves for the outcomes of OEDIPUS trial: A) probability of no first unscheduled contact with asthma exacerbation at time t; B) probability of no asthma review in primary care at time t



Note: The median time to first unscheduled contact with an asthma exacerbation in the intervention and the control groups is 171 and 189 days, respectively. The median time to first asthma review in primary care in the intervention and the control groups is 71 and 324 days, respectively.