UNIVERSITÉ DE TOURS

ÉCOLE DOCTORALE : Santé, Sciences Biologiques et Chimie du Vivant

UMR INSERM 1246 - SPHERE

THÈSE présentée par :

Floriane LE VILAIN--ABRAHAM

soutenue le : **18 octobre 2023**

pour obtenir le grade de : **Docteur de l'université de Tours**
Discipline/ Spécialité : Sciences de la Vie et de la Santé / Biostatistiques

Analyse d'un essai randomisé en clusters avec un critère de jugement de type survie

THÈSE dirigée par :
Mme CAILLE Agnès Professeure des universités – Praticienne hospitalière, Université de Tours

THÈSE co-dirigée par :
Mme DESMÉE Solène Maîtresse de conférences des universités, Université de Tours

RAPPORTEURS :
M. LATOUCHE Aurélien Professeur des universités, Conservatoire National des Arts et Métiers
Mme RONDEAU Virginie Directrice de recherche, INSERM U1219 Bordeaux Population Health

PRÉSIDENTE DU JURY :
Mme BIERME Hermine Professeure des universités, Université de Tours

---

JURY :
Mme BIERME Hermine Professeure des universités, Université de Tours
Mme CAILLE Agnès Professeure des universités – Praticienne hospitalière, Université de Tours
Mme DESMÉE Solène Maîtresse de conférences des universités, Université de Tours
M. LAMBERT Jérôme Professeur des universités – Praticien hospitalier, Université Paris Cité
M. LATOUCHE Aurélien Professeur des universités, Conservatoire national des arts et métiers
Mme RONDEAU Virginie Directrice de recherche, INSERM U1219 Bordeaux Population Health

Remerciements

Je tiens à remercier et à exprimer toute ma reconnaissance à Agnès Caille. Je te remercie de m'avoir permis de travailler sur le projet QUARTET au cours de cette thèse et de pouvoir continuer à le faire après la fin de celle-ci. Merci pour le temps que tu m'as consacré malgré tes nombreuses responsabilités. Merci pour ta confiance, ta bienveillance, ton soutien, tes relectures minutieuses, le partage de tes connaissances et de ton expérience qui m'ont donné un cadre de travail idéal.

Je tiens également à remercier et à exprimer toute ma reconnaissance à Solène Desmée. J'ai eu énormément de chance de pouvoir travailler avec toi. Merci pour ta disponibilité, ta gentillesse, tes précieux conseils, ta rigueur et également ton soutien, qui m'ont permis d'avoir d'aussi bonnes conditions de travail pour réaliser cette thèse. Tes connaissances et ton accompagnement au cours de ces trois années m'ont été plus que bénéfiques.

Je tiens à remercier et à exprimer toute ma gratitude envers les membres du jury d'avoir accepté de prendre le temps d'évaluer ce travail. Je remercie le Professeur Aurélien Latouche et Madame Virginie Rondeau de m'avoir fait l'honneur d'accepter de rapporter cette thèse. Je remercie également les Professeurs Hermine Biermé et Jérôme Lambert d'avoir accepté de juger cette thèse et de participer à mon jury de thèse.

Je tiens à remercier Elsa Tavernier. Merci pour le partage de tes connaissances très précieuses en R et en survie (entre autres). Merci pour ta disponibilité constante, ta bonne humeur, tes encouragements, tes questionnements et le temps passé à m'aider à résoudre toutes sortes de problèmes.

Je tiens à remercier Étienne Dantan. Merci d'avoir pris le temps de prendre part et de t'être investi dans les différents travaux de cette thèse. Merci pour tes conseils pertinents et tes connaissances qui ont été d'une très grande utilité et qui ont, de manière certaine, amélioré les travaux de cette thèse.

I also want to thank Jennifer Thompson for her help, useful advice and valuable expertise on cluster randomized trials with a small number of clusters.

Je tiens à remercier Bruno Giraudeau. Merci de m'avoir encadré dans le cadre de mon stage en 2019, durant lequel j'ai pu découvrir le monde des essais randomisés en clusters. J'ai énormément appris au cours de ces quelques mois. Merci, car ce stage m'a donné le goût pour la recherche et l'envie de faire cette thèse.

Je tiens à remercier toute l'équipe de SPHERE et en particulier les personnes avec qui j'ai partagé le bureau des doctorants à Tours : merci pour les échanges, les coups de mains et le soutien. Merci particulièrement à Jules pour ton aide et nos nombreuses discussions, la plus part du temps concernant les essais randomisés en clusters.

Je tiens à remercier tous les membres de l'équipe du CIC du CHRU de Tours que j'ai pu croiser au cours des dernières années pour leur accueil chaleureux, leur bienveillance, leur aide, leur écoute, leur soutient moral, leur patience et pour les très nombreux repas, pauses improvisées et douceurs partagées.

Enfin, je tiens à remercier tous mes proches pour leur patience, leur confiance, leur présence et leur soutien sans faille.

Ce travail a été financé par l'Agence Nationale de la Recherche dans le cadre du projet QUARTET (QUantification of clustering and Analysis of cluster Randomised Trials with time-to-EvenT outcomes) du programme Jeunes Chercheuses et Jeunes Chercheurs.

[ANR-19-CE36-0002-QUARTET]

Valorisation scientifique

Publications issues de la thèse

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Restricted mean survival time to estimate an intervention effect in a cluster randomized trial. *Statistical Methods in Medical Research*. 2023.

Le Vilain-Abraham F, Desmée S, Thompson JA, Tavernier A, Dantan E, Caille A. Restricted mean survival time in cluster randomized trials with a small number of clusters: Improving variance estimation for pseudo-values regression. [En préparation]

Autres publications en lien avec la thèse

Caille A, Taljaard M, **Le Vilain-Abraham F**, Le Moigne A, Copas AJ, Tubach F, Dechartres A. Recruitment and implementation challenges were common in stepped-wedge cluster randomized trials: Results from a methodological review. *Journal of Clinical Epidemiology*. 2022.

Marce D, **Le Vilain-Abraham F**, Bridou M, Quereux G, Dupuy A, Lesimple T, Le Corre Y, Wierzbicka-Hainaut E, Legoupil D, Celerier P, Maillard H, Machet L, Caille A. Impact of Distribution of a Tip Sheet to Increase Early Detection and Prevention Behavior Among First-degree Relatives of Melanoma Patients: a Cluster Randomized Trial. *Cancers*. 2022.

Communications orales

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Analysis of time-to-event outcomes using restricted mean survival time in cluster randomised trials. *7^{th} cluster meeting on Current Developments in Cluster Randomised Trials and Stepped Wedge Designs*, online conference, 15-17 novembre 2021.

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Différence de temps de survie moyen restreint pour évaluer un effet traitement dans un essai randomisé en clusters. *$16^{\text{ème}}$ conférence francophone d'épidémiologie clinique (EPICLIN) et $29^{\text{èmes}}$ journées des statisticiens des centres de lutte contre le cancer (CLCC)*, Paris, France, 18-20 mai 2022.

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Estimating the intervention effect using restricted mean survival time in a cluster randomized trial. *43^{rd} Conference of the International Society for Clinical Biostatistics (ISCB)*, Newcastle, Angleterre, 21-25 août 2022.

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Improving variance estimation for pseudo-values regression for restricted mean survival time in small sample cluster randomized trials. *$17^{\text{ème}}$ conférence francophone d'épidémiologie clinique (EPICLIN) et $30^{\text{èmes}}$ journées des statisticiens des centres de lutte contre le cancer (CLCC)*, Paris, France, 10-12 mai 2023.

Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A. Improving variance estimation for pseudo-values regression for restricted mean survival time in small sample cluster randomized trials. *44^{th} Conference of the International Society for Clinical Biostatistics (ISCB)*, Milan, Italie, 27-31 août 2023.

Communication affichée

Le Vilain-Abraham F, Tavernier E, Desmée S, Caille A. Analysis of time-to-event outcomes using restricted mean survival time in cluster randomised trials. *$5^{\text{ème}}$ Journée de la recherche biomédicale Angers-Tours*, Angers, France, 10 décembre 2021.

Récompense scientifique

Lauréate du prix "Jeune chercheur" de la meilleure communication orale. *$16^{\text{ème}}$ conférence francophone d'épidémiologie clinique (EPICLIN) et $29^{\text{èmes}}$ journées des statisticiens des centres de lutte contre le cancer (CLCC)*, Paris, France, 18-20 mai 2022.

Diffusion de la culture scientifique

Participation au concours *Ma thèse en 180 secondes*, sélection locale et régionale, édition 2021

Table des matières

I Introduction **1**

1 Essai randomisé en clusters (ERC) 1

1.1 Définition 1

1.2 Pourquoi randomiser des clusters ? 2

1.3 Particularités statistiques des essais randomisés en clusters 3

1.3.1 Coefficient de corrélation intra-classe 3

1.3.2 Calcul du nombre de sujets nécessaires 4

1.3.3 Méthodes d'analyses statistiques 5

2 Analyse de données de survie 10

2.1 Notations et définitions 10

2.1.1 Terminologie 10

2.1.2 Distribution du temps de survie 10

2.1.3 Données censurées 11

2.1.4 Notations 11

2.2 Estimations non paramétriques 11

2.3 Estimations semi-paramétriques et paramétriques 12

2.3.1 Modèles à risques proportionnels 12

2.3.2 Modèle de Cox : une approche semi-paramétrique 13

2.3.3 Approche paramétrique 13

2.3.4 Vérification de l'hypothèse des risques proportionnels 13

2.3.5 Alternatives en cas de risques non-proportionnels 13

2.4 Différence de temps de survie moyen restreint (RMST) 14

2.4.1 Définitions 15

2.4.2 Méthodes d'estimations 15

3 Qu'en est-il de l'analyse des données de survie dans les ERC ? 18

3.1 Méthodes pour estimer un HR dans le cadre des ERC 18

3.1.1 Modèles de survie conditionnels 18

3.1.2 Modèles de survie marginaux 19
3.2 Autres méthodes 19
3.3 Méthodes pour estimer un RMST dans le cadre des ERC 19
4 Objectifs de la thèse 21

II Estimation d'une différence de RMST dans le cadre des ERC 23
1 Présentation du travail 23
1.1 Contexte 23
1.2 Méthodes 23
1.3 Résultats 24
1.4 Application 24
1.5 Conclusion 25
2 Article n°1 25

III Estimation d'une différence de RMST avec la régression sur des pseudo-valeurs pour les ERC avec un faible nombre de clusters 61
1 Présentation du travail 61
1.1 Contexte 61
1.2 Méthodes 62
1.3 Résultats 62
1.4 Application 63
1.5 Conclusion 63
2 Article n°2 (en préparation) 64

IV Travaux annexes liés aux ERC 123
1 Analyse de l'essai randomisé en clusters FADEMELA 123
1.1 Rationnel et objectif de l'étude 123
1.2 Méthodologie de l'étude 124
1.2.1 Schéma d'étude 124
1.2.2 Déroulement de l'étude 124
1.2.3 Critère de jugement principal 124
1.2.4 Nombre de sujets nécessaires 124
1.2.5 Analyses statistiques 125
1.3 Résultats 125
1.4 Conclusion 126
2 Difficultés de recrutement et d'implémentation de l'intervention dans les ERC en *"stepped-wedge"* 126

2.1 Rationnel et objectif 126
2.2 Méthodes 127
2.3 Résultats 127
2.4 Conclusion 128

V Discussion générale **129**

1 Synthèse des principaux travaux 129
2 Recommandations 129
3 Limites et perspectives 130
3.1 Ajustement sur des covariables 130
3.2 Autre méthode d'estimation du RMST dans les ERC 130
3.3 Utilisation des méthodes proposées en pratique 131
3.4 Extensions des méthodes 132
3.5 Revue de la littérature des méthodes d'analyses de survie dans les ERC 133
4 Conclusion générale 133

Références **142**

A Algorithme d'estimation des paramètres avec les équations d'estimation généralisées (GEE) **143**

B Analyse de l'essai randomisé en clusters FADEMELA : Article **145**

C Difficultés de recrutement et d'implémentation de l'intervention dans les ERC en *"stepped-wedge"* : Article **159**

Liste des figures

I.1 Allocation des clusters aux groupes intervention et contrôle dans un essai randomisé en clusters en deux bras parallèles 1

I.2 Interprétation du coefficient de corrélation intra-classe (CCI) 4

I.3 Synthèse des différences entre l'approche conditionnelle et marginale pour une analyse au niveau individuel d'un ERC 8

IV.1 Résumé de l'étude FADEMELA 125

IV.2 Essai randomisé en clusters en *stepped-wedge* 126

Abréviations

CCI	Coefficient de Corrélation Intraclasse
CONSORT	*CONsolidated Standards of Reporting Trials*
EM	*Expectation Maximisation*
ERC-SW	Essai Randomisé en Clusters en *Stepped-Wedge*
ERC	Essai Randomisé en Clusters
GEE	*Generalized Estimating Equations* - équations d'estimation généralisées
HR	*Hazard ratio*
IPCW	*Inverse probability of censoring weighting*
RMST	*Restricted Mean Survival Time* - temps de survie moyen restreint

Chapitre I

Introduction

1 Essai randomisé en clusters (ERC)

1.1 Définition

Un essai randomisé en clusters (ERC) est un essai dans lequel on randomise des groupes d'individus, et non pas des individus [1]. Ces groupes d'individus sont nommés clusters ou parfois grappes. Les clusters peuvent être, par exemple, des familles, des hôpitaux, des écoles ou encore des zones géographiques [2]. La plupart du temps, les clusters sont des entités naturellement présentes, c'est-à-dire qu'ils ne sont pas créés spécifiquement pour les besoins de l'essai [3].

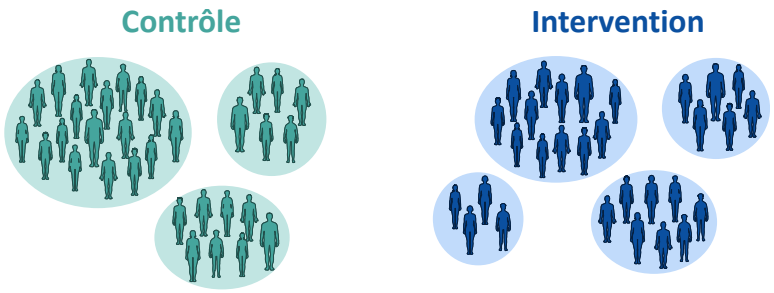


FIGURE I.1 – Allocation des clusters aux groupes intervention et contrôle dans un essai randomisé en clusters en deux bras parallèles

Les ERC sont communément utilisés dans la recherche interventionnelle en santé publique [4]. Ils sont particulièrement mis en place afin d'évaluer des interventions de prévention, organisationnelles ou visant à changer les pratiques des professionnels de santé mais rarement pour évaluer des traitements pharmacologiques.

L'utilisation de ce type d'étude a connu une forte augmentation au cours des dernières décennies [5]. À titre d'illustration, une rapide recherche sur *Web of Science* a conduit aux résultats suivants : le nombre d'articles par an avec le terme "cluster randomi..." dans leur titre est passé de 2 en 1992 à 35 en 2002, puis 312 en 2012, pour atteindre 1080 en 2022. Une extension du CONSORT (CONsolidated

Standards of Reporting Trials) pour les ERC a d'ailleurs été publiée en 2004 [6], puis mise à jour en 2012 [7], afin de fournir des recommandations pour la rédaction d'articles rapportant des ERC.

De nombreux schémas d'étude existent pour les ERC comme dans les essais à randomisation individuelle (en deux bras parallèles, en cross-over, stepped-wedge) [8].

Le présent travail de thèse se focalise sur les ERC en deux bras parallèles (Figure I.1).

1.2 Pourquoi randomiser des clusters ?

Le choix d'un ERC peut être justifié par plusieurs raisons :

- ***Application de l'intervention au niveau du cluster***
  Dans certains cas, l'intervention est délivrée directement à l'ensemble du cluster et il est impossible de la délivrer uniquement à une partie des individus. Par exemple, afin de réduire les blessures liées aux chutes des personnes âgées, Drahota *et al.* [9] ont voulu évaluer l'effet d'un changement du revêtement des sols absorbant les chocs dans les services des hôpitaux accueillant des personnes âgées. Le revêtement du sol affectait donc l'ensemble des patients d'un même service. Dans cette étude, la randomisation des patients était impossible et l'unité de randomisation ne pouvait être que les hôpitaux ou les services même si les retombées potentielles étaient mesurées au niveau des individus. Le choix du schéma d'étude peut donc être directement influencé par la nature de l'intervention elle-même.

- ***Limitation de la contamination au niveau des professionnels de santé***
  Les ERC permettent de limiter une potentielle contamination entre les groupes lorsque l'intervention est délivrée aux professionnels de santé. Par exemple, dans l'essai OEDIPUS visant à évaluer l'efficacité d'un programme d'éducation sur le contrôle de l'asthme destiné à la minorité ethnique d'Asie du Sud à Londres, l'intervention incluait un programme de formation dédié aux professionnels de santé [10]. Dans cette étude, si un essai à randomisation individuel avait été choisi, les patients du groupe contrôle auraient été pris en charge par les mêmes professionnels de santé que les patients recevant l'intervention et auraient probablement été également affectés par l'intervention. En effet, il est difficile d'imaginer qu'un professionnel de santé soigne certains patients comme s'il avait reçu le programme de formation et d'autres comme s'il ne l'avait pas reçu, donnant des recommandations différentes pour des patients ayant le même profil. Dans cet exemple, même s'il était possible de randomiser les patients, il ne semblait pas souhaitable de le faire à cause du risque de contamination intergroupe, et les cabinets médicaux ont été randomisés.

- ***Limitation de la contamination entre les individus d'un cluster***
  Randomiser des clusters plutôt que des individus permet également d'éviter la contamination entre les membres d'un même cluster. Par exemple, Naldi *et al.* [11] ont mené une étude afin d'évaluer l'impact d'un programme de sensibilisation à la photoprotection sur la survenue de coups de soleil. Des écoles étaient sélectionnées aléatoirement pour bénéficier ou non d'un programme éducatif dispensé par des enseignants à l'école. Si les élèves avaient été répartis aléatoirement entre le groupe recevant l'intervention et le groupe contrôle, les élèves d'une même école auraient pu faire partie des deux groupes. L'interaction entre les élèves à l'école aurait entraîné la diffusion de la sensibilisation aux élèves du groupe contrôle, qui auraient été exposés, même

partiellement, à l'intervention.

La contamination peut aussi être au sens médical du terme. Kelly *et al.* [12] ont testé l'hypothèse selon laquelle la défense contre les infections intestinales dépend du statut en micronutriments. Des familles étaient randomisées pour recevoir des comprimés de micronutriments ou des comprimés placebos. Le fait que tous les membres d'une même famille reçoivent les mêmes comprimés a permis de prévenir, dans le groupe intervention, des infections intestinales dues à l'échec du placebo. L'ERC a donc permis de minimiser le risque de contamination au sein de la famille.

- ***Évaluation de l'effet indirect d'une intervention***
  Les ERC ont un intérêt particulier dans le milieu de la vaccination. Ils permettent d'évaluer l'effet indirect chez les individus non-vaccinés, en plus de l'effet direct du vaccin chez les individus vaccinés. En effet, les individus non-vaccinés évoluant avec des personnes vaccinées sont moins susceptibles de contracter la maladie. Par exemple, Lemaitre *et al.* [13] ont évalué l'impact de la vaccination contre la grippe du personnel des maisons de retraite sur la mortalité des résidents à l'aide d'un ERC.

- ***Faisabilité logistique, administrative ou financière***
  Le choix d'un ERC peut être fait afin de faciliter la réalisation et la mise en place de l'essai, en termes de coût ou de logistique. Par exemple, Murira *et al.* [14] ont comparé deux systèmes de prise en charge anténatale au Zimbabwe : le système actuel qui prévoit 12 visites durant une grossesse et un nouveau système avec 6 visites. La mise en place d'un seul système au sein d'une même clinique était plus simple. Par ailleurs, si les interventions nécessitent des équipements coûteux, randomiser des clusters permet de ne pas équiper chaque cluster et de réduire les coûts.

1.3 Particularités statistiques des essais randomisés en clusters

Dans un ERC, les observations ne sont pas indépendantes au sein d'un même cluster. Les réponses de deux individus d'un même cluster sont plus semblables que celles de deux individus appartenant à deux clusters différents. Cette corrélation doit être quantifiée et prise en compte lors de la planification et l'analyse statistique de l'étude.

1.3.1 *Coefficient de corrélation intra-classe*

La ressemblance entre les réponses des individus d'un même cluster est généralement quantifiée par le coefficient de corrélation intra-classe (CCI). Le CCI est défini comme la part de la variance totale du critère de jugement due à la variance des réponses entre clusters. Théoriquement, le CCI prend donc ses valeurs entre 0 et 1 et n'a pas d'unité [15].

Interprétation du CCI

Le CCI est une mesure de l'homogénéité des réponses au sein d'un cluster. Il peut être interprété comme la corrélation entre les réponses de deux individus qui appartiennent à un même cluster, quels que soient les individus et quel que soit le cluster [16]. Un CCI nul signifie que les observations à l'intérieur des clusters sont indépendantes. Il n'y a pas plus de ressemblance entre les observations dans un cluster que dans l'ensemble des observations (Figure I.2). On n'observe pas de structure particulière au sein de chaque cluster. Au contraire, si le CCI vaut 1, toutes les observations à l'intérieur d'un cluster sont

identiques. Dans ce cas, l'information apportée par un cluster est la même que celle apportée par un seul des individus du cluster.

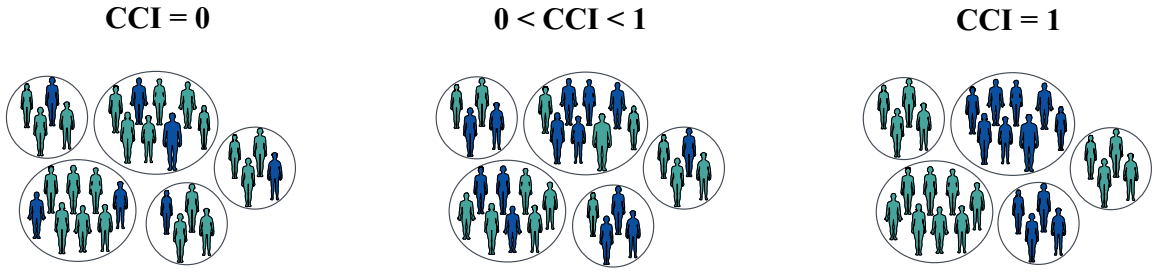


FIGURE I.2 – Interprétation du coefficient de corrélation intra-classe (CCI)
On suppose une situation où le critère de jugement est binaire et la couleur de l'individu reflète sa réponse. Par exemple, un individu est en vert s'il a un critère de jugement égal à 0 et en bleu s'il a un critère de jugement est égal à 1.

Facteurs influençant le CCI

Différents facteurs influençant le CCI ont été identifiés. Campbell *et al.* [17] ont montré que le CCI diminue lorsque la taille du cluster augmente (la ressemblance à l'échelle d'une famille est plus grande que celle à l'échelle d'un hôpital). De plus, dans le cas d'un critère de jugement binaire, le CCI dépend de la prévalence de ce même critère [18, 19, 20]. Les prévalences extrêmes (proche de 0 ou 1), sont associées à de plus faibles CCI. Enfin, le CCI dépend également du type du critère de jugement. Le critère de jugement peut être de type "process", c'est-à-dire qu'il décrit un comportement (par exemple la prise en charge des patients), ou de type "outcome", c'est-à-dire qu'il décrit l'impact direct sur le patient (par exemple la tension). Les CCI pour des critères de type "outcome" sont inférieurs aux CCI associés aux critères de type "process" [21]. En effet, une standardisation des pratiques professionnelles n'entraîne pas forcément une uniformisation des réponses individuelles. Les critères de type "outcome", qui dépendent à la fois des pratiques et des caractéristiques individuelles, vont avoir une plus grande variabilité que les critères de type "process" qui dépendent uniquement des pratiques.

Estimation du CCI

La méthode d'estimation ponctuelle la plus couramment utilisée est celle de l'analyse de la variance d'un modèle à effets aléatoires à un facteur. D'autres méthodes d'estimation ponctuelle existent. Donner [22] a détaillé différents estimateurs pour les critères de jugement continus et Ridout [23] pour les critères de jugement binaires. Concernant les différentes méthodes d'estimation de l'intervalle de confiance, elles ont été décrites par Ukoumunne [24]. À noter qu'en pratique, le CCI est généralement faible dans les ERC [25, 21, 26, 27].

1.3.2 *Calcul du nombre de sujets nécessaires*

Les informations apportées par n individus dans un ERC sont moindres que par n individus d'un essai à randomisation individuelle, à cause de la corrélation entre les réponses des individus au sein d'un même cluster. Un effectif plus grand est donc nécessaire dans un ERC pour atteindre la même puissance que dans un essai à randomisation individuelle. Le nombre de sujets nécessaires pour un

ERC est généralement calculé en multipliant le nombre de sujets nécessaires d'un essai à randomisation individuelle par un facteur d'inflation, noté IF, défini de la manière suivante [28] :

$$IF = 1 + (m - 1)\rho \tag{1.1}$$

avec m le nombre d'individus par cluster et ρ le CCI attendu a priori. On peut remarquer que plus le CCI est élevé, c'est-à-dire que plus les réponses des individus d'un même cluster se ressemblent, plus le nombre de sujets nécessaires sera important. Ainsi, si la valeur du CCI retenue lors du calcul du nombre de sujets nécessaires est inférieure à celle effectivement obtenue, la puissance sera inférieure à la puissance planifiée [29].

Le facteur d'inflation (1.1) repose sur l'hypothèse que la taille des clusters, m, est identique pour chaque cluster. Cette hypothèse semble peu réaliste et si la taille des clusters est variable alors le nombre de sujets nécessaires sera sous-estimé en utilisant le facteur d'inflation (1.1). Si la variabilité de la taille des clusters est faible, alors le facteur d'inflation peut être calculé en remplaçant m par la taille moyenne des clusters, $\bar{m}$ dans (1.1) [8]. Dans le cas contraire, il est préférable d'utiliser le facteur d'inflation suivant [30] :

$$IF = 1 + \left[(1 + CV_m^2)\bar{m} - 1\right]\rho \tag{1.2}$$

avec CV_m le coefficient de variation de la taille des clusters qui est défini comme $CV_m = \frac{s_m}{\bar{m}}$ où s_m est l'écart-type de la taille des clusters.

Finalement, lorsque le nombre total de clusters est faible (en général en dessous de 30), le nombre de sujets nécessaires pour un essai randomisé individuel, qui est ensuite multiplié par le facteur d'inflation, doit être calculé à partir des valeurs critiques d'une distribution de Student plutôt que d'une distribution normale, afin d'éviter une sous-estimation du nombre de sujets nécessaires de l'ERC [8].

1.3.3 *Méthodes d'analyses statistiques*

Les ERC ont la particularité d'avoir une unité de randomisation, le cluster, en général différente de l'unité observée, l'individu. Ainsi, les observations ne sont pas indépendantes et la corrélation entre les réponses des individus au sein d'un même cluster doit être prise en compte lors de l'analyse afin d'éviter une sous-estimation de la variance de l'effet intervention et donc une inflation de l'erreur de type I [8]. Les ERC ont la particularité de pouvoir être analysés à deux niveaux : au niveau du cluster ou au niveau de l'individu.

1.3.3.1 Analyse au niveau cluster

L'analyse peut être réalisée au niveau du cluster si le critère de jugement est mesuré directement au niveau du cluster ou alors en utilisant une mesure synthétisant les réponses de chaque individu (valeur moyenne ou proportion du critère de jugement pour chaque cluster par exemple). Les clusters étant indépendants les uns des autres, des méthodes statistiques conventionnelles peuvent alors être utilisées sur les mesures de chaque cluster (par exemple un test de Student [8]). Les tests peuvent aussi être pondérés pour prendre en compte la variabilité de la taille des clusters [8]. Un ajustement sur les covariables au niveau du cluster est également possible, mais il est compliqué d'ajuster sur des covariables individuelles [31].

1.3.3.2 Analyse au niveau individuel

La pratique la plus courante, et sur laquelle porte ce travail, est l'analyse au niveau individuel [32]. Dans ce cas, le critère de jugement a été mesuré pour chaque individu. Il n'est plus possible de supposer que les réponses sont indépendantes au sein des clusters et par conséquent d'utiliser les méthodes statistiques habituelles. On peut alors utiliser des modèles de régression pour prendre en compte la corrélation entre les réponses des individus d'un même cluster. Deux types de modèles de régression sont utilisés pour analyser un ERC : les modèles conditionnels ou les modèles marginaux [8].

Deux approches, deux modélisations, deux méthodes d'estimation

Afin d'illustrer la différence entre les deux familles de modèles de régression, nous nous restreignons au cas précis d'un critère de jugement continu sans ajustement sur des covariables autres que le groupe. Des approches similaires peuvent être utilisées pour des critères de jugement d'une autre nature en adaptant la modélisation.

Par la suite, nous considérons un échantillon de K clusters avec m_k individus dans le cluster k ($k = 1, ..., K$).

Approche conditionnelle

Avec l'approche conditionnelle, la dépendance des réponses des sujets d'un même cluster est prise en compte grâce à l'introduction d'un effet aléatoire "cluster" dans le modèle. Ainsi, le modèle considéré est le suivant :

$$Y_{ki} = \beta_0 + \beta_1 Z_{ki} + b_k + e_{ki} \tag{1.3}$$

où

- Y_{ki} est la valeur du critère de jugement de l'individu i du cluster k
- β_0 est l'intercept
- β_1 est l'effet intervention
- Z_{ki} est une variable binaire qui vaut 1 si le cluster k appartient au groupe intervention et 0 si le cluster k appartient au groupe contrôle
- b_k l'effet moyen du cluster k (effet aléatoire)
- e_{ki} correspond au résidu de l'individu i du cluster k.

On suppose que $b_k \sim \mathcal{N}(0, \sigma_b^2)$ où σ_b^2 est la variance entre les clusters et $e_{ki} \sim \mathcal{N}(0, \sigma_w^2)$ avec σ_w^2 la variance intra-cluster. Les effets aléatoires "cluster", b_k, sont supposés indépendants. On fait également l'hypothèse que b_k et e_{ki} sont indépendants. Dans l'approche conditionnelle, les résidus intra-clusters, e_{ki}, sont supposés indépendants.

L'approche conditionnelle correspond finalement à un modèle à effets mixtes classique. Plusieurs méthodes d'estimation existent, toutes basées sur le maximum de vraisemblance [33, 34].

Approche marginale

Avec l'approche marginale, en gardant les mêmes notations que précédemment, le modèle suivant est considéré :

$$Y_{ki} = \beta_0 + \beta_1 Z_{ki} + e_{ki} \tag{1.4}$$

Contrairement au modèle conditionnel, les résidus intra-clusters, e_{ki}, sont corrélés.

Les estimations des paramètres du modèle marginal sont généralement obtenues à l'aide d'équations d'estimation généralisées, ou *generalized estimating equations* (GEE), développées par Liang et Zeger [35, 36] :

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \mathbf{D}_k' \mathbf{V}_k^{-1} \left(\mathbf{Y}_k - \boldsymbol{\mu}_k\right) = 0 \tag{1.5}$$

avec $Y_k = (Y_{k1}, .., Y_{km_k})$, $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{km_k})$, $\mu_{ki} = \beta_0 + \beta_1 Z_{ki}$, $\mathbf{D}_k = \frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ et $\boldsymbol{V}_k$ la matrice de covariance de travail du cluster k de taille $m_k \times m_k$. Elle est définie comme :

$$\mathbf{V}_k = \phi \mathbf{A}_k^{\frac{1}{2}} \mathbf{R}(\rho) \mathbf{A}_k^{\frac{1}{2}} \tag{1.6}$$

où ϕ est un paramètre de dispersion, $\mathbf{R}(\rho)$ une matrice de corrélation de travail de taille $m_k \times m_k$ et $\mathbf{A}_k$ une matrice diagonale de taille $m_k \times m_k$ avec comme éléments diagonaux $Var(\mu_{ki})$.

La structure de la matrice de corrélation de travail, $\mathbf{R}(\rho)$, qui décrit la corrélation entre les réponses des individus au sein d'un même cluster, doit être spécifiée. En général, dans un essai randomisé en clusters, il n'y a pas de raison de penser que la réponse d'un individu est plus corrélée à la réponse de certains individus du cluster que d'autres. Ainsi, une matrice de corrélation de travail échangeable est généralement utilisée, qui suppose que la corrélation des réponses pour toutes les paires d'individus d'un même cluster est identique et que cette corrélation est la même pour tous les clusters [8]. Elle est définie comme suit :

$$\mathbf{R}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \tag{1.7}$$

avec $\rho = Corr(Y_{ki}, Y_{kj}) \forall i \neq j, \forall k$ [35, 36]. D'autres matrices ont également été proposées par Liang et Zeger comme la matrice de corrélation de travail indépendante qui fait l'hypothèse que les réponses au sein d'un clusters sont indépendantes et qui correspond à la matrice identité [35, 36].

Les estimations des paramètres sont obtenues en résolvant les équations (1.5) avec l'algorithme itératif disponible en Annexe A. La matrice de variances-covariance des paramètres est généralement estimée à l'aide de l'estimateur dit "robuste" ou "sandwich", défini de la manière suivante :

$$\hat{\mathbf{V}} = \boldsymbol{\Omega}^{-1} \left[\sum_{k=1}^{K} \mathbf{D}_k' \mathbf{V}_k^{-1} \left(\mathbf{Y}_k - \hat{\boldsymbol{\mu}}_k\right) \left(\mathbf{Y}_k - \hat{\boldsymbol{\mu}}_k\right)' \mathbf{V}_k^{-1} \mathbf{D}_k \right] \boldsymbol{\Omega}^{-1} \tag{1.8}$$

avec $\boldsymbol{\Omega} = \sum_{k=1}^{K} \mathbf{D}_k' \mathbf{V}_k^{-1} \mathbf{D}_k$.

Les estimations des paramètres $\boldsymbol{\beta}$ et l'estimateur de variance sandwich $\hat{\mathbf{V}}$ sont robustes au choix de la matrice de corrélation de travail $\mathbf{R}(\rho)$. Ils sont consistants même si la matrice de corrélation est mal spécifiée [35, 36].

Finalement, on peut noter qu'il est facile d'ajuster sur des covariables avec les deux approches, conditionnelle et marginale, en les intégrant dans les modèles (1.3) et (1.4).

Deux approches, deux interprétations

Les deux méthodes aboutissent à des estimations de l'effet intervention qui n'ont pas la même interprétation (Figure 1.3). Le modèle à effets mixtes évalue l'impact sur le critère de jugement principal

qu'aurait le passage d'un individu du bras contrôle au bras intervention, conditionnellement à l'appartenance au cluster. C'est un effet dit *"cluster-specific"*, car l'effet intervention est conditionnel au cluster et diffère selon le cluster auquel appartient l'individu. Dans le cas du modèle GEE, l'impact sur le critère de jugement principal concerne le passage d'un individu du bras contrôle au bras intervention quel que soit le cluster auquel il appartient. C'est un effet marginal (*"population-average"*) car l'effet intervention est directement interprétable à l'échelle de la population entière.

Les deux approches mènent donc à des interprétations différentes et le choix de l'une d'entre elles doit se faire en tenant compte de la question posée par l'essai.

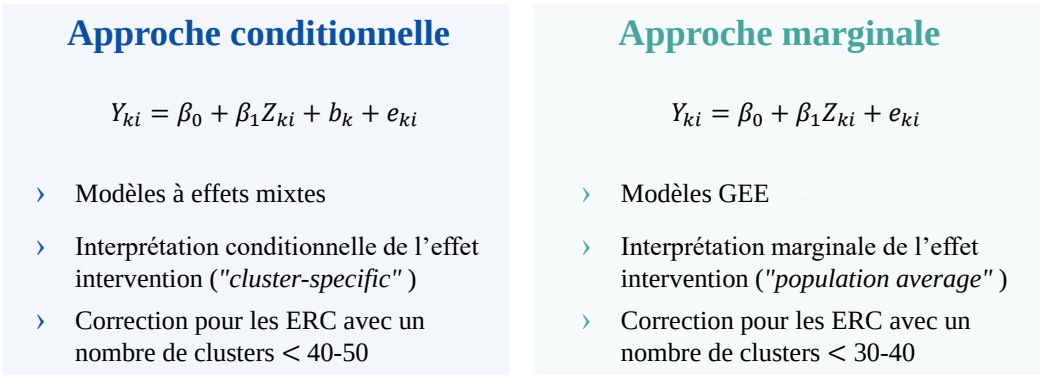


FIGURE I.3 – Synthèse des différences entre l'approche conditionnelle et marginale pour une analyse au niveau individuel d'un ERC

1.3.3.3 Analyse d'un ERC avec un petit nombre de clusters

L'analyse au niveau du cluster est robuste au nombre total de clusters [37]. Cependant, les deux approches pour l'analyse au niveau individuel ne sont valides que lorsqu'il y a un nombre de clusters suffisamment grand. Si l'étude comporte un très faible nombre de clusters, il y a un risque d'inflation de l'erreur de type I. Il n'y a pas de limite claire, mais il est généralement recommandé d'avoir au moins 30 à 40 clusters pour les modèles à effets mixtes et au moins 40 à 50 clusters pour la méthode GEE [8, 38, 1, 37] (Figure I.3). Cependant, il n'est pas rare que les ERC randomisent très peu de clusters. En effet, plusieurs revues ont montré que les ERC comptent souvent moins de 40 clusters au total [39, 40, 41]. Des solutions ont été proposées pour les deux approches.

Pour les modèles mixtes, il est suggéré de changer le nombre de degrés de liberté de la distribution de Student utilisée lors de l'inférence statistique pour estimer l'effet intervention. Différentes méthodes ont été proposées pour déterminer le nombre de degrés de liberté : choisir le nombre total de clusters moins la taille du vecteur des effets fixes [42], la méthode de Sattherwaite [43] (approximation du nombre de degrés de liberté basée sur les deux premiers moments) ou encore la méthode de Kenward-Roger [44] (utilise l'approximation de Satterthwaite après avoir appliqué à la statistique de test un facteur d'échelle basé sur une estimation de la matrice de covariance pour un petit échantillon).

Dans le cadre de la méthode GEE, des estimateurs permettant de corriger la variance ont été proposés et comparés dans de nombreuses études de simulation pour différents types de critères de jugement [45, 46, 47, 48, 49, 50, 51, 52, 53, 54]. Une autre possibilité, qui peut être combinée avec les méthodes de correction de la variance, est d'estimer l'intervalle de confiance et la p-valeur à l'aide de la distribution

de Student plutôt que normale [53, 54, 55]. Le nombre de degrés de liberté peut être choisi comme le nombre total de clusters mois la taille du vecteur des coefficients de régression ou par la méthode Sattherwaite [52].

2 Analyse de données de survie

Cette thèse se focalise sur les critères de jugement de type survie, c'est-à-dire le temps écoulé jusqu'à la survenue d'un événement d'intérêt (communément le décès, d'où le terme "survie"). Dans le domaine de la recherche clinique, plusieurs événements peuvent être étudiés, comme par exemple la guérison, l'apparition ou la rechute d'une maladie ou encore le décès.
L'analyse de données de survie correspond à l'étude du délai de survenue de cet événement d'intérêt. On cherche à estimer la distribution des temps de survie, à comparer ces distributions entre différents groupes d'individus et/ou déterminer l'effet de variables explicatives sur ces distributions.

2.1 Notations et définitions

2.1.1 *Terminologie*

Soit T^*, le temps de survie, défini comme une variable aléatoire réelle, positive et continue. Ce temps de survie correspond au temps écoulé entre une date d'origine et la date de survenue de l'événement. La date d'origine correspond à l'origine de la durée étudiée qui peut être par exemple la date d'entrée dans l'étude. Cette date est donc propre à chaque individu. On définit également la date de dernière nouvelle comme la dernière date à laquelle on dispose d'une information sur l'individu, à savoir, s'il a connu l'évènement ou non. Le délai écoulé entre la date d'origine et la date de dernière nouvelle est appelé le délai de suivi. Enfin, on définit la date de point comme la date d'arrêt de l'étude. On ne tient compte d'aucune information ultérieure à cette date, commune à l'ensemble des individus.

2.1.2 *Distribution du temps de survie*

La distribution du temps de survie peut être spécifiée par différentes fonctions de manière équivalente :

- ***La fonction de densité*** f
  Elle correspond à la probabilité que l'événement survienne entre t et $t+dt$ (dt tendant vers 0) :

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T^* < t+dt)}{dt}$$

- ***La fonction de répartition*** F
  Elle correspond à la probabilité de présenter l'événement avant le temps $t \geq 0$:

$$F(t) = P(T^* < t) = \int_0^t f(u)du$$

  La fonction de répartition est positive, monotone et croissante avec $F(0) = 0$ et $\lim_{t \to +\infty} F(t) = 1$.

- ***La fonction de survie*** S
  Elle correspond à la probabilité de ne pas présenter l'événement avant le temps $t \geq 0$:

$$S(t) = P(T^* \geq t) = 1 - F(t) = \int_t^{+\infty} f(u)du$$

  La fonction de survie est monotone et décroissante avec $S(0) = 1$ et $\lim_{t \to +\infty} S(t) = 0$.

- ***La fonction de risque instantané*** h (ou *hazard function*)
  Elle correspond à la probabilité de survenue de l'événement entre t et $t+dt$ (dt tendant vers 0) conditionnellement au fait que l'événement ne soit pas survenu avant t :

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T^* < t + dt | T^* \geq t)}{dt} = \frac{f(t)}{S(t)} = -\frac{d \ln(S(t))}{dt}$$

- ***La fonction de risque cumulé*** H
  Elle est définie comme

$$H(t) = \int_0^t h(u)du = -\ln(S(t))$$

  On en déduit :

$$S(t) = exp(-H(t)) = exp\left(-\int_0^t h(u)du\right)$$

2.1.3 *Données censurées*

Les données de survie ont comme particularité que l'événement d'intérêt ne peut pas toujours être observé : on parle de données censurées. Cette particularité est à prendre en compte lors de l'analyse statistique. Le cas le plus fréquent est la censure à droite qui correspond à la situation où l'individu n'a pas connu l'événement avant à la fin de l'étude ou a été "perdu de vue". Dans ce cas, les temps de survie ne sont pas tous connus et pour certains individus, on sait juste que le temps de survie est supérieur au délai du suivi. Dans le cas où l'individu a présenté l'événement avant le début du suivi, on parle de censure à gauche. Dans ce cas, on sait uniquement que la date de l'événement est antérieure à la date du début du suivi. Enfin, le dernier cas est celui de la censure par intervalle, où la date de survenue de l'événement a eu lieu entre deux dates connues (par exemple entre deux visites de suivi d'un individu) mais elle n'est pas précisément connue.

Ce travail se place dans le cadre de censure à droite qui est la situation la plus courante.

2.1.4 *Notations*

Par la suite, on considère un échantillon de N individus. Chaque individu i ($i = 1, ..., N$) a un temps de survie T_i^* et on note C_i le temps jusqu'à la censure. Les données de survie observées pour l'individu i sont (T_i, δ_i) où $T_i = min(T_i^*, C_i)$ et δ_i est l'indicatrice de l'événement avec $\delta_i = I_{T_i^* \leq C_i}$. Si $\delta_i = 1$, l'événement est survenu et observé. Si $\delta_i = 0$, l'individu est censuré et on sait uniquement que le temps de survie est supérieur à son délai de suivi.

2.2 Estimations non paramétriques

Dans le cadre de l'analyse de données de survie, de nombreux estimateurs non paramétriques ont été développés. L'estimateur non paramétrique le plus connu pour estimer la fonction de survie est l'estimateur de Kaplan-Meier, défini comme suit [56] :

$$\hat{S}_{KM}(t) = \prod_{i=1,...,n; T_i \leq t} \left(1 - \frac{d_i}{N_i}\right) \tag{2.1}$$

où d_i est le nombre de patients présentant l'événement au temps T_i et N_i le nombre de patients encore à risque au temps T_i. L'estimateur $\hat{S}_{KM}(t)$ est une fonction en escalier décroissante continue à droite. La variance de l'estimateur de Kaplan-Meier est généralement estimée avec l'estimateur de Greenwood [57]. Si l'on s'intéresse plutôt à la fonction de risque cumulé, l'estimateur de Nelson-Aalen [58, 59] est principalement utilisé. D'autres estimateurs existent comme l'estimateur de Breslow [60] pour la fonction de risque cumulé ou l'estimateur de Harrington-Fleming [61] pour la fonction de survie.

2.3 Estimations semi-paramétriques et paramétriques

2.3.1 *Modèles à risques proportionnels*

En analyse de données de survie, les modèles à risques proportionnels sont les modèles les plus couramment utilisés. Ils permettent d'étudier l'effet de covariables sur le risque de survenue d'un événement. La fonction de risque instantané de l'individu i, h_i, est modélisée en fonction d'un vecteur de covariables $\mathbf{Z}_i$, pour tout $t > 0$, comme suit :

$$h_i(t|\mathbf{Z}_i) = h_0(t)g(\boldsymbol{\beta}'\mathbf{Z}_i) \tag{2.2}$$

où h_0 est la fonction de risque instantané de base, $\boldsymbol{\beta}$ le vecteur des coefficients de régression associés aux covariables et g une fonction positive. La fonction h_0 est commune à tous les individus et correspond au risque instantané de survenue de l'événement pour un individu dont toutes les covariables sont fixées à leur valeur de référence. On peut remarquer que la fonction de risque instantané a deux composantes : une partie dépendant uniquement de l'effet du temps (h_0) et une autre partie dépendant uniquement des covariables ($g(\boldsymbol{\beta}'\mathbf{Z}_i)$).

Le plus souvent, une fonction exponentielle est choisie pour g afin d'obtenir une fonction de risque instantané positive et sans contrainte sur les valeurs des coefficient $\boldsymbol{\beta}$, amenant au modèle suivant :

$$h_i(t|\mathbf{Z}_i) = h_0(t)exp(\boldsymbol{\beta}'\mathbf{Z}_i) \tag{2.3}$$

L'interprétation des résultats avec les modèles à risques proportionnels se fait à l'aide du rapport des risques instantanés, ou *hazard ratio* (HR). Ce HR correspond au rapport des deux fonctions de risque instantané pour deux individus distincts i et j, qui ont pour covariables $\mathbf{Z}_i$ et $\mathbf{Z}_j$, respectivement :

$$\frac{h_i(t|\mathbf{Z}_i)}{h_j(t|\mathbf{Z}_j)} = \frac{h_0(t)g(\boldsymbol{\beta}'\mathbf{Z}_i)}{h_0(t)g(\boldsymbol{\beta}'\mathbf{Z}_j)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\exp(\boldsymbol{\beta}'\mathbf{Z}_j)}$$

Ce rapport mesure le risque relatif à l'instant t de présenter un événement pour l'individu présentant les caractéristiques $\mathbf{Z}_i$ par rapport à l'individu présentant les caractéristiques $\mathbf{Z}_j$. Ainsi, dans le cas où on a une unique covariable qui est le groupe auquel appartient l'individu ($Z_i = 1$ si l'individu appartient au groupe intervention et 0 au groupe contrôle), l'effet intervention est alors quantifié par $HR = exp(\beta)$ qui représente le rapport des risques instantanés entre les individus des deux groupes. On peut noter que ce rapport est constant au cours du temps. Les fonctions de risque instantané sont donc proportionnelles, d'où l'appellation de modèle "à risques proportionnels". Cette conséquence inhérente au modèle est aussi une hypothèse qu'il faut vérifier pour pouvoir l'utiliser (voir section 2.3.4).

2.3.2 *Modèle de Cox : une approche semi-paramétrique*

Le modèle le plus couramment utilisé en analyse de survie est le modèle de Cox [62]. C'est un modèle à risques proportionnels semi-paramétrique car il suppose que la fonction g du modèle 2.2 est la fonction exponentielle mais il ne fait aucune hypothèse sur la forme de la fonction de risque instantané de base h_0. Les coefficients de régression $\boldsymbol{\beta}$ sont estimés en maximisant la vraisemblance dite partielle du modèle, qui ne dépend pas de la fonction h_0 [63]. Il n'est donc pas nécessaire de spécifier et d'estimer la fonction de risque instantané de base. L'estimation de $\boldsymbol{\beta}$ par une telle méthode repose sur l'hypothèse qu'il n'y a pas d'ex aequo, c'est-à-dire que les temps d'événement sont distincts. S'il y a des événements simultanés, la vraisemblance totale peut être appromixée par l'approximation de Breslow [64], le plus couramment, ou d'Efron [65].

2.3.3 *Approche paramétrique*

Avec des modèles à risques proportionnels, il est également possible de spécifier la fonction de risque instantané de base h_0 dans le modèle (2.3) afin d'obtenir un modèle paramétrique. La fonction h_0 peut prendre différentes formes paramétriques en suivant par exemple une distribution exponentielle, qui suppose une fonction de risque instantané de base constante au cours du temps, de Weibull ou Gompertz [66]. On peut également spécifier la forme de h_0 à l'aide de splines [66].

2.3.4 *Vérification de l'hypothèse des risques proportionnels*

Les modèles à risques proportionnels reposent sur l'hypothèse de proportionnalité des risques qui doit être vérifiée. Pour cela, si l'on compare deux groupes, on peut tracer les courbes $\log(-\log(S(t)))$ de chaque groupe en fonction du temps sur un même graphique [67]. Si les deux courbes sont parallèles, alors l'hypothèse de proportionnalité des risques est respectée. Pour un modèle de Cox, il est également possible de représenter les résidus de Schoenfeld pour une variable en fonction du temps [67]. Si l'on obtient une droite horizontale, alors l'hypothèse de proportionnalité des risques est satisfaite.

2.3.5 *Alternatives en cas de risques non-proportionnels*

Si l'hypothèse des risques proportionnels n'est pas respectée, plusieurs approches peuvent être utilisées. Une première possibilité est de stratifier sur la covariable pour laquelle l'hypothèse n'est pas respectée [67]. Cependant, cette méthode ne permet pas de quantifier ou de tester l'effet des covariables sur lesquelles on stratifie, ce qui est généralement ce que l'on souhaite faire lorsque l'on s'intéresse à l'effet d'une intervention.

Une autre solution est de partitionner le temps en intervalles [67]. Dans chacun des intervalles prédéfinis, les risques sont supposés proportionnels, mais le HR peut prendre des valeurs différentes entre les intervalles. Il est également possible d'inclure une covariable dépendante du temps en introduisant une interaction entre la covariable et une fonction continue du temps [67]. La fonction dépendante du

temps peut prendre n'importe quelle forme. Le choix de la forme de la fonction peut s'appuyer sur le graphique des résidus du Schoenfeld en fonction du temps.

Une autre possibilité est d'utiliser un autre modèle comme un modèle de vie accélérée ou à risques additifs [67].

Le principe d'un modèle de vie accélérée, si on s'intéresse à la survie entre deux groupes, est que la survie du groupe intervention au temps t est égale à la survie du groupe contrôle au temps ϕt, avec ϕ le paramètre d'accélération qui est strictement positif. Si $\phi < 1$, la survie dans le groupe intervention est plus longue. La fonction de risque instantané pour un individu i est alors modélisée de la manière suivante [67] :

$$h_i(t|\mathbf{Z}_i) = \phi h_0(\phi t)$$

avec $\phi = \exp(\boldsymbol{\beta}'\mathbf{Z}_i)$. Les covariables ont un effet sur la fonction de risque instantané de base, qui n'est pas commune à tous les individus. Le rapport des fonctions de risque instantané pour deux individus n'est donc pas proportionnel au cours du temps. Comme pour les autres modèles, il est possible de définir ou non une forme pour la fonction de risque instantané de base h_0.

Le modèle à risques additifs, quant à lui, a été proposé par Aalen [68]. Il modélise la fonction de risque instantané de l'individu i de la manière suivante :

$$h_i(t|\mathbf{Z}_i) = I_{t \leq T_i} \left(h_0(t) + \boldsymbol{\beta}(t)'\boldsymbol{Z}_i(t)\right)$$

avec $\boldsymbol{\beta}(t)$ le vecteur des fonctions de régression dépendant du temps et mesurant l'effet additionnel des covariables sur le risque instantané de base, et $\boldsymbol{Z}_i(t)$ le vecteur des covariables de l'individu i pouvant dépendre du temps. D'autres modèles similaires ont par la suite été proposés, comme celui de Lin et Ying, [69] qui suppose que les covariables ont un effet constant au cours du temps, ou celui de McKeague et Sasieni [70], pour lequel certaines covariables peuvent avoir un effet variant au cours du temps et d'autres un effet constant.

Une alternative en cas de risques non proportionnels est de quantifier l'effet intervention par une différence de durée de survie moyenne restreinte à un temps d'intérêt t^*.

2.4 Différence de temps de survie moyen restreint (RMST)

La différence de durée de survie moyenne restreinte, ou *restricted mean survival time* (RMST) en anglais, à un temps d'intérêt t^*, nommé parfois temps d'horizon, est une mesure de l'effet intervention d'un critère de type survie. C'est une alternative au HR, qui reste la mesure de l'effet intervention la plus communément utilisée et qui est généralement estimée à l'aide de modèles à risques proportionnels. Cependant, lorsque l'hypothèse de proportionnalité des risques n'est pas respectée, l'utilisation de tels modèles est impossible. En outre, le HR est une mesure relative de l'effet intervention qui peut être difficile à interpréter cliniquement et son interprétation causale a été remise en question [71, 72, 73, 74, 75, 76, 77]. La différence de RMST est donc une alternative intéressante, même dans le cas où l'hypothèse de proportionnalité des risques est satisfaite.

2.4.1 *Définitions*

Le RMST est défini comme l'espérance de $X(t^*) = min(T^*, t^*)$ et correspond à l'aire sous la courbe de la fonction de survie, S, entre 0 et t^* :

$$RMST(t^*) = E[X(t^*)] = E[min(T^*, t^*)] = \int_0^{t^*} S(t)\,\mathrm{d}t$$

La variance de $X(t^*)$ est égale à [78] :

$$Var[X(t^*)] = 2\int_0^{t^*} tS(t)\,\mathrm{d}t - \left\{\int_0^{t^*} S(t)\,\mathrm{d}t\right\}^2$$

Le choix du temps d'intérêt est important, mais il n'existe pas de règle spécifique. Il est conseillé de préspécifier t^* au stade de la conception de l'étude et de choisir une valeur cliniquement pertinente en fonction de l'objectif de l'étude ainsi que de sa faisabilité [73, 78, 79].

Dans un essai en deux bras parallèles, l'effet intervention peut être quantifié par la différence des RMST au temps t^* entre les groupes intervention et contrôle :

$$\Delta RMST(t^*) = RMST_1(t^*) - RMST_0(t^*) = \int_0^{t^*} S_1(t)\,\mathrm{d}t - \int_0^{t^*} S_0(t)\,\mathrm{d}t \tag{2.4}$$

avec $RMST_j(t^*)$ et S_j, respectivement, le RMST et la fonction de survie dans le groupe j ($j = 1$ pour le groupe intervention et $j = 0$ pour le groupe contrôle). La variance de la différence de RMST correspond à la somme des variances des RMST des deux groupes [78].

Avec une différence de RMST, l'effet de l'intervention correspond au gain d'espérance de vie obtenu grâce à l'intervention sur une durée de suivi t^*. En plus d'être facilement interprétable, la différence de RMST a l'avantage de ne pas reposer sur l'hypothèse des risques proportionnels.

2.4.2 *Méthodes d'estimations*

De nombreuses méthodes ont été développées pour estimer une différence de RMST pour des données de survie indépendantes. Elles peuvent être classées en deux grandes catégories : (1) les méthodes intégrant directement une estimation de la fonction de survie, et (2) les méthodes modélisant directement le RMST en fonction de covariables grâce à des modèles de régression linéaires généralisés. La section suivante décrit de manière succincte et non exhaustive quelques-unes des méthodes existantes. Certaines seront détaillées dans une des publications en lien avec cette thèse [80].

2.4.2.1 Intégration directe de la fonction de survie

La première approche consiste à estimer la fonction de survie pour ensuite l'intégrer entre 0 et t^*. Les fonctions de survie de chaque groupe sont donc substituées par leur estimation dans l'équation 2.4 afin d'obtenir une différence de RMST. Différentes méthodes suivant cette approche sont décrites ci-dessous.

Intégration des courbes de Kaplan-Meier

Une méthode couramment utilisée est d'estimer la fonction de survie par l'estimateur de Kaplan-

Meier (2.1) [56]. La variance de l'estimation du RMST obtenue en intégrant la fonction de survie de Kaplan-Meier, $\widehat{RMST}(t^*)$, dans un groupe est estimée par [78] :

$$Var(\widehat{RMST}(t^*)) = \sum_{l=1}^{D}\left[\sum_{k=l}^{D_l}(t_{k+1}-t_k)\hat{S}(t_k)\right]^2 \frac{d_l}{N_l(N_l-d_l)} \tag{2.5}$$

avec t_1,...,t_D les D temps d'événements distincts avant t^*, $t_0 = 0$, $t_{D+1} = t^*$, d_l le nombre d'événements au temps t_l et N_l le nombre d'individus encore à risque au temps T_i.

Courbes de Kaplan-Meier ajustées

Conner *et al.* [81] ont proposé une méthode basée sur l'intégration des courbes de Kaplan-Meier permettant d'ajuster sur des covariables à l'aide d'une pondération par l'inverse du score de propension. Un poids est alloué à chaque individu, correspondant à l'inverse de la probabilité d'appartenir au groupe observé. Ces poids sont ensuite appliqués à la distribution de survie pour obtenir un estimateur de Kaplan-Meier ajusté.

Modèles à risques proportionnels

Certaines méthodes ont été proposées pour estimer la fonction de survie à l'aide de modèles à risques proportionnels, permettant d'ajuster sur des covariables. L'idée générale de ces méthodes est d'estimer les paramètres du modèle à risques proportionnels et de la fonction de risque instantané de base, puis d'en déduire la fonction de risque cumulé de laquelle la fonction de survie est obtenue et finalement intégrée pour estimer le RMST. Karrison [82] a proposé d'utiliser un modèle à risques proportionnels exponentiel par morceaux, c'est-à-dire en supposant que la fonction de risque instantané de base est constante dans des intervalles de temps prédéfinis, et stratifié sur le groupe. Zucker [83] a proposé d'utiliser un modèle de Cox stratifié par groupe, en estimant les fonctions de risque cumulés de base dans chaque groupe à l'aide de l'estimateur de Breslow [60]. Chen et Tsiatis [84] ont proposé une approche assez similaire à celle de Zucker en considérant deux modèles : le premier suppose que la fonction de risque instantané de base et les coefficients de régression sont communs aux deux groupes (modèle plus restrictif que celui Zucker), et le second qu'à la fois la fonction de risque instantané de base et les coefficients de régression peuvent être différents entre les deux groupes (modèle plus général que celui Zucker). Enfin, Zhang et Schaubel [85] ont proposé une méthode similaire au second modèle de Chen et Tsiatis, mais permettant de prendre en compte les censures dépendant des covariables à l'aide d'une pondération par l'inverse des probabilités de censure, ou *inverse probability of censoring weighting* (IPCW) en anglais. On peut noter que ces méthodes ont l'avantage que l'hypothèse de proportionnalité des risques n'a pas besoin d'être respectée pour le groupe, du fait de la stratification, mais elle doit l'être pour les covariables sur lesquelles on ajuste.

Modèle paramétrique flexible

Royston et Parmar [75, 86] ont proposé un modèle paramétrique permettant de prendre en compte de manière flexible un large nombre de distributions de la fonction de risque instantané de base. Le modèle utilisé est le suivant :

$$\ln H(t|\mathbf{Z}_i) = \ln H_0(t) + \mathbf{Z}_i'\boldsymbol{\beta} = s(\ln t) + \mathbf{Z}_i'\boldsymbol{\beta} \tag{2.6}$$

où le logarithme de la fonction de risque cumulé de base, $\ln H_0(t) = s(\ln t)$, est modélisé par une spline cubique en logarithme du temps. La fonction $s(\ln t)$ correspond à une combinaison linéaire de fonctions

de bases et de paramètres. Ainsi, cette méthode permet de spécifier des fonctions de distribution de base flexibles. Le RMST, pour les covariables $\mathbf{Z}_i$, est obtenu en prédisant le logarithme de la fonction de risque cumulé (équation (2.6)) pour un nombre de temps t suffisamment important, afin d'en déduire la fonction de survie qui est ensuite intégrée entre 0 et t^*. Royston et Parmar proposent d'estimer les erreurs standards par bootstrap ou avec la méthode delta [75].

2.4.2.2 Modèles de régression

D'autres méthodes ont été développées afin de modéliser directement le RMST en fonction des covariables à l'aide de modèles de régression linéaires généralisés.

Régression sur des pseudo-valeurs

Andersen *et al.* [87, 88] ont proposé une approche par pseudo-valeurs pour estimer une différence de RMST. Leur méthode permet de modéliser $E[min(T_i^*, t^*)]$ en fonction des covariables $\mathbf{Z}_i$ de la manière suivante :

$$E[min(T_i^*, t^*)|\mathbf{Z}_i] = g(\boldsymbol{\beta}'\mathbf{Z}_i) \tag{2.7}$$

Premièrement, une pseudo-valeur est calculée pour chaque individu i comme suit :

$$Y_i = n \times \widehat{RMST}(t^*) - (n-1) \times \widehat{RMST}^{-i}(t^*) = n \int_0^{t^*} \hat{S}(t)\,\mathrm{d}t - (n-1) \int_0^{t^*} \hat{S}^{-i}(t)\,\mathrm{d}t$$

où $\widehat{RMST}(t^*)$ est l'estimateur intégrant la fonction de survie de Kaplan-Meier pour l'ensemble de l'échantillon de taille n et $\widehat{RMST}^{-i}(t^*)$ pour l'échantillon sans l'individu i de taille $n-1$. Les pseudo-valeurs sont ensuite utilisées comme variable réponse dans le modèle (2.7). Andersen *et al.* [87, 88], suggèrent d'utiliser une fonction de lien g identité ou log. Les coefficients de régression sont estimés à l'aide de GEE et leur variance à l'aide de l'estimateur sandwich.

Autres modèles de régression

Tian *et al.* [79] ont proposé un modèle de régression qui permet de modéliser le RMST en fonction des covariables à baseline en construisant des équations d'estimation à l'aide de méthode IPCW. Wang et Schaubel [89] ont développé une méthode similaire permettant de prendre en compte les censures non-indépendantes (dépendantes du vecteur de covariables ou des temps d'événement) à l'aide de double poids inversés. Zhong et Schaubel [90] ont proposé une méthode permettant de modéliser le RMST comme une fonction continue du temps. Cette dernière approche permet d'explorer les effets des covariables dans le temps sur le RMST et de pouvoir estimer un RMST à plusieurs temps d'intérêt à l'aide d'un seul modèle.

3 Qu'en est-il de l'analyse des données de survie dans les ERC ?

Dans le cadre des ERC, les méthodes d'analyse de survie classiques, présentées dans la section précédente, ne peuvent pas être utilisées du fait de la non-indépendance des réponses des individus au sein des clusters. Des méthodes statistiques adaptées prenant en compte la corrélation intra-cluster doivent être utilisées. Cependant, il n'existe pas de recommandations sur les stratégies d'analyse optimales pour les ERC avec un critère de jugement de type survie, contrairement aux critères de jugement continu ou binaire, même si en pratique ils ne sont pas rares. En effet, dans le cadre du projet QUARTET (QUantification of clustering and Analysis of cluster Randomised Trials with time-to-EvenT outcomes), dans lequel s'inscrit cette thèse, une revue de la littérature a été menée afin d'estimer la prévalence des critères de jugement de type survie dans les ERC. Parmi 184 articles inclus rapportant des ERC publiés entre 2013 et 2018 dans six principales revues de médecine générale (*Annals of Internal Medicine*, *British Medical Journal*, *Journal of the American Medical Association*, *The Lancet*, *The New England Journal of Medicine* and *PLOS Medicine*), 17% (32/184) des ERC présentaient un critère de jugement principal ou secondaire de type survie [91]. Cette revue a également montré qu'en pratique, les stratégies utilisées n'étaient pas toujours adaptées puisque seulement 44% (14/32) des ERC avec un critère de jugement de type survie prenaient en compte la corrélation intra-cluster dans l'analyse statistique. L'utilisation de méthodes non optimales peut être en partie expliquée par le manque de littérature sur les critères de type survie dans les ERC.

Certaines méthodes d'analyse adaptées pour des données de survie corrélées existent néanmoins et peuvent être utilisées dans le cadre des ERC. Dans cette section, nous présentons certaines d'entre elles.

3.1 Méthodes pour estimer un HR dans le cadre des ERC

Les modèles à risques proportionnels ont été adaptés pour des données corrélées. Dans le cadre des ERC, ils permettent une analyse au niveau individuel, soit avec l'approche conditionnelle (*"cluster-specific"*), soit avec l'approche marginale (*"population average"*).

3.1.1 *Modèles de survie conditionnels*

L'approche conditionnelle en survie correspond aux modèles incluant un effet aléatoire pour prendre en compte la corrélation entre les réponses des membres d'un même cluster, de la même manière que les autres méthodes conditionnelles pour les ERC. Ces modèles sont nommés "modèles à fragilité". Généralement, les modèles à fragilité sont des extensions des modèles à risques proportionnels :

$$h_{ki}(t|\mathbf{Z}_{ki}) = h_0(t)u_k \exp(\boldsymbol{\beta}'\mathbf{Z}_{ki}) \tag{3.1}$$

où h_{ki} est la fonction de risque conditionnel pour le sujet i du cluster k et u_k est le terme de fragilité du cluster k. Le terme de fragilité est partagé par l'ensemble des individus d'un même cluster et il diffère d'un cluster à l'autre. Il permet d'introduire de la corrélation entre les temps de survie des individus au sein d'un cluster et de prendre en compte l'ensemble des facteurs de risque non observés communs au cluster.

Une autre formulation équivalente possible est la suivante :

$$h_{ki}(t|\mathbf{Z}_{ki}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_{ki} + w_k) \tag{3.2}$$

avec $w_k = exp(u_k)$. On fait la distinction entre u_k le terme de fragilité et w_k l'effet aléatoire.

Distribution du terme de fragilité

Ce modèle requiert la spécification de la distribution du terme de fragilité. Plusieurs choix existent mais les distributions les plus utilisées sont la distribution gamma à un paramètre de moyenne 1 et de variance θ, pour ses propriétés mathématiques intéressantes, et log-normale [66, 92].

Méthode d'estimation

Les méthodes d'estimation varient en fonction des hypothèses du modèle sur la forme de la fonction de risque instantané de base et sur la distribution du terme de fragilité. Pour un modèle semi-paramétrique de Cox, un algorithme *Expectation Maximisation* (EM) peut être utilisé pour maximiser la vraisemblance partielle lorsqu'une distribution gamma est utilisée [66, 92]. L'estimation des paramètres peut également se faire en maximisant la vraisemblance partielle pénalisée pour une distribution gamma ou log-normale [66]. Pour des modèles paramétriques, la vraisemblance marginale est maximisée. Dans le cas d'une distribution gamma, la forme analytique est disponible [66, 92]. Dans le cas d'une distribution log-normale, la vraisemblance marginale est approchée par méthode numérique, comme par exemple par quadratures de Gauss [66, 92].

3.1.2 *Modèles de survie marginaux*

L'approche marginale a été principalement développée pour le modèle de Cox [93, 94, 95, 96]. Dans le cadre des ERC, le modèle considéré est le suivant :

$$h_{ki}(t|\mathbf{Z}_{ki}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_{ki}) \tag{3.3}$$

avec $h_0(t)$ non spécifiée. Les estimations des paramètres sont obtenues à l'aide des GEE avec une matrice de corrélation indépendante et en estimant la variance à l'aide de la variance sandwich.

Des méthodes de correction de la variance ont été proposées récemment afin de pouvoir analyser un ERC avec un petit nombre de clusters à l'aide d'un modèle de Cox marginal [49].

3.2 Autres méthodes

D'autres méthodes ne reposant pas sur un modèle à risques proportionnels ont été développées . Blaha *et al.* [48] ont proposé un modèle à risques additifs à effets mixtes avec une alternative pour les ERC avec un faible nombre de clusters. Il est également possible d'utiliser un modèle de vie accéléré à effets mixtes [66].

3.3 Méthodes pour estimer un RMST dans le cadre des ERC

Au début de cette thèse, aucune méthode, à notre connaissance, n'avait été développée pour estimer une différence de RMST pour des données corrélées dans un ERC. Entre temps, en 2022, Chen *et*

al. [97] ont étendu aux ERC une méthode initialement proposée par Zhong et Schaubel [90] pour des données indépendantes, qui estime le RMST comme une fonction continue du temps d'intérêt $t^\star$ et permet de prendre en compte les censures dépendantes des covariables. A notre connaissance, seule cette méthode est disponible.

4 Objectifs de la thèse

L'objectif principal de cette thèse est de proposer des méthodes adaptées à l'analyse ERC avec un critère de jugement de type survie évaluant l'effet d'une intervention à l'aide d'une différence de temps de survie moyen restreint.

Dans un premier temps, nous avons proposé deux approches pour estimer une différence de RMST pour les ERC en étendant des méthodes déjà existantes pour des données indépendantes (l'intégration des courbes de Kaplan-Meier et la régression sur des pseudo-valeurs). Les méthodes proposées ont ensuite été évaluées et comparées à l'aide d'une étude de simulation. Ce premier travail est présenté dans le Chapitre II.

Dans un second temps, nous nous sommes concentrés sur les ERC avec un petit nombre de clusters, pour lesquels le premier travail a montré que l'erreur de type I n'était pas contrôlée avec les deux approches proposées. Nous avons comparé plusieurs approches visant à corriger le biais de l'estimateur de la variance afin de réduire l'erreur de type I obtenue avec la régression sur des pseudo-valeurs. Ce second travail est présenté dans le Chapitre III.

Chapitre II

Estimation d'une différence de RMST dans le cadre des ERC

1 Présentation du travail

1.1 Contexte

Au début de la thèse, à notre connaissance, aucune méthode développée pour estimer une différence de RMST n'était adaptée aux données corrélées d'un ERC. Notre objectif a donc été de proposer des méthodes permettant de prendre en compte de manière appropriée la ressemblance des réponses des individus d'un même cluster et qui soient facilement utilisables en pratique.

1.2 Méthodes

Nous avons étendu aux données corrélées des ERC deux approches développées initialement pour des données de survie indépendantes. La première de ces approches est celle qui consiste à estimer le RMST en intégrant les courbes de survie de Kaplan-Meier. Pour des données indépendantes, la variance correspond à la somme des variances estimées dans chaque groupe par l'estimateur (1.8) (méthode KM_{indep}). Pour des données corrélées, la corrélation intra-cluster a été prise en compte en estimant la variance par bootstrap, en rééchantillonant les clusters (méthode KM_{clust}). La seconde approche étendue est la régression sur des pseudo-valeurs. Elle consiste à calculer une pseudo-observation pour chaque individu et à les considérer comme les variables réponses d'un modèle linéaire estimé à l'aide des GEE. Pour des données indépendantes, la variance est estimé avec l'estimateur sandwich en considérant que les individus sont indépendants (méthode PV_{indep}). Elle a été étendue aux ERC en spécifiant une matrice de corrélation de travail échangeable (1.7) (méthode PV_{ECM}). Cependant, la matrice de corrélation échangeable étant généralement sujette à des problèmes de convergence lorsque qu'il y a peu de clusters, nous avons également étudié une structure de corrélation alternative en utilisant une matrice de corrélation indépendante (méthode PV_{ICM}).

Nous avons conduit une étude de simulation pour évaluer et comparer ces cinq méthodes, trois

prenant en compte la corrélation intra-cluster (KM_{clust}, PV_{ICM}, PV_{ECM}) et deux ne la prenant pas en compte (KM_{indep} et PV_{indep}). Nous avons étudié 140 scénarios, avec différents nombres totaux de clusters ($K = \{10; 20; 40; 50; 100\}$), tailles moyennes des clusters ($m = \{25; 80\}$), effets intervention ($HR = \{1; 0,8; 0,5\}$) et degrés de corrélation intra-cluster, quantifié par le tau de Kendall ($\tau = \{0,001; 0,01; 0,05; 0,1; 0,2\}$). Nous avons également évalué les méthodes en simulant des données de survie sous une hypothèse de risques proportionnels et non-proportionnels. Pour chaque scénario, 1000 jeux de données ont été simulés.

Les simulations ont été conduites sous R. Les codes pour simuler et analyser les données sont disponibles sur GitHub.

1.3 Résultats

La méthode PV_{ECM} présentait des problèmes de convergence qui étaient d'autant plus fréquents que le nombre de clusters ou le degré de corrélation intra-cluster étaient faibles. Le pourcentage de problèmes de convergence pouvait atteindre 15,4% dans le pire des scénarios avec le plus faible nombre de clusters et degré de corrélation intra-cluster. Les autres méthodes ne présentaient pas de problème de convergence.

La variance de l'effet intervention était sous-estimée par les méthodes KM_{indep} et PV_{indep}, ne prenant pas en compte la nature corrélée des données, et ce d'autant plus que le degré de corrélation intra-cluster était important. La sous-estimation de la variance menait à une erreur de type I non contrôlée.

En revanche, les méthodes étendues au ERC (KM_{clust}, PV_{ICM}, KM_{ECM}) estimaient bien la variance et l'erreur de type I était proche de 5%, s'il y avait un nombre total de clusters suffisant (≥ 50). Dans le cas où le nombre de clusters était faible, la variance de l'effet intervention était sous-estimée conduisant à une inflation de l'erreur de type I. Pour tenter de contrôler l'erreur de type I, nous avons évalué un test de permutation appliqué à la régression sur des pseudo-valeurs. Le test de permutation a donné des résultats satisfaisants quand $K < 50$ avec des erreurs de type I proches de 5%. Ce test de permutation a également été inversé afin de calculer un intervalle de confiance, qui présentait un taux de couverture proche de 95%.

Finalement, les méthodes étendues aux ERC présentaient un très faible biais dans l'estimation de l'effet intervention quel que soit le nombre de clusters. Globalement, les 3 méthodes prenant en compte la corrélation intra-cluster avaient des performances similaires. On peut cependant noter que la régression sur des pseudo-valeurs avec la matrice échangeable tendait à produire une erreur de type I légèrement plus proche de 5% qu'avec la matrice indépendante. La taille des clusters, le degré de corrélation intra-cluster et l'effet intervention ne semblaient pas affecter les résultats. Les performances des méthodes étaient similaires dans le cas des risques proportionnels et non-proportionnels.

1.4 Application

Pour illustrer les méthodes proposées, nous avons utilisé les données de l'essai OEDIPUS [10]. Cet ERC avait pour objectif d'évaluer un programme d'éducation sur le contrôle de l'asthme dédié à la minorité éthnique d'Asie du sud de deux quartiers londoniens (Newham et Tower Hamlets). L'intervention

évaluée incluait un programme de formation à destination des professionnels de santé et un programme d'éducation thérapeutique destiné aux patients. Les clusters étaient les cabinets médicaux qui ont été randomisés afin de recevoir l'intervention évaluée ou les soins habituels. Au total, 84 cabinets médicaux ont été randomisés : 44 dans le groupe intervention avec 183 patients et 40 dans le groupe contrôle avec 192 patients.

Deux critères de jugement de type survie étaient étudiés dans l'étude : 1) le temps jusqu'à une consultation en urgence (un des co-critère des jugement principaux) et 2) le temps jusqu'au premier rendez-vous de suivi de l'asthme (un des critères de jugement secondaires). Nous avons estimé une différence de RMST à t^* = 365 jours et son intervalle de confiance avec les trois méthodes (KM_{clust}, PV_{ICM}, KM_{ECM}). Pour les trois méthodes, nous avons estimé une différence non significative de 4 jours entre le groupe d'intervention et le groupe contrôle pour le temps jusqu'à une consultation en urgence. En revanche, pour les trois méthodes, le programme d'éducation au contrôle de l'asthme a contribué à raccourcir le temps jusqu'au premier rendez-vous de suivi de l'asthme, avec une diminution significative d'environ trois mois.

1.5 Conclusion

L'étude de simulation a permis de montrer que les méthodes étendues aux données corrélées des ERC estiment correctement la variance de l'effet intervention et contrôlent l'erreur de type I pour les ERC avec au moins 50 clusters. Pour des ERC avec un plus faible nombre de clusters, la régression sur des pseudo-valeurs avec un test de permutation peut être utilisée, afin d'obtenir à la fois une p-valeur et un intervalle de confiance.

Ce travail a fait l'objet d'un article publié dans *Statistical Methods in Medical Research* [80].

2 Article n°1

Check for updates

Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
1–17
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802231192960
journals.sagepub.com/home/smm
Sage

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] and Agnès Caille[1,2]

Abstract

For time-to-event outcomes, the difference in restricted mean survival time 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 restricted mean survival time estimation for independent data to clustered data in the framework of cluster randomized trials: 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 (≥ 50) under both proportional and non-proportional hazards assumption. For cluster randomized trials 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 cluster randomized trial 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 levels 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 an 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

[1]INSERM, SPHERE, U1246, Tours University, Nantes University, Tours, France
[2]INSERM, SPHERE, U1246, Tours University, Nantes University, Tours, France
[3]NSERM, SPHERE, U1246, Nantes University, Tours University, Nantes, France

Corresponding author:
Floriane Le Vilain–Abraham, Univ Tours, Nantes Université, INSERM UMR 1246, Methods in Patients-Centered Outcomes and Health Research, SPHERE, 2 boulevard tonnellé, Tours 37000, France.
Email: floriane.levilainabraham@univ-tours.fr

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 by 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 and Parmar[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 the 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^*.

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

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.

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

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 is 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 $\{X_l, \delta_l, Z_l; l = 1, \ldots, n\}$, with the intervention indicator $Z_l = l$ 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 $\hat{S}(\cdot)$:

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

where $t_1, \ldots, t_E$ 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

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

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 $(l = 1, \ldots, n)$ 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 $E[\min(T_l, t^*)|\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}}]$ depends on $\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}}$:

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

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

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

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 the 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 $\hat{\boldsymbol{\beta}}$ could be estimated by using the sandwich estimator detailed by Andersen et al.[22] and provided in Supplemental Appendix A.1. The $\Delta\widehat{RMST}(t^*)$ corresponds to $\hat{\beta}_1$ and the associated $100(1-\alpha)\%$ CI is calculated as $\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)}$, where $\widehat{Var}(\hat{\beta}_1)$ 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} is the event indicator, and Z_{kl} is 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 $100(1-\alpha)\%$ 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 $\hat{\boldsymbol{Y}}_k = (\hat{Y}_{k1},\ldots,\hat{Y}_{km_k})$ and $\boldsymbol{\mu}_k = (\mu_{k1},\ldots,\mu_{km_k})$, where $\mu_{kl} = g^{-1}(\boldsymbol{\beta}'\tilde{\boldsymbol{Z}}_{kl})$. The regression coefficients were estimated from the following GEE[33,34]:

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

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 $\boldsymbol{A}_k$ is 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 $\rho = Corr(\hat{Y}_{kl}, \hat{Y}_{ki})\ \forall k, i \neq l$.[2] The estimations of ρ and ϕ are detailed in Supplemental Appendix A.2.

The variance of the regression coefficient estimates was estimated with the following sandwich estimator[33]:

$$\Sigma = I(\hat{\boldsymbol{\beta}})^{-1} \widehat{var}(\boldsymbol{U}(\hat{\boldsymbol{\beta}})) I(\hat{\boldsymbol{\beta}})^{-1} \tag{8}$$

where

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

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

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 an 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 $100(1-\alpha)\%$ 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 Supplemental 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 two-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 $m_k (k = 1, \ldots, K)$ 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 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

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)

T_{delay} and a benefit after: $h_{kl}(t) = u_k \lambda \rho t^{\rho-1} \exp(Z_{kl}\beta_G 1\{t > T_{delay}\})$.[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 the same cluster were independent. More details about the simulation of the censoring times are available in Supplemental 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 number 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 $v = 48^2$ or $m = 25$ and variance $v = 15^2$, 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 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 (Supplemental 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 the number of clusters, size of the 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 Supplemental Appendix C.3 and provide the true value of $\Delta RMST(t^*)$ for each scenario in Supplemental 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 a test

size of 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 Supplemental Appendix B.4.

The code for the simulation study is available at 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 (Supplemental Table D1). The percentage of convergence problems was low when $K \geq 40$, between 0.1% and 2%, 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 the null hypothesis $H_0 : \beta_1 = \beta^*$, for many different values of β^*, and therefore, fitting multiple times model (9) of Supplemental 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 $m = 25$ (Supplemental 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 $\tau = 0.01$, 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 $\tau = 0.2$.

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 rates 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 $m = 25$ (Supplemental Table F1) are qualitatively similar.

All three methods lead to negligible bias in 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 −6.8% to 2.8% for the methods accounting for clustering. The relative error was close to 0, regardless of 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

Table 2. Type I error rate (%) for different combinations of number of clusters ($K = 10, 20, 40, 50, 100$) and Kendall's tau ($\tau = 0.001, 0.01, 0.05, 0.1, 0.2$) 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.

		Methods for independent data		Extended methods for clustered data			Extended methods and permutation test	
Number of clusters K	Kendall's τ	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		

bias explained the inflated type I error rates obtained in Section 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 $m = 25$ (Supplemental 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 a negligible bias in estimating the true $\Delta RMST(t^*)$ across all scenario settings. When $K \geq 50$, the relative error ranged from -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 a 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] These two parallel-group CRTs aimed at evaluating the effectiveness of an asthma-control education program dedicated to South Asian ethnic minority people in two London districts (Newham and

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

			Relative bias			Relative error			Coverage rate			
True HR	K	τ	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ICM} with 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	

(continued)

Table 3. Continued

			Relative bias			Relative error			Coverage rate			
True HR	K	τ	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ICM} with permutation
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 assumption for different combinations of a number of clusters ($K = 10, 20, 40, 50, 100$). Kendall's tau ($\tau = 0.001, 0.01, 0.05, 0.1, 0.2$) with true HR = 0.5 after 90 days and mean cluster size $m = 80$.

		Relative bias			Relative error			Coverage rate			
K	τ	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	KM_{clust}	PV_{ECM}	PV_{ICM}	PV_{ICM} with 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	

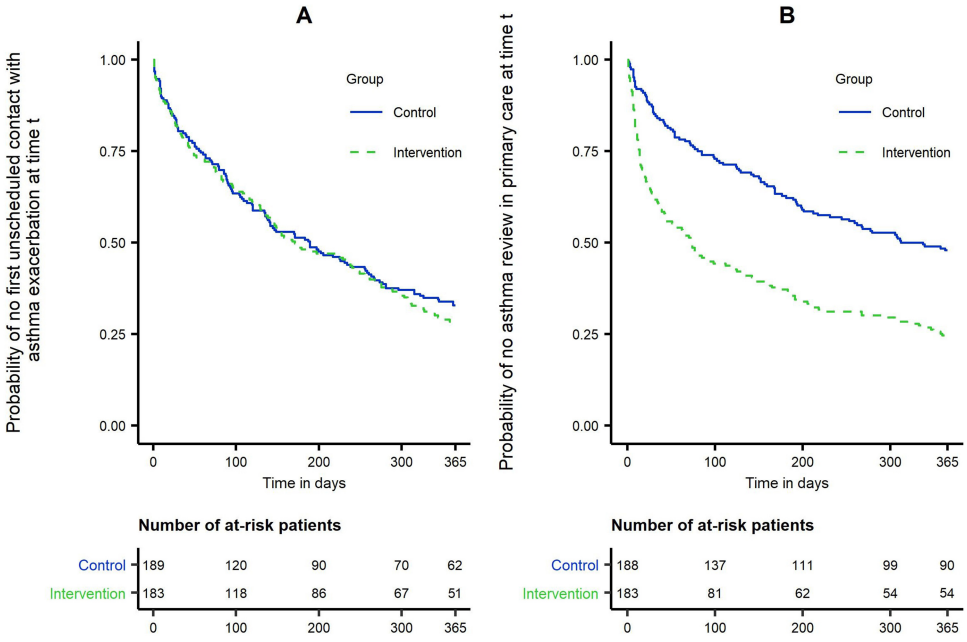


Figure 1. Kaplan-Meier curves for the outcomes of the OEDIPUS trial: (A) probability of no first unscheduled contact with asthma exacerbation at the time *t*; (B) probability of no asthma review in primary care at the time *t*.
Note: The median time to first unscheduled contact with an asthma exacerbation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

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 by 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 shown in Figure 1. The proportional hazards assumption, checked graphically by plotting the $\log(-\log(S(t)))$ curves, was met for both outcomes (Supplemental Figure G1). 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 0.85–1.41) but a significant intervention effect on time to the first asthma review in primary care (HR = 2.73, 95% CI 1.54–2.78). The asthma-control education program may significantly contribute to reducing 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 Supplemental 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 3 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.

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 ($n = 371$).

		95% CI	
Method	$\Delta RMST(365)$	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][a]
PV_{ICM}	−4.056	[−35.088; 26.977]	[−35.907; 27.788][a]
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][b]
PV_{ICM}	−89.524	[−123.717; −55.331]	[−126.179; −53.566][b]

[a]The computation times for the PV_{ECM} and PV_{ICM} are 2 min and 30 s and 1 min and 48 s respectively.
[b]The computation times for the PV_{ECM} and PV_{ICM} are 2 min and 12 s and 1 min and 48 s, respectively.

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 (≥ 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.[28] 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 of censoring 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 CI.

One might be discouraged from estimating a CI 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 2 h and 20 min, for 10 clusters, to 4 h and 40 min, for 40 clusters. Therefore, on average, the computational time to estimate a permutation-based CI for a single dataset was between 8.4 and 16.8 s. It is important to note that we benefited from a computing center, that allowed us to parallelize the calculations with 30 cores. In the OEDIPUS trial, the situation was closer to practical use since the calculations were not parallelized and we estimated the CI 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 CI is used when $K < 50$. Therefore, the computation time seems reasonable to use permutation-based CI 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 a 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 a jackknife. For independent data, Jacobsen and Torben[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 by Jacobsen and Torben,[70] Overgaard et al.,[71] and Zheng 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 the $\Delta RMST(t^*)$. Because to our knowledge, no sample-size formula based on the $\Delta RMST(t^*)$ 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 the use an 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 (≥ 50) and that we can perform a permutation test in case of a small number of clusters (< 50).

Acknowledgements

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 their 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.

Declaration of conflicting interests

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

Funding

The author(s) 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].

ORCID iDs

Floriane Le Vilain–Abraham https://orcid.org/0000-0002-5601-0246
Etienne Dantan https://orcid.org/0000-0001-7137-5051
Agnès Caille https://orcid.org/0000-0003-4715-6144

Supplemental material

Supplemental material for this article is available online.

References

1. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
2. Eldridge S, 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**: 125–137.
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, 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**: 1065–1073.
8. Therneau TM, Grambsch PM. *Modeling Survival Data*. Berlin: Springer, 2000.
9. Kalbfleich JD, 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**: 1813–1819.
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**: 2380–2385.
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**: 127–134.
15. Royston P, 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**: 2409–2421.
16. Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010; **2**: 13–15.
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**: 188–189.
19. Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* 2001; **57**: 1030–1038.
20. Royston P, 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, 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**: 335–350.
23. Andersen PK, 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* 2022; **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**: 222–233.
26. Wang X, Schaubel DE. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime Data Anal* 2018; **24**: 176–199.
27. Zhong Y, Schaubel DE. Restricted mean survival time as a function of restriction time. *Biometrics* 2022; **78**: 192–201.
28. Chen X, MO H and Li F. Clustered restricted mean survival time regression. *Biom J* 2022; **00**: 1–17.

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**: 436–453.
30. Ying Z, Wei LJ. The Kaplan-Meier estimate for dependent failure time observations. *J Multivar Anal* 1994; **50**: 17–29.
31. Efron B, 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, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
34. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
35. Sharples K, 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, 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. *Br Med J* 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, 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**: 1069–1092.
45. Li F, Lokhnygina Y, Murray DM, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Stat Med* 2016; **35**: 1565–1579.
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, 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, 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 Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2010. https://www.R-project.org/
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, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013; **32**: 4118–4134.
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**: 1292–1300.
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, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**: 126–134.
59. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; **96**: 1387–1396.
60. Fay MP, 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; **65**: 2200113.
64. Chen X, Li F. Finite-sample adjustments in variance estimators for clustered competing risks regression. *Stat Med* 2022; **41**: 2645–2664.
65. Binder N, Gerds TA and Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Anal* 2014; **20**: 303–315.

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* 2021; **9**: e701–e710.
68. Parner ET, 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, 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. *Stat Sin* 2021. https://arxiv.org/abs/2103.00605.
73. Uno H, 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.

Web Appendix for "Restricted mean survival time to estimate an intervention effect in a cluster randomized trial"

Le Vilain--Abraham F, Tavernier E, Dantan E, Desmée S, Caille A

Web Appendix A: Generalized estimating equation

A.1. Generalized estimating equation and sandwich variance estimator for the pseudo-values regression for independent data

In section 2.2.3 in the main text, we consider the following generalized linear regression model:

$$g\left(E\left[\min(T_l, t^*) \mid \widetilde{\boldsymbol{Z}}_l\right]\right) = \boldsymbol{\beta}' \widetilde{\boldsymbol{Z}}_{\boldsymbol{l}} \quad (1)$$

where t^* is the time horizon, T_l is the true time-to-event for individual l; $\widetilde{\boldsymbol{Z}}_{\boldsymbol{l}} = (1, Z_l)'$, where Z_l is the intervention variable = 1 if the individual l received the intervention and 0 if the individual is in a control group; $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, where β_0 is the intercept and β_1 is the intervention effect, and $g(.)$ a link function. Because the survival time can be censored, $\min(T_l, t^*)$ is not observed for all individuals and is replaced by the pseudo-values Y_l, defined in equation (6) in the main text, in equation (1). The regression coefficients may be estimated from the following generalized estimating equation (GEE):[1,2]

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{l=1}^{N} \boldsymbol{U}_l(\boldsymbol{\beta}) = \sum_{l=1}^{N} \frac{\partial \mu_l}{\partial \boldsymbol{\beta}} V_l^{-1}\left(\widehat{Y}_l - \mu_l\right) = 0 \quad (2)$$

where V_l is a working variance and $\mu_l = g^{-1}(\boldsymbol{\beta}' \boldsymbol{Z}_{\boldsymbol{l}})$. Andersen et al.[3] showed that the estimates of the regression coefficients are consistent and the variance for $\widehat{\boldsymbol{\beta}}$ could be estimated with the following sandwich estimator:

$$\boldsymbol{\Sigma} = \boldsymbol{I}\left(\widehat{\boldsymbol{\beta}}\right)^{-1} \widehat{var}\left(\boldsymbol{U}\left(\widehat{\boldsymbol{\beta}}\right)\right) \boldsymbol{I}\left(\widehat{\boldsymbol{\beta}}\right)^{-1} \quad (3)$$

where

$$I(\boldsymbol{\beta}) = \sum_{l=1}^{N} \left(\frac{\partial \mu_l}{\partial \boldsymbol{\beta}}\right)' V_l^{-1} \left(\frac{\partial \mu_l}{\partial \boldsymbol{\beta}}\right),$$

$$\widehat{var}\left(\boldsymbol{U}(\widehat{\boldsymbol{\beta}})\right) = \sum_{l=1}^{N} \boldsymbol{U}_l(\widehat{\boldsymbol{\beta}})\boldsymbol{U}_l(\widehat{\boldsymbol{\beta}})'$$

A.2. Generalized estimating equation and exchangeable working correlation matrix: estimation of ρ and ϕ parameters

In section 2.3.3 in the main text, we consider a sample of *K* clusters of size m_k (*k=1,…,K*) and then $n = \sum_{k=1}^{K} m_k$ individuals. The following generalized estimating equation is considered:

$$\boldsymbol{U}(\beta) = \sum_{k=1}^{K} \boldsymbol{U}_k(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}\right)' \boldsymbol{V}_k^{-1}\left(\widehat{\boldsymbol{Y}_k} - \boldsymbol{\mu}_k\right) = 0 \qquad (4)$$

where $\widehat{\boldsymbol{Y}}_k = (\hat{Y}_{k1}, \ldots, \hat{Y}_{km_k})$, $\mu_{kl} = g^{-1}(\boldsymbol{\beta}'\widetilde{\boldsymbol{Z}}_{kl})$ with $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{km_k})$ and $\boldsymbol{V}_k$ 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. If the working correlation matrix is exchangeable, then $\boldsymbol{R}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$ where, $\rho = Corr\left(\hat{Y}_{kl}, \hat{Y}_{ki}\right)\ \forall k, i \neq l$. The parameters ρ and ϕ are estimated from the Pearson residuals defined as following:

$$\hat{r}_{kl} = \frac{\hat{Y}_{kl} - \mu_{kl}}{\sqrt{Var(\mu_{kl})}} \qquad (5)$$

The dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{1}{n-p} \sum_{k=1}^{K} \sum_{l=1}^{m_k} \hat{r}_{kl} \qquad (6)$$

The parameter ρ is estimated by

$$\hat{\rho} = \frac{1}{(n^* - p)\phi} \sum_{k=1}^{K} \sum_{i \neq k} \hat{r}_{kl} \hat{r}_{ki} \qquad (7)$$

where $n^* = \sum_{k=1}^{K} m_k(m_k - 1)$.

Web Appendix B: Permutation-based confidence intervals

We consider the model (1) described in the Appendix A.1, with g(·) the identity link function, for the pseudo-values regression.

$$\mathrm{E}[\min(\mathrm{T_l}, \mathrm{t}^*) \mid \tilde{\mathbf{Z}}_\mathrm{l}] = \boldsymbol{\beta}' \tilde{\mathbf{Z}}_\mathbf{l} = \beta_0 + \beta_1 \mathrm{Z_l} \qquad (8)$$

A 100(1-α)% permutation-based confidence interval for the intervention effect β_1 can be obtained by inverting the permutation test. This requires testing the non-zero null hypothesis $\mathrm{H_0}: \beta_1 = \beta^*$, where $\beta^* \neq 0$, with permutation tests and collecting the set of values that are not rejected by these tests. To do so, Rabideau and Wang[4] proposed to rewrite the model (8) as follows:

$$\mathrm{E}[\min(\mathrm{T_l}, \mathrm{t}^*) \mid \tilde{\mathbf{Z}}_\mathrm{l}] = \beta_0 + \beta^* \mathrm{Z_l} + (\beta_1 - \beta^*)\, \mathrm{Z_l} = \beta_0 + \beta^* \mathrm{Z_l} + \delta_{\mathrm{offset}} \mathrm{Z_l} \qquad (9)$$

and to test an equivalent zero null hypothesis $\mathrm{H_0}: \delta_{\mathrm{offset}} = (\beta_1 - \beta^*) = 0$. The permutation test is carried out comparing the test statistic obtained by fitting model (9) using the observed dataset and the permuted test statistics obtained by fitting the following model under different permuted intervention allocations:

$$\mathrm{E}[\min(\mathrm{T_l}, \mathrm{t}^*) \mid \tilde{\mathbf{Z}}_\mathrm{l}] = \beta_0 + \beta^* \mathrm{Z_l^{obs}} + \delta_{\mathrm{offset}}\, \mathrm{Z_l^{perm}} \qquad (10)$$

with the fixed offset term $\beta^* \mathrm{Z_l^{obs}}$ where $\mathrm{Z_l^{obs}}$ is the intervention allocation in the observed dataset, and $\mathrm{Z_l^{perm}}$ is a permuted intervention allocation.

This method could be computationally intensive because permutation tests are performed on a large number of β^* and each permutation test is carried out on many permuted intervention allocations. Rabideau and Wang[4] proposed an efficient procedure, based on a method suggested by Garthwaite[5], to reduce the computation time. It is a sequential search, performed separately for the lower and upper bounds, which considers only a single permuted intervention allocation for each permutation test. We computed permutation-based confidence interval using this approach, which is briefly described below. For further details, see Rabideau and Wang[4] and Garthmaite[5].

B.1. Search procedure

Suppose we considered a I-step search for the upper bound of the 100(1-α)% confidence interval for β_1, denoted U. At the step i, we set $\beta^* = U_i$, the current value of U, and randomly permuted the allocation of the clusters to the control and intervention groups. We fit the model (10) to obtain the permuted test statistic T_i^{perm}. We also fit the model (9) with the observed dataset to obtain the observed test statistic T_i^{obs}. The upper bound is updated based on the observed and permuted test statistics.

$$U_{i+1} = \begin{cases} U_i - \dfrac{c(\alpha/2)}{i}, & \text{if } T_i^{perm} > T_i^{obs} \\ U_i + \dfrac{c(1-\alpha/2)}{i}, & \text{if } T_i^{perm} \leq T_i^{obs} \end{cases}$$

where c is a positive constant, called the step length constant.

Independently, the search for the lower bound L is carried out in a similar way, using

$$L_{i+1} = \begin{cases} L_i + \dfrac{c(\alpha/2)}{i}, & \text{if } T_i^{perm} < T_i^{obs} \\ L_i - \dfrac{c(1-\alpha/2)}{i}, & \text{if } T_i^{perm} \geq T_i^{obs} \end{cases}$$

The final updated value, at the step I, are used as the bounds of the CI.

B.2. Implementation

Following the recommendations of Rabideau and Wang[4] and Garthwaite[5], at the step i, we set $c = \kappa(U_i - \hat{\beta}_1)$ and $c = \kappa(\hat{\beta}_1 - L_i)$ for the upper and lower bounds, respectively, where $\kappa = 2/(z_{\alpha/2}(2\pi)^{-\frac{1}{2}}\exp(-z_{\alpha/2})/2)$ with α the significance level of the permutation test, $z_{\alpha/2}$ the $\alpha/2$-quantile of the standard normal distribution and $\hat{\beta}_1$ the estimation of β_1 obtained by fitting the model (8) with the observed dataset. The lower and upper bounds were initialized at $\hat{\beta}_1 - \frac{t_2 - t_1}{2}$ and $\hat{\beta}_1 + \frac{t_2 - t_1}{2}$, respectively, where t_1 and t_2 are the second smallest and second largest permuted estimations of δ_{offset} from a permutation test testing $H_0: \beta_1 = \hat{\beta}_1$, with $\lceil (4 - \alpha)/\alpha \rceil$ permutated allocations. To avoid an early large change, i is set to $\min(\lceil 0.3(4 - \alpha)/\alpha \rceil, 50)$ at the first step. Finally, we set I to 5000 iterations to obtain an asymptotic variance of the coverage of the 95% confidence interval of 1.3×10^{-5}.

Web Appendix C: Additional details for the simulation study

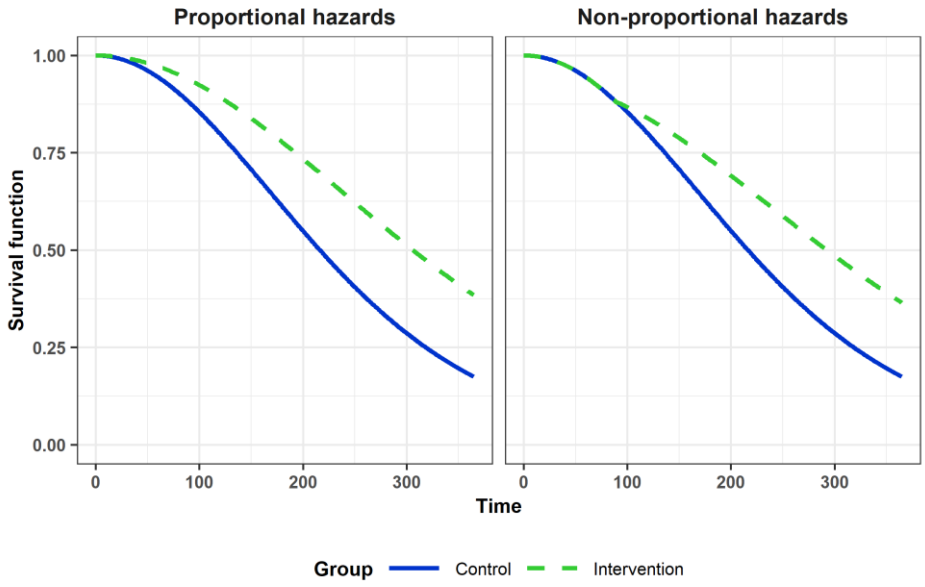
C.1. Simulation of censoring times

We simulated random censoring times independent of the event times. We assumed that the censoring times in a cluster were independent. Considering a fixed censoring rate ψ, we generated R_{kl} (for $l = 1, \ldots, m_k, k = 1, \ldots, K$) from a continuous uniform distribution on 0 to 1 and censored the observation if $R_{kl} < \psi$. When an observation was censored,

the censoring time for this individual l from the cluster k was generated with a uniform distribution on 0 to T_{kl}, the time-to-event of this individual.[6]

C.2. Simulated survival function

Figure C2: Simulated survival functions under proportional and non-proportional hazards assumptions for true HR = 0.5 and Kendall's tau τ = 0.1. The survival functions for the other scenarios are similar.



C.3. Computation of the true difference in RMST

We computed the true difference in RMST up to t^* by integrating the marginal survival function in each group between 0 and t^*.

$$\int_0^{t^*} S(t, Z = 1)dt - \int_0^{t^*} S(t, Z = 0)dt \qquad (8)$$

where $S(t, Z = j)$ is the marginal survival function for the group j (j=0 for the control group, 1 for the intervention group). The marginal survival function in each group was derived from the conditional survival function by integrating out the frailty with respect to the gamma density function.[7]

$$S(t, Z = j) = \int_0^{+\infty} S(t, Z = j|u_k) f(u_k) du_k$$

where $S(t, Z = j \mid u_k)$ is the conditional survival function for individuals in group j ($j = 0,1$) and $f(\cdot)$ is the density function of the distribution gamma $\Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$ with a mean of 1 and variance θ. An analytic form was obtained by using the Laplace transform.[7]

$$S(t, Z = j) = \mathcal{L}(H(t)) = \frac{1}{\left(1 + \frac{\lambda t^{\rho} \exp(Z\beta_G)}{1/\theta}\right)^{1/\theta}}$$

where $\mathcal{L}(.)$ is the Laplace transform and $H(.)$ the cumulative hazard function.

The integrals in (4) were approximated by using the numerical integration method of the right Riemann sum.

For non-proportional hazards assumption, we summed the true ΔRMST computed separately over the 2 periods (0 to 90 days, 90 days to t^*) with the corresponding treatment effect β.

Table C3: True difference in RMST up to 365 days for each scenario setting

Hazards assumption	True hazard ratio $exp(\beta_G)$	Kendall's tau τ	True ΔRMST up to 365
Null hypothesis	1	0.001	0
		0.01	0
		0.05	0
		0.1	0
		0.2	0
Proportional hazards	0.5	0.001	55.15
		0.01	54.78
		0.05	53.11
		0.1	51.00
		0.2	46.70
	0.8	0.001	18.72
		0.01	18.60
		0.05	18.04
		0.1	17.33
		0.2	15.87
Non-proportional hazards	0.5 (after 90 days)	0.001	42.03

	0.01	41.72
	0.05	40.33
	0.1	38.58
	0.2	35.01

C.4. Definition of the performance measures reported in the simulations

Let's denote $\Delta\widehat{RMST}(t^*)_d$ and $\hat{V}(\Delta\widehat{RMST}(t^*)_d)$ the $\Delta\widehat{RMST}(t^*)$ and its variance estimation in the simulated dataset d (d = 1,…, 1000) and $\Delta^{t^*}_{RMST}$ the true RMST difference. The performance of methods maintaining the type I error rate was assessed with the following criteria:

(1) The relative bias calculated as $\frac{\overline{\Delta RMST}(t^*)-\Delta^{t^*}_{RMST}}{\Delta^{t^*}_{RMST}} \times 100$, where $\overline{\Delta RMST}(t^*) = \frac{1}{1000}\sum_{d=1}^{1000} \Delta\widehat{RMST}(t^*)_d$,

(2) the relative error (RE) of the asymptotic standard errors (ASE), defined as $\frac{ASE-ESE}{ESE} \times 100$, where $ASE = \sqrt{\frac{1}{1000}\sum_{d=1}^{1000} \hat{V}(\Delta\widehat{RMST}(t^*)_d)}$ and the empirical standard error $ESE = \sqrt{\frac{1}{999}\sum_{d=1}^{1000}\left(\Delta\widehat{RMST}(t^*)_d - \overline{\Delta RMST}(t^*)\right)^2}$. The RE corresponds to the bias in the estimation of the true standard error, and thus the variance, of the $\Delta RMST(t^*)$. A value of the RE close to 0 reflects a good estimation of the asymptotic standard error of the intervention effect. When the RE deviates from 0, the variance is underestimated if RE is < 0 and overestimated otherwise.

(3) The empirical coverage rate of the nominal 95% confidence intervals defined as the proportion of 95% confidence intervals containing the true RMST difference.

References

1. 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.

2. Andersen PK and Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010; 19: 71-99.

3. Andersen P, Klein J and Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 2003; 90: 15-27.

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

5. Garthwaite, PH. Confidence Intervals from Randomization Tests. *Biometrics* 1996; 52: 1387–13936. Srivastava DK, George EO, Lu Z, et al. Impact of unequal censoring and insufficient follow-up on comparing survival outcomes: Applications to clinical studies. *Stat Methods Med Res* 2021; 30: 2057-2074.

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

Web Appendix D: Convergence

Table D1: Scenario settings and percentage of simulation iterations that did not converge for the pseudo-values regression with an exchangeable correlation matrix. There were no convergence issues for the scenarios not reported in the Table.

Total number of clusters	Mean cluster size	Kendall's tau τ	True hazard ratio $exp(\beta_G)$	Percentage of datasets with convergence issues
Null hypothesis				
10	80	0.001	1	14.0
10	80	0.01	1	3.0
10	80	0.05	1	0.2
20	80	0.001	1	6.4
20	80	0.01	1	0.2
40	80	0.001	1	2.0
50	25	0.001	1	1.4
50	80	0.001	1	0.6
100	25	0.001	1	0.2
Proportional hazards				
10	80	0.001	0.8	14.9
10	80	0.001	0.5	15.4
10	80	0.01	0.8	3.2
10	80	0.01	0.5	5.4
10	80	0.05	0.8	0.2
10	80	0.05	0.5	0.2
10	80	0.1	0.5	0.1
20	80	0.001	0.8	4.2
20	80	0.001	0.5	4.4
20	80	0.01	0.8	0.4
20	80	0.01	0.5	0.7
40	80	0.001	0.8	1.5
40	80	0.001	0.5	1.6
50	25	0.001	0.8	1.6
50	25	0.001	0.5	1.0
50	25	0.01	0.8	0.2
50	80	0.001	0.8	0.3
50	80	0.001	0.5	0.6
100	25	0.001	0.8	0.2
100	25	0.001	0.5	0.2
Non-proportional hazards				
10	80	0.001	0.5 (after 90 days)	14.3
10	80	0.01	0.5 (after 90 days)	4.4
10	80	0.05	0.5 (after 90 days)	0.2
20	80	0.001	0.5 (after 90 days)	5.2
20	80	0.01	0.5 (after 90 days)	0.5
40	80	0.001	0.5 (after 90 days)	1.2
50	25	0.001	0.5 (after 90 days)	1.5
50	25	0.01	0.5 (after 90 days)	0.5
50	80	0.001	0.5 (after 90 days)	0.3
100	25	0.001	0.5 (after 90 days)	0.2
100	25	0.01	0.5 (after 90 days)	0.1
100	80	0.001	0.5 (after 90 days)	0.2

Web Appendix E: Simulation results for the type I error rate

Table E1: Type I error rate (%) under the proportional hazards assumption for different combinations of number of clusters (K = 50, 100) and Kendall's tau (τ = 0.001, 0.01, 0.05, 0.1, 0.2) with mean cluster size m = 25. 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		
		KM_{indep}	PV_{indep}	KM_{clust}	PV_{ECM}	PV_{ICM}
50	0.001	**5.2**	**5.1**	**5.6**	6.9	**5.8**
	0.01	10.8	10.7	**6.2**	7.3	**6.3**
	0.05	25.3	25.3	**5.5**	**5.9**	**6.1**
	0.1	39.9	39.9	**5.4**	**6.0**	**5.8**
	0.2	51.6	51.6	7.4	7.8	7.7
100	0.001	**6.0**	**6.0**	**6.1**	**5.8**	**5.9**
	0.01	11.5	11.5	**5.3**	**5.2**	**5.6**
	0.05	27.3	27.3	**6.1**	**5.7**	**6.0**
	0.1	39.1	39.1	**6.1**	**5.6**	6.5
	0.2	50.1	50.1	6.7	**5.6**	7.1

Web Appendix F: Simulation results with mean cluster size $m = 25$

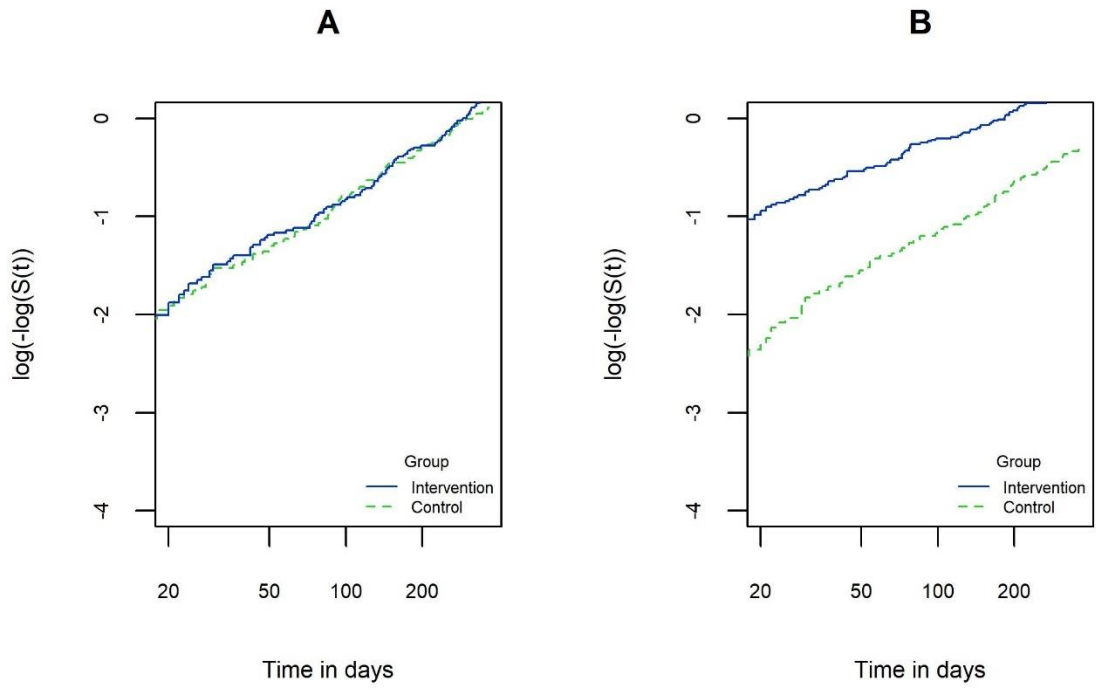
Table F1: Relative bias (%), relative error (%) and coverage rate (%) under proportional and non-proportional hazards assumptions for different combinations of number of clusters (K = 50, 100), Kendall's tau (τ = 0.001, 0.01, 0.05, 0.1, 0.2) and true HR (HR = 0.8, 0.5) with mean cluster size m = 25.

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}
0.8	50	0.001	-2.29	-2.20	-2.31	0.4	-1.4	-0.4	93.6	92.4	93.1
		0.01	2.08	2.13	2.05	-1.5	-2.9	-2.1	94.3	93.9	94.1
		0.05	-0.19	-0.53	-0.23	-5.0	-3.2	-5.1	91.3	92.2	91.2
		0.1	0.18	0.12	0.15	-4.8	-1.0	-4.7	93.0	93.5	92.5
		0.2	-4.01	-0.83	-4.05	-6.4	-3.8	-6.3	93.1	92.7	91.8
	100	0.001	1.11	1.16	1.08	-2.4	-3.3	-2.7	94.0	93.3	94.1
		0.01	-0.70	-0.68	-0.72	-5.5	-6.0	-5.6	93.4	93.4	93.2
		0.05	-0.68	-0.75	-0.70	-0.2	1.5	-0.1	94.6	94.3	94.6
		0.1	-5.19	-4.36	-5.21	0.0	0.1	0.1	93.9	93.7	93.7
		0.2	-0.36	-0.55	-0.38	-2.3	-1.8	-2.1	94.7	94.7	94.5
0.5	50	0.001	-1.27	-1.51	-1.42	0.0	-3.2	-0.2	93.7	93.3	93.7
		0.01	-0.82	-0.98	-0.94	-6.4	-6.6	-6.7	93.0	93.0	93.0
		0.05	-1.94	-1.61	-2.05	-2.9	-2.0	-3.3	94.1	94.2	93.5
		0.1	-2.48	-2.61	-2.63	-6.3	-6.1	-6.4	92.6	92.9	91.9
		0.2	-2.78	-2.89	-2.90	-6.2	-6.3	-6.5	93.0	92.9	91.8
	100	0.001	-0.45	-0.57	-0.59	-1.7	-2.4	-1.4	94.0	94.0	94.1
		0.01	-0.56	-0.67	-0.69	2.6	2.7	2.8	95.7	95.2	95.6
		0.05	-1.04	-1.10	-1.15	0.2	0.3	0.1	94.9	94.7	95.2
		0.1	-2.14	-2.16	-2.24	-2.4	-4.0	-2.6	94.2	92.8	93.9
		0.2	-3.58	-3.62	-3.67	-3.2	-2.2	-3.4	93.1	93.9	93.0
Non-proportional hazards											
0.5 (after 90 days)	50	0.001	2.58	2.59	2.47	-2.8	-4.8	-3.0	92.9	91.6	92.7
		0.01	2.64	2.74	2.55	-5.4	-5.6	-5.6	92.8	92.4	92.9
		0.05	1.69	1.67	1.59	-1.9	-1.1	-1.9	93.8	94.0	93.5
		0.1	2.09	2.57	1.99	-3.1	-2.4	-3.2	93.0	92.4	92.7
		0.2	0.72	0.91	0.62	-4.8	-3.5	-4.8	93.1	93.1	92.1
	100	0.001	2.35	2.27	2.24	-1.0	-0.4	-0.8	93.4	93.9	93.6
		0.01	3.10	2.96	3.01	-1.4	-2.3	-1.2	93.2	92.5	93.1
		0.05	1.62	1.69	1.56	-2.2	-1.7	-2.0	94.3	93.9	94.2
		0.1	1.81	1.98	1.72	-1.7	-3.5	-1.5	94.6	94.0	94.4
		0.2	1.07	0.94	0.99	-1.9	1.1	-1.8	94.4	94.8	93.6

Web Appendix G: Checking the proportional hazards assumption and R code for the illustrative example OEDIPUS

G.1. Checking the proportional hazards assumption

Figure G1: The $\log\left(-\log\left(S(t)\right)\right)$ curves for the outcomes of the OEDIPUS trial: A) time to first unscheduled contact with an asthma exacerbation; B) time to first asthma review in primary care



G.2. R code to estimate the difference in RMST

The R code used to compute the difference in RMST and its confidence interval for the 2 outcomes of the OEDIPUS trial with all the proposed methods in Table 5 is detailed in the following section. The "ci" function used to estimate the permutation-based confidence interval is available on https://github.com/Le-Vilain-Abraham/Simulation-study-with-R-software/tree/main/Proportional-hazards/1%20-%20Generation%20of%20the%20datasets%20and%20statistical%20analysis/Permutation%20test/confidence%20interval. The functions called by the "ci" function ("allocation", "ci_permuation", "initialization", "initialization_ci", "update_bound") are also available on this link. To use the "ci" function, the variable indicating

the group of intervention and the cluster ID must be denoted *arm* and *cluster*, respectively, due to the way the function was coded.

We consider a dataset, denoted *dataset*, with the following variables: the observed survival time (denoted *time)*, the event indicator = 1 if the observed time corresponds to the time between the origin and the occurrence of the survival event and 0 otherwise (denoted *indicator*), the trial group = 0 for the control group and 1 for the intervention group (denoted *group*) and the cluster ID (denoted *cluster)*. The horizon time t^* (denoted *t_star*) is fixed to 365.

G.2.1. Kaplan-Meier–based method

```
t_star = 365

#function to compute the bootstrap variance
variance_bootstrap <- function(dataset, t_star) {
     cluster.id.0 <-
     unique(as.numeric(dataset[which(dataset$group == 0),
     ]$cluster))
     cluster.id.1 <-
     unique(as.numeric(dataset[which(dataset$group == 1),
     ]$cluster))

     K0 <- length(cluster.id.0)
     K1 <- length(cluster.id.1)

     #sample K0 and K1 clusters in control and intervention
groups
     c <-c(sample(cluster.id.0, K0, replace = TRUE),
     sample(cluster.id.1, K1, replace = TRUE))

     #select the clusters in the data.frame
     data.boot <- dataset[unlist(lapply(c, function(x)
     which(dataset$cluster == x))), ]

     # Check
     while(min(max(data.boot[which(data.boot$group==1),]$time),
     max(data.boot[which(data.boot$group == 0),
]$time))<t_star){
     cluster.id.0 <-
     unique(as.numeric(dataset[which(datase$group == 0),
     ]$cluster))
     cluster.id.1 <-
     unique(as.numeric(dataset[which(dataset$group == 1),
     ]$cluster))
```

```
     K0 <- length(cluster.id.0)
     K1 <- length(cluster.id.1)

     #sample K0 and K1 clusters in control and intervention
group
     c <- c(sample(cluster.id.0, K0, replace=TRUE),
     sample(cluster.id.1, K1, replace=TRUE))


     data.boot <- dataset[unlist(lapply(c , function(x)
     which(dataset$cluster == x))), ]
     }

     #Analyze with the bootstrap dataset
     results <- rmst2(data.boot$time, data.boot$indicator,
     data.boot$group, tau= t_star)

     #Return the difference in RMST
     return(results$unadjusted.result[1,1])
     }


kaplan_meier <- rmst2(dataset$time, dataset$indicator,
dataset$group, tau = t_star)


bootstrap <- replicate(10000, variance_bootstrap(dataset,
t_star))


#Difference in RMST
kaplan_meier$unadjusted.result[1,"Est."]

#Confidence interval
quantile(bootstrap, 0.025)
quantile(bootstrap, 0.975)
```

G.2.2. Pseudo-values regression based methods

```
t_star = 365

#Computation of the pseudo-values
dataset$pv <- pseudomean(dataset$time,
                         dataset$indicator,
                         tmax=t_star)

#Exchangeable matrix
PV1 <-summary(gee(pv ~ group,
              data = dataset,
              id = cluster,
```

```
               family = gaussian,
               corstr = "exchangeable))

#Independent matrix
PV2 <- summary(gee(pv ~ group,
                data = dataset,
                id = cluster,
                family = gaussian,
                corstr = "independence"))

#Difference in RMST
PV1$coefficients["group","Estimate"]
PV2$coefficients["group","Estimate"]

#Model-based confidence interval
PV1$coefficients["group","Estimate"]-
qnorm(0.975)*sqrt(PV1$coefficients["group","Robust S.E."]^2)
PV1$coefficients["group","Estimate"]+qnorm(0.975)*sqrt(PV1$coe
fficients["group","Robust S.E."]^2)

PV2$coefficients["group","Estimate"]-
qnorm(0.975)*sqrt(PV2$coefficients["group","Robust S.E."]^2)
PV2$coefficients["group","Estimate"]+qnorm(0.975)*sqrt(PV2$coe
fficients["group","Robust S.E."]^2)

#Permutation-based confidence interval
base$arm <- base$group

ci_exc <- ci(dataset, "exchangeable")
ci_ind <- ci(dataset, "independence")
```

Chapitre III

Estimation d'une différence de RMST avec la régression sur des pseudo-valeurs pour les ERC avec un faible nombre de clusters

1 Présentation du travail

1.1 Contexte

La première étude de simulation a montré que les méthodes proposées étaient adaptées pour des ERC avec au moins 50 clusters au total. Cependant il n'est pas rare que les ERC randomisent très peu de clusters. En effet, plusieurs revues ont montré que les ERC comptent souvent moins de 40 clusters au total [39, 40, 41]. Ainsi, en pratique, l'utilisation des méthodes proposées dans le premier travail est fortement limitée.

Une solution proposée lors du premier travail, pour la méthode de régression sur des pseudo-valeurs, est d'utiliser un test de permutation. Cependant, cette méthode peut être complexe à appréhender pour des statisticiens non-experts et prendre du temps à implémenter.

La régression sur des pseudo-valeurs repose sur la méthode GEE qui est connue pour sous-estimer la variance lorsque le nombre de clusters est faible et donc produire une erreur de type I trop élevée. De nombreuses approches ont été proposées afin de réduire l'erreur de type I obtenue avec l'estimateur de la variance sandwich (1.8) dans cette situation. Une première approche est d'utiliser un autre estimateur de la variance, qui corrige le biais de l'estimateur de la variance sandwich. Plusieurs méthodes de correction de ce biais ont été proposées comme celle de Mancl et DeRouen [53], Kauermann et Carroll [54], Fay et Graubard [45] ou encore Morel et *al.* [46]. Une autre approche consiste à utiliser une distribution de Student, plutôt que la distribution normale usuelle, pour estimer la p-valeur et

l'intervalle de confiance, afin de prendre en compte la variabilité de la variance [53, 54, 55].

L'objectif du second travail était d'évaluer ces deux approches, combinées ou non, dans le cadre de la régression sur des pseudo-valeurs pour estimer une différence de RSMT pour des ERC avec un faible nombre de clusters.

1.2 Méthodes

Nous avons mené une étude de simulation afin de comparer les méthodes de correction proposées par Mancl et DeRouen (MD) [53], Kauermann et Carroll (KC) [54], Fay et Graubard (FG) [45] et Morel et *al.* (MBN) [46], combinées à une distribution normale ou de Student à $K-2$ degrés de liberté, où K est le nombre total de clusters de l'étude. Nous avons également comparé ces méthodes de correction avec le test de permutation évalué dans notre premier travail, comme les intervalles de confiance obtenus en inversant le test de permutation ont montré de bonnes performances. Pour les mêmes raisons que celles évoquées dans la première étude de simulation, les matrices de corrélation échangeable et indépendante ont été considérées. Au total, 96 scénarios ont été évalués avec différents nombres totaux de clusters ($K = \{10; 20; 30; 40\}$), tailles moyennes de clusters ($m = \{80; 450\}$), coefficients de variation de la taille des clusters ($CV = \{0,6; 0,8\}$), degrés de corrélation intra-cluster, quantifié avec un tau de Kendall, ($\tau = \{0,001; 0,01; 0,1\}$) et effets intervention (nul ou menant à une puissance de 80% avec une variance sandwich, une distribution normale et une matrice indépendante). Pour chaque scénario, 1000 jeux de données ont été simulés.

Les simulations ont été conduites sous R. Les codes pour simuler et analyser les données sont disponibles sur GitHub.

1.3 Résultats

L'étude de simulation a permis de montrer que la distribution de Student était plus performante en terme d'erreur de type I que la distribution normale, pour toutes les méthodes de correction, quel que soit le scénario. L'utilisation d'une méthode de correction permettait de réduire l'erreur de type I, comparé à l'estimateur de la variance sandwich. Les méthodes de correction doivent donc être utilisées de manière conjuguée avec une distribution de Student.

L'estimateur de Mancl et DeRouen peut être utilisé lorsque le nombre de clusters est supérieur ou égal à 30, celui de Kauerman et Carroll et Morel *et al.* pour des ERC avec au moins 20 clusters et celui de Fay et Graubart quel que soit le nombre de clusters. A part le nombre total de clusters, les paramètres de l'étude de simulation n'avait pas d'effet sur les performances des méthodes.

Il est important de noter que les tests de permutation exigent beaucoup de ressources et sont chronophages. Le temps de calcul dépend fortement du nombre de clusters et de la taille moyenne des clusters. L'évaluation d'un seul scénario avec 1000 itérations peut prendre jusqu'à plus d'une semaine à l'aide des ressources à notre disposition (accès au cluster CaScimModOT [98] permettant la parallélisation des calculs avec 30 coeurs). Ainsi, les résultats pour les scénarios avec les plus grands nombres de clusters et les plus grandes tailles de clusters ne sont pas encore disponibles au moment de la rédaction de ce travail, ce qui explique qu'ils ne sont pas présentés dans l'article en cours de

préparation (section 2).

Pour les résultats disponibles, la méthode de permutation présente une erreur de type I proche de 5% et un taux de couverture proche de 95%. On peut s'attendre à ce que les résultats soient similaires pour les scénarios manquants, qui correspondent aux nombres de clusters les plus grands, et donc à des situations moins problématiques. De plus, lors du premier travail, le test de permutation a montré de bonne performances pour des scénarios similaires.

Finalement, les méthodes de correction, comparées à la méthode de permutation, ont des performances similaires en termes d'erreur de type I. Les méthodes de correction avec une matrice indépendante ont une puissance similaire au test de permutation, mais moindre que la puissance obtenue avec les méthodes de correction avec une matrice échangeable.

1.4 Application

Nous avons illustré les méthodes avec l'essai DEMETER (Drainage des sécrétions sous-glottiques et prévention des pneumonies acquises sous ventilation mécanique en réanimation, Etude Médico-Economique multicentrique randomisée, en clusTer en crossovER). L'objectif principal de cette étude médico-économique était d'évaluer si la mise en place d'un drainage des sécrétions sous-glottiques, en plus des mesures habituelles de prévention des pneumonies acquises sous ventilation mécanique, était associée à un rapport coût-utilité à 1 an, qui rendrait sa mise en œuvre acceptable dans les services de réanimation et d'urgences. Le schéma de cette étude était un essai randomisé en clusters en cross-over. Les patients de la période d'intervention étaient intubés avec une sonde endotrachéale spécifique permettant le drainage des sécrétions sous-glottiques, tandis que ceux de la période de contrôle étaient intubés avec une sonde endotrachéale standard ne permettant pas un tel drainage.

Pour illustrer les méthodes, nous avons uniquement pris en compte les patients de la première période. L'analyse incluait donc 12 clusters avec 565 patients dans le groupe intervention et 14 clusters avec 722 patients dans le groupe contrôle. Nous avons étudié le décès à 90 jours après l'admission en réanimation. Nous avons estimé une différence de RMST à $t^* = 90$ jours. Les intervalles de confiance ont été estimés à partir des 4 estimateurs de variances corrigées et avec une distribution de Student à 24 degrés de liberté. Les matrices indépendante et échangeable ont été utilisées.

Nous avons estimé une différence de 3,55 jours entre le groupe d'intervention et le groupe de contrôle pour le décès dans les 90 premiers jours après l'admission en soins intensifs. Les 4 estimateurs de variance corrigés du biais et le test de permutation ont produit des intervalles de confiance très similaires. Les intervalles de confiance basés sur une matrice échangeable étaient légèrement plus larges que ceux obtenus à partir d'une matrice indépendante. Toutes les méthodes ont abouti à la même conclusion, à savoir que l'ajout du drainage des sécrétions sous-glottiques aux mesures habituelles de prévention de la pneumonie sous ventilation mécanique n'avait pas d'effet significatif sur le décès à 90 jours après l'admission en réanimation.

1.5 Conclusion

L'étude de simulation a montré que la méthode de correction de Fay et Graubard associée a une distribution de Student à $K - 2$ degrés de liberté pouvait être utilisée pour corriger l'erreur de type I

trop élevée pour un ERC avec un faible nombre de clusters. Une matrice échangeable devrait être utilisée de préférence, quand cela est possible, afin de bénéficier d'une puissance plus importante. Si cela n'est pas possible, une matrice indépendante peut être utilisée.

L'écriture d'un article sur ce travail est actuellement en cours de préparation.

2 Article n°2 (en préparation)

Restricted mean survival time in cluster randomized trials with a small number of clusters: Improving variance estimation of the pseudo-values regression

Floriane Le Vilain-Abraham[1], Solène Desmée[1], Jennifer A Thompson[2], Elsa Tavernier[1,4], Etienne Dantan[3], Agnès Caille[1,4]

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

2 Department of Infectious Disease, London School of Hygiene & Tropical Medicine, London, UK.

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

4 INSERM CIC1415, CHRU de Tours, Tours, France

Abstract

In randomized clinical trials with a time-to-event outcome, the intervention effect could be quantified by a difference in restricted mean survival time (ΔRMST) between the intervention and control groups up to time t^*, defined as the expected survival duration gain due to intervention over t^*. In cluster randomized trials (CRTs), social units are randomized to intervention or control groups, introducing a correlation between the survival times of the subjects of a same cluster that must be taken into account in the statistical analysis. In a previous work, we proposed the use of pseudo-values regression for estimating the ΔRMST in CRTs. It consists in computing pseudo-values for each individual and considering them as the dependent variable of a linear model fitted by generalized estimating equations (GEE). From a simulation study, we showed this method requires at least 50 clusters to control the type I error rate but can be corrected using a permutation test. Here, we aimed to propose alternative methods for small sample CRTs. We evaluated the performance of four bias-corrections of the sandwich variance estimator of the GEE developed by Mancl and DeRouen, Kauermann and Carroll, Fay and Graubard and Morel et al. with either a Student or a normal distribution of the Wald test statistic through a simulation study. We also compared these methods to the permutation test. Based on the simulation results, we recommend the use of

the Fay and Graubard estimator with a Student distribution and, if possible, an exchangeable correlation matrix for the GEE regression model. Finally, we provided an illustrative analysis, using each of the methods considered, in the DEMETER trial evaluating the use of a specific endotracheal tube allowing the subglottic secretion drainage for the prevention of ventilator-associated pneumonia.

Keywords

Cluster randomized trial; time-to-event outcome; restricted mean survival time; pseudo-values; small samples

1. Introduction

In randomised clinical trials with a time-to-event outcome, the intervention effect is conventionally quantified by the hazard ratio (HR), which is generally estimated using a Cox proportional hazards model[1]. Despite its common use in survival analysis, the clinical interpretation of the HR can be challenging, particularly when the proportional hazards assumption is not met[2-6]. Moreover, its causal interpretation is not straightforward[7,8]. The use of the restricted mean survival time (RMST), which does not rely on the proportional hazards assumption, has been recommended as an alternative measure of the intervention effect[3,9]. The RMST is defined as the average survival from time 0 to a specific time horizon t* and corresponds to the area under the survival function up to t^* [10]. The intervention effect summarized by a difference in RMST ($\Delta RMST(t^*)$) between the intervention and control groups is easily interpretable as the expected survival gain until t^* due to intervention.

In cluster randomized trials (CRTs), entire social groups are randomly assigned to intervention and control groups, rather that individuals themselves[11]. In such study design, the individuals within the same cluster are more likely to have similar survival times than individuals from distinct clusters. For valid inference, statistical analysis must take into account this clustering[12]. In a previous work[13], we showed that $\Delta RMST(t^*)$ can be estimated in CRTs using two distinct methods: one based on the direct integration of Kaplan-Meier curves[14] and one based on pseudo-values regression[15,16]. The later consists in computing pseudo-values for each individual and considering them as the dependent variable of a linear model fitted by generalized estimating equations (GEE). We recommended the pseudo-values regression based-method over the Kaplan-Meier based-method because it is already implemented in standard software, such as R[17] or SAS[17]and because it offers the possibility to adjust for covariates easily.

For CRTs with less than 50 clusters, we observed an inflated type I error rate. For the pseudo-values regression method, it is related to the underestimation of the variance when the number of clusters is limited[13,18-25]. Since several reviews suggested that CRTs frequently involve the randomization of fewer than 40 clusters, this is a severe practical limitation[25-27]. As a solution, we implemented a permutation test for the pseudo-values regression which corrected the type I error. We also inverting the permutation test, using an efficient approach to reduce the computation time[28,29], to provide confidence intervals. Although the permutation test showed good performance for CRTs with a limited number of clusters, it could be complex for a non-expert statistician to apprehend and time-consuming to implement.

Meanwhile, several methods have been proposed to correct the small-sample bias of the sandwich variance estimator of the GEE for binary and continuous outcomes such as those of Mancl and DeRouen[22], Kauermann and Carroll[23], Fay and Graubard[24] and Morel, Bokossa and Neerchal[30]. The performance of these methods has been compared for binary[19,20,31] and continuous[19,31] outcomes in CRTs, and more recently for the analysis of time-to-event outcomes with a HR[32-34]. In addition to the bias-correction methods, the use of a Student distribution instead of a normal distribution of the GEE Wald test statistic for testing the intervention effect and constructing the confidence interval, has been suggested in order to account for the variability in the standard error estimator[21-23]. Previous simulation studies showed that combining a bias-corrected variance estimator with a Student distribution can improve the performance of the GEE Wald test for CRTs with a limited number of clusters and non-censored outcomes[19-22]. However, to the best of our knowledge, the performance of these approaches has not been investigated for pseudo-values regression for $\Delta RMST(t^*)$ for CRTs with a small number of clusters.

We aim to evaluate and compare the performance of different bias-corrected sandwich estimators, combined with a normal or a Student distribution, in pseudo-values regression for $\Delta RMST(t^*)$ in CRTs with a limited number of clusters. The structure of the article is as follows. In section 2, the pseudo-values regression method extended to clustered survival data in our previous work is briefly described, followed by the different investigated approaches to correct the inflated type I error rate observed with a small number of clusters. In section 3, we compare and evaluate the small-sample bias correction approaches in a simulation study. In section 4, we illustrate the proposed approaches with the DEMETER trial. Section 5 provides a discussion.

2. Methods

2.1. Notation and definitions

To make inference, we consider K independent clusters, where the cluster k is of size m_k (for $k = 1,\ldots, K$). The total number of individuals is $n = \sum_{k=1}^{K} m_k$. Let T_{kl} and C_{kl} be the true time-to-event and the right-censoring time, respectively, of the individual l from cluster k. The observed time is $X_{kl} = \min(T_{kl}, C_{kl})$, i.e. the time between the origin and the occurrence of the survival event or censoring. The event indicator is defined as $\delta_{kl} = \mathbb{1}\{T_{kl} \leq C_{kl}\}$. We denote the intervention indicator Z_{kl}, which is equal to 1 if the individual l from cluster k is in

the intervention group and 0 if the individual is in the control group. Therefore, we consider the observed data $\{X_{kl}, \delta_{kl}, Z_{kl}; k = 1, \dots, K, l = 1, \dots, m_k, \}$. We assume that the event and censoring times are independent. The time horizon of interest is denoted t^*.

2.2. Pseudo-values regression for clustered data

In a previous work[13], we extended to clustered survival data the pseudo-values regression approach proposed by Andersen et al.[15,16] to estimate a $\Delta RMST(t^*)$. We briefly described the method in this section.

We are interested in modelling the RMST up to t^* given the intervention indicator, that is $E\left[\min(T_{kl}, t^*) \mid \widetilde{\boldsymbol{Z}}_{kl}\right]$ with $\widetilde{\boldsymbol{Z}}_{kl} = (1, Z_{kl})'$. Using a linear regression model, we can write:

$$E\left[\min(T_{kl}, t^*) \mid \widetilde{\boldsymbol{Z}}_{kl}\right] = \boldsymbol{\beta}' \widetilde{\boldsymbol{Z}}_{kl} \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ with β_0 the intercept coefficient and β_1 the intervention effect coefficient. To estimate the coefficients, a pseudo-value is first computed for each individual. Let $\widehat{RMST}(t^*)$ be the estimator of the RMST which integrates the Kaplan-Meier estimator $\hat{S}(\cdot)$ between 0 and the time of interest t^*[14].

$$\widehat{RMST}(t^*) = \int_0^{t^*} \hat{S}(t)\, dt$$

The pseudo-value for individual *l* in cluster *k* is defined as:

$$\widehat{Y_{kl}} = n \times \widehat{RMST}(t^*) - (n-1)\, \widehat{RMST}^{-kl}(t^*)$$

where $\widehat{RMST}(t^*)$ and $\widehat{RMST}^{-kl}(t^*)$ are the Kaplan-Meier-based estimators of the RMST for the whole dataset of size n and for the dataset of size $n-1$ that excludes the individual l from cluster k, respectively. Then, the pseudo-value $\widehat{Y_{kl}}$ is used as the response in the linear regression model (1). The regression coefficients are estimated using the following generalized estimating equations (GEE):

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

where $\widehat{\boldsymbol{Y}}_k = (\hat{Y}_{k1}, \dots, \hat{Y}_{km_k})$; $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{km_k})$ with $\mu_{kl} = \boldsymbol{\beta}' \widetilde{\boldsymbol{Z}}_{kl}$; $\boldsymbol{D}_k = \left(\frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}}\right)$; and $\boldsymbol{V}_k$ is a $m_k \times m_k$ working covariance matrix for cluster k. The following standard sandwich estimator is used to estimate the variance of the regression coefficient estimates:

$$\mathbf{V} = \boldsymbol{\Omega}^{-1} \left(\sum_{k=1}^{K} \boldsymbol{D}_k' \boldsymbol{V}_k^{-1} \boldsymbol{r}_k \boldsymbol{r}_k' \boldsymbol{V}_k^{-1} \boldsymbol{D}_k \right) \boldsymbol{\Omega}^{-1}$$

where $\boldsymbol{\Omega} = \sum_{k=1}^{K} \boldsymbol{D}_k' \boldsymbol{V}_k^{-1} \boldsymbol{D}_k$ and $\boldsymbol{r}_k = \left(\widehat{\boldsymbol{Y}}_k - \widehat{\boldsymbol{\mu}}_k\right)$. The $\Delta RMST(t^*)$ and its associated 100(1-α)% confidence interval are estimated by $\widehat{\beta_1}$ and $\widehat{\beta_1} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\widehat{\beta_1})}$, respectively, where $\widehat{Var}(\widehat{\beta_1})$ corresponds to the robust sandwich variance estimate of $\widehat{\beta_1}$.

2.2.1. Working correlation matrix

The working covariance matrix 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 diagonals elements $Var(\mu_{kl})$. An exchangeable working correlation matrix is commonly used in CRTs[18]. It assumes the correlation between any two pseudo-values within a given cluster is the same and identical 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, $\rho = Corr\left(\hat{Y}_{kl}, \hat{Y}_{ki}\right) \forall k, i \neq l$[35,36]. However, in our previous simulation study[13], we observed non-convergence problems with an exchangeable correlation matrix when the number of clusters is limited and clustering is low. Since the GEE estimates are robust to misspecification of the correlation matrix[36], we also investigate an alternative independent working correlation structure, i.e. identity matrix, which has previously shown good statistical performance for pseudo-values regression in CRTs[13].

2.3. Small-sample bias corrections for the sandwich estimator

Although the sandwich variance estimator **V** is consistent[36], we observed in our previous simulation study that it tends to underestimate the variance for pseudo-values regression for CRTs when the number of clusters is limited[13]. This negative bias led to undercoverage of the confidence intervals and inflated type I error. Other studies, conducted in different contexts than pseudo-values regression, have reported similar results for GEE with correlated binary and continuous outcomes[19-25]. Several methods have been proposed to correct the small-sample bias of the sandwich variance estimator **V**. In the following section, we described four bias-corrected variance estimators.

2.3.1. Mancl and DeRouen (MD) estimator

Mancl and DeRouen proposed the following estimator[22]:

$$\boldsymbol{V}_{MD} = \boldsymbol{\Omega}^{-1}\left(\sum_{k=1}^{K} \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1}\boldsymbol{r}_k\boldsymbol{r}_k'(\boldsymbol{I}_k - \boldsymbol{H}_k')^{-1}\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\right)\boldsymbol{\Omega}^{-1}$$

where $\boldsymbol{I}_k$ is an $m_k \times m_k$ identity matrix and $\boldsymbol{H}_k$ is the leverage matrix of the cluster k defined as $\boldsymbol{H}_k = \boldsymbol{D}_k\boldsymbol{\Omega}^{-1}\boldsymbol{D}_k'\boldsymbol{V}_k^{-1}$. Because the elements of the $\boldsymbol{H}_k$ are between 0 and 1, $\boldsymbol{V}_{MD}$ is expected to inflate **V.**

2.3.2. Kauermann and Carroll (KC) estimator

Kauermann and Carroll suggested the following estimator[23]:

$$\boldsymbol{V}_{KC} = \boldsymbol{\Omega}^{-1}\left(\sum_{k=1}^{K} \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-\frac{1}{2}}\boldsymbol{r}_k\boldsymbol{r}_k'(\boldsymbol{I}_k - \boldsymbol{H}_k')^{-\frac{1}{2}}\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\right)\boldsymbol{\Omega}^{-1}$$

Compared to the MD estimator, the KC estimator use the correction $(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1/2}$ instead of $(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1}$. Therefore, the KC estimator inflates the standard robust sandwich estimator $\mathbf{V}$ less than the MD estimator.

Since the calculation of $(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1/2}$ tends to be unstable compared to $(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1}$, we approximate the KC-estimator, as suggested by Gallis et al.[37], by:

$$\boldsymbol{V}_{KC} = \boldsymbol{\Omega}^{-1}\left(\frac{1}{2}\left(\sum_{k=1}^{K} \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}(\boldsymbol{I}_k - \boldsymbol{H}_k)^{-1}\boldsymbol{r}_k\boldsymbol{r}_k'\boldsymbol{V}_k^{-1}\boldsymbol{D}_k + \sum_{k=1}^{K} \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}\boldsymbol{r}_k\boldsymbol{r}_k'(\boldsymbol{I}_k - \boldsymbol{H}_k')^{-1}\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\right)\right)\boldsymbol{\Omega}^{-1} \quad (2)$$

2.3.3. Fay and Graubard (FG) estimator

Fay and Graubard proposed the following variance estimator[24]:

$$\boldsymbol{V}_{FG} = \boldsymbol{\Omega}^{-1}\left(\sum_{k=1}^{K} \boldsymbol{G}_k\boldsymbol{D}_k'\boldsymbol{V}_k^{-1}\boldsymbol{r}_k\boldsymbol{r}_k'\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\boldsymbol{G}_k\right)\boldsymbol{\Omega}^{-1}$$

where $\boldsymbol{G}_k$ is a diagonal matrix, whose jth diagonal element is $\left\{1 - \min\left(b, \{\boldsymbol{Q}\}_{jj}\right)\right\}^{-\frac{1}{2}}$ with $\boldsymbol{Q} = \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\boldsymbol{\Omega}^{-1}$ and $b < 1$ a user-specified constant[24]. The bound parameter b is used to avoid extreme adjustments when the jjth element of $\boldsymbol{Q}$ is very close to 1[24]. Fay and Graubard recommend to set arbitrarily $b = 0.75$, value that is rarely reached[24].

2.3.4. Morel, Bossoka and Neerchal (MBN) estimator

Morel, Bossoka and Neerchal suggested the following correction[30]:

$$\mathbf{V}_{\mathrm{MBN}} = \boldsymbol{\Omega}^{-1}\left(\sum_{k=1}^{K} \boldsymbol{D}_k'\boldsymbol{V}_k^{-1}(c\boldsymbol{r}_k\boldsymbol{r}_k' + \hat{\kappa}_n\hat{\phi}\boldsymbol{V}_k)\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\right)\boldsymbol{\Omega}^{-1}$$

where $c = \frac{n-1}{n-q}\frac{K}{K-1}$, n is the number of individuals, q is the dimension of the vector of parameters $\boldsymbol{\beta}$, $\hat{\phi} = \max(w, trace\{\boldsymbol{\Omega}^{-1}\sum_{k=1}^{K}\boldsymbol{D}_k'\boldsymbol{V}_k^{-1}\boldsymbol{r}_k\boldsymbol{r}_k'\boldsymbol{V}_k^{-1}\boldsymbol{D}_k\}/q)$ with $0 \leq w \leq 1$ and $\hat{\kappa}_n = \begin{cases} \frac{q}{K-q} & \text{if } K > (a+1)q \\ \frac{1}{a} & \text{otherwise} \end{cases}$. Within this paper, we followed the recommendations of Morel et al.[30] and set $a = 2$ and $w = 1$.

2.4. Distribution of the Wald test statistic

The use of a Student distribution instead of a normal distribution of the Wald statistic has been proposed to take into account the variability of the sandwich estimator[21-23]. Previous works showed that using a Student distribution can improve the performance of the GEE Wald test for CRTs with a limited number of clusters[19-22]. Several propositions have been made to determine the degree of freedom such as Satterthwaite-type degree of freedom[21,24,38] or approximation based on the number of clusters[20,22,31]. For ease of use, we approximated the degree of freedom by $K - q$, which has shown good performance in previous simulation studies[20,22,31].

2.5. Permutation test

Another alternative to reduce the inflated type I error rate for CRTs with a small number of clusters is the permutation test, that showed good performance in our previous simulation study.[13] More specifically, we inverted the permutation test to obtain confidence intervals. It requires to determine the batch of values β^* for which the null hypothesis $H_0: \beta_1 = \beta^*$ is not rejected. To obtain a reasonable computation time, we implemented an efficient method, described in Appendix A.

2.6. Implementation

The pseudo-values were computed with the *pseudomean* function from the R package "pseudo"[17]. We fitted the generalized linear regression model and estimated the standard variance $\boldsymbol{V}$ with the *gee* function from the R package "gee"[17]. The MD estimators were estimated with the functions *GEE.var.md* from the R package "geesmv"[20]. To compute the FG and the MBN estimators, we used the *GEE.var.fg* and *GEE.var.mbn* functions,

respectively, from the R package "geesmv", that we slightly modified to be used for CRTs with variable cluster sizes. To approximate the KC estimator by the formula (2), we used a R function based on a function from Thompson et al.[39] These R functions for the FG, MBN and KC estimators are available on https://github.com/Le-Vilain-Abraham/RMST-estimation-with-pseudo-values-regression-in-small-sample-CRTs.

3. Simulation study

3.1. Simulation study design

3.1.1. Data generation

We simulated two-arm parallel CRTs with a total of K clusters and a 1:1 randomisation ratio. The cluster sizes m_k (for $k = 1, \dots, K$) were variable across clusters and drawn from a negative binomial distribution with mean m and coefficient of variation $CV = \sqrt{v}/m$, with v the variance of the distribution.

We simulated clustered time-to-event data, measured in days, using the following Weibull gamma frailty regression model[40]:

$$h_{kl}(t) = \mathrm{u_k}\lambda\theta t^{\theta-1}\exp(\beta_G Z_{kl}) \quad (3)$$

where $h_{kl}(t)$ is the hazard function for individual l from cluster k at time t; $\mathrm{u_k}$ is the frailty term of the cluster k drawn from a gamma distribution $\Gamma(\frac{1}{\theta}, \frac{1}{\theta})$ with a mean of one and variance θ; λ and θ are, respectively, the scale and shape parameters of the Weibull distribution; and β_G is the intervention effect. We defined the true HR as $\exp(\beta_G)$. We quantified the degree of clustering by the Kendall's tau, $\tau = \frac{\theta}{\theta+2}$.

The time-to-events were generated from the inverse of the cumulative incidence function of the Weibull regression model (3). We simulated random censoring times independently of the event times, assuming that the censoring times within a cluster were independent. We detailed the censoring time generation mechanism in Appendix B.1.

3.1.2. Scenarios

We simulated CRTs with a total number of clusters K = {10, 20, 30, 40}, a mean cluster size m = {80, 450}[41] and a coefficient of variation $CV = \{0.6, 0.8\}$[42]. The considered Kendall's tau were $\tau = \{0.001, 0.01, 0.1\}$. To achieve a survival rate of about 0.2 in the control group at the end of the 365-days follow-up period, the parameters (ρ, λ) of the Weibull distribution were fixed to (2, 0.000016). We simulated scenarios without and with intervention effect. To assess the null hypothesis (i.e., absence of intervention effect), we set $\beta_G = 0$ ($HR = 1$). For

the scenarios with an intervention effect ($HR \neq 1$), we selected β_G to achieve 80% power for each combination of parameters, with the pseudo-value regression with the standard sandwich estimators, an independent correlation matrix and the usual normal distribution. Since no sample-size formula is available for $\Delta RMST(t^*)$ in CRTs, we used simulation. More details of the procedure are available in Appendix B.2. The censoring rate ψ, reflecting the loss to follow-up, was fixed at 20%.

In total, we evaluated 96 scenarios, with different total number of clusters, mean cluster size, coefficient of variation of the cluster size, degree of clustering and intervention effect (Table 1). For each scenario, 1000 datasets were generated.

Table 1: Summary of the settings of the 96 scenarios of the simulation study

Parameters	***Number of values***	***Values***
Total number of clusters (K)	4	10, 20, 30, 40
Mean cluster size (m)	2	80, 450
Coefficient of variation of the cluster size (CV)	2	0.6, 0.8
Kendall's tau (τ)	3	0.001, 0.01, 0.1
True HR ($exp(\beta_G)$)	2	1 (no intervention effect) or HR between 0.40 and 0.93 (intervention effect)

3.1.3. Analysis methods

We set the time horizon t^* to 365 days and the significance level α to 5%. We analysed each simulated dataset with the five following variance estimators : the standard sandwich estimator $\mathbf{V}$, the MD estimator $\boldsymbol{V}_{MD}$, the KC estimator $\boldsymbol{V}_{KC}$, the FG estimator $\boldsymbol{V}_{FG}$ and the MBN estimator $\boldsymbol{V}_{MBN}$. For the FG estimator we set the bound parameter b to 0.75 and for the MBN estimator $q = 2$, $a = 2$ and $w = 1$, as recommended by the respective authors[24,30]. Both exchangeable and independent working correlation matrices were used. We combined the five variance estimators with the normal and Student distributions of K-2 degrees of freedom. In addition, we studied the permutation-based confidence intervals, which previously showed good performance for CRTs with a limited number of clusters[13], to compare its performance with the proposed approaches. We only provided permutation-based confidence intervals for the pseudo-values regression with an independent matrix, because of the important non-convergence problems encountered with the exchangeable correlation matrix when the number of clusters is small[13].

3.1.4. Performance measures

The percentage of simulation iterations that did not converge was reported for each scenario and each analysis method. Then, the following performance measures were computed based on the analyses that converged: the type I error, the relative error of the asymptotic standard error, the coverage rate of the nominal 95% confidence intervals and the power. We detailed the performance measures in Appendix B.3.2.

The code of the simulation is available on https://github.com/Le-Vilain-Abraham/RMST-estimation-with-pseudo-values-regression-in-small-sample-CRTs.

3.2. Results

3.2.1. Convergence

Only the pseudo-values regression with an exchangeable correlation matrix had convergence problems (Web Table C.1). Overall, the percentage of non-convergence increased with decreasing number of clusters, mean cluster size and degree of clustering and with increasing coefficient of variation of cluster size. The number of non-convergences more than doubled when the mean cluster size decreased from 450 to 80 and increased by around 40% when the coefficient of variation of the cluster size increased from 0.6 to 0.8. The percentage of non-convergence was low for $K \geq 30$ clusters, not exceeding 5.2%, but increased strongly for smaller numbers of clusters, reaching 21.5% in the worst scenario. Finally, the Kendall's tau had a large effect on non-convergence, as the percentage of non-convergence was at least multiplicated by 3 when τ was divided by 10. Of note, convergence problems for $\tau = 0.1$ only occurred for a very small number of cluster ($K = 10$) and a small mean cluster size ($m = 80$).

3.2.2. Normal distribution versus Student distribution

Figure 1 compares the type I error rates between the normal and the Student distributions for the five variance estimators with both independent and exchangeable correlation matrices for all scenarios.

The Student distribution clearly outperformed the normal distribution for all the variance estimators. Indeed, the type I error rates with all methods were not appropriate with a normal distribution, except for $K = 40$ or the MD estimator when $K \geq 20$.

Although the Student distribution reduced the inflation of the type I error rate compared to the normal distribution, the standard sandwich estimator $\boldsymbol{V}$ produced uncontrolled type I errors,

especially for the very small numbers of clusters. Compared to the standard sandwich estimator, the 4 small-sample bias corrections led to smaller type I error rate.

These findings suggested that the usual normal distribution should be avoided when the total number of clusters is limited, even when a small-sample bias correction is used. Therefore, thereafter, we excluded the normal distribution and the standard sandwich variance estimator.

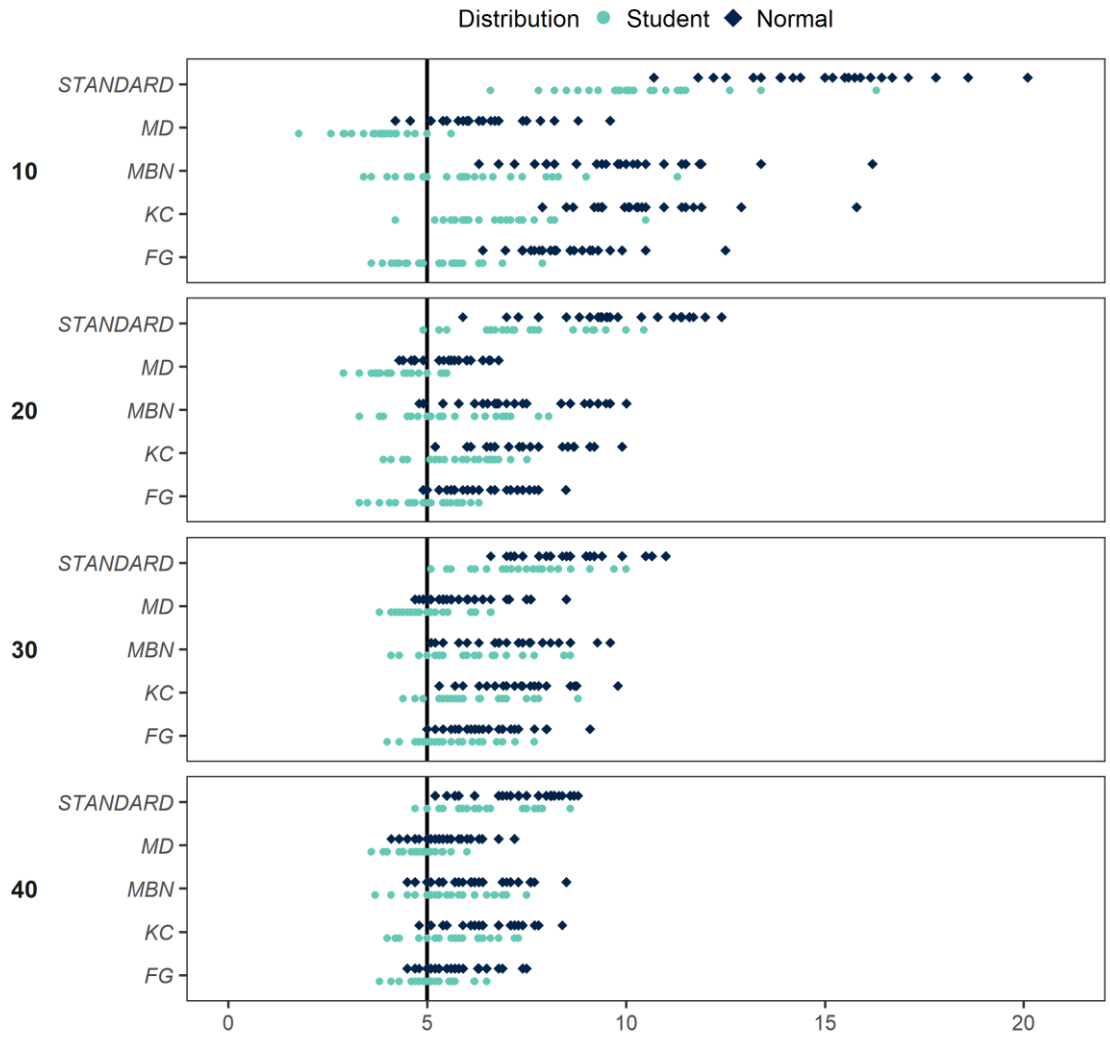


Figure 1: Type I error rate of the standard variance and the four correction estimators by number of clusters (in rows) according to the distribution of the Wald test statistic (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios without intervention effect are displayed.

3.2.3. Performance measures in absence of intervention effect

The type I error rate, the relative error and the coverage rate under the null hypothesis are compared between the independent and exchangeable correlation matrices for the 4 correction methods with a Student distribution and the permutation test in Figure 2.

Overall, the small-sample bias corrections controlled the type I error in most of the scenarios. The MD estimator tended to produce conservative type I error, especially for a very small

number of clusters and/or with an exchangeable correlation matrix. For the KC, FG and MBN estimators, the exchangeable correlation matrix led to smaller type I error rate compared to the independent matrix. The type I error rates of the KC and the MBN estimators were occasionally increased with an independent correlation matrix when $K = 10$, reaching up to 10.5% and 11.3%, respectively.

The relative error with the MD estimator exceeded 10% with both correlation matrices in most of scenarios when $K = 10$. Therefore, for a very limited number of clusters, the MD estimator overestimated the variance which explains the conservative type I error rate and coverage rates. For $K = 20$, results were similar, but for at least 30 clusters, the relative error of the MD estimator was lower than 10% and the coverage rate was close to 95%.

Overall, the three other correction methods, KC, FG and MBN, had small relative errors and coverage rates close to the nominal rate. Of note, when $K = 10$, the independent matrix led once to relative errors lower than -10% with the MBN and KC estimators, resulting in coverage rates lower to 90%.

When $K = 10$, the type I error rate of the MD estimator seems to increase with increasing Kendall's tau (Web Figures D3). Otherwise, there was no apparent pattern of the type I error, the relative error and the coverage rate according to Kendall's tau, mean cluster size and coefficient of variation for the 4 corrections (Web Figures D1, D2 and D3).

The permutation-based confidence intervals produced an appropriate type I error rate and coverage rate. The values of K, m, CV and τ did not seem to affect the performance measures of the permutation test (Web Figures D1, D2 and D3).

3.2.4. Performance measures in presence of intervention effect

Figure 3 compares the power, the relative error and the coverage rate under the scenarios with an intervention effect between independent and exchangeable correlation matrices for the 4 variance estimators combined with a Student distribution and the permutation test for all scenarios.

The power of the 4 correction methods was higher with an exchangeable matrix than with an independent matrix when $K \geq 20$ whereas for $K = 10$, both matrices correlation gave similar power. Of note, the variability in the power was more important for the exchangeable matrix than for the independent matrix, for all values of the total number of clusters. When $K = 10$, the power tended to be higher for the KC, FG and MBN estimators than for the MD estimator,

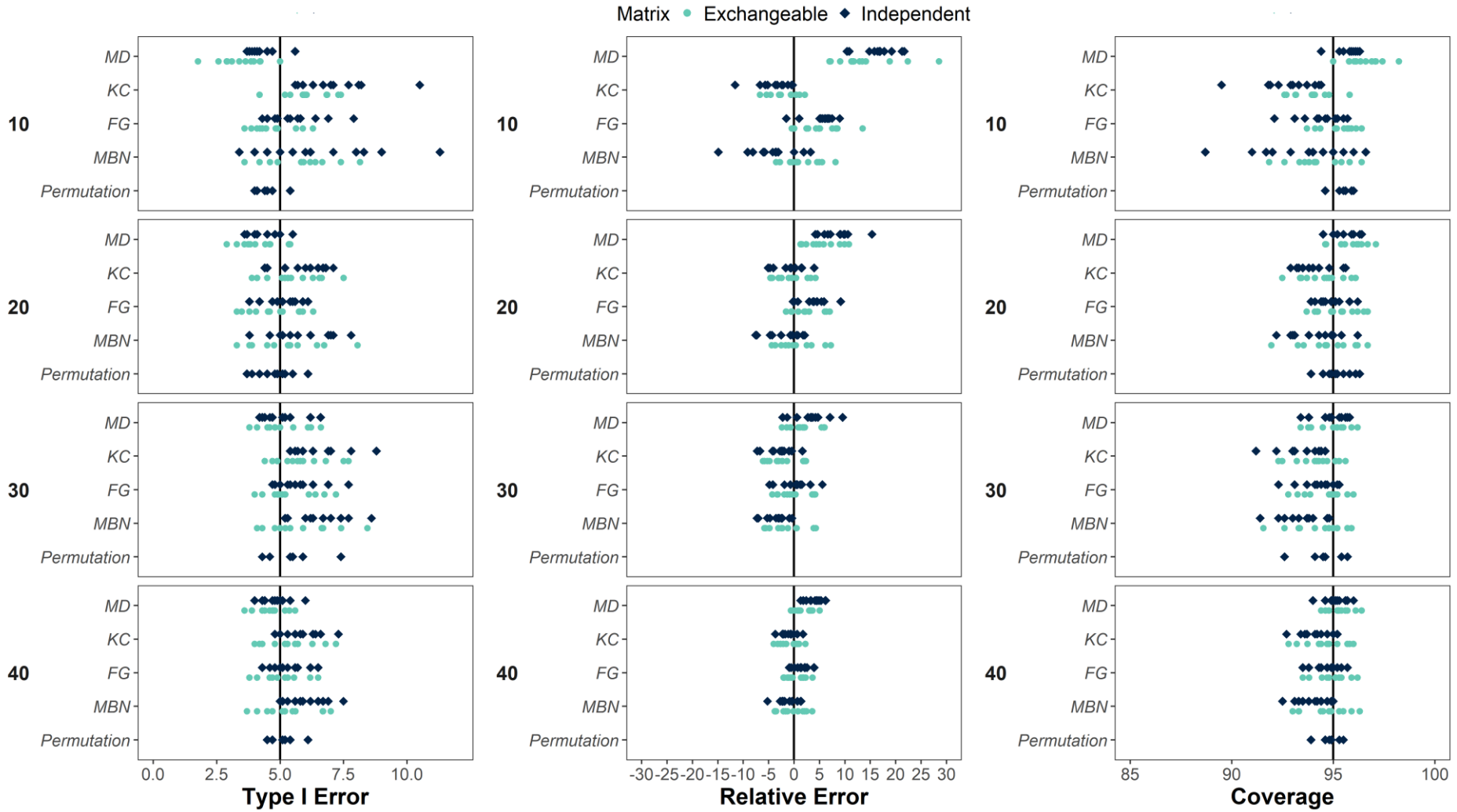


Figure 2: Type I error rate, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the matrix of correlation (colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios without intervention effect are displayed for each performance measure.

Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

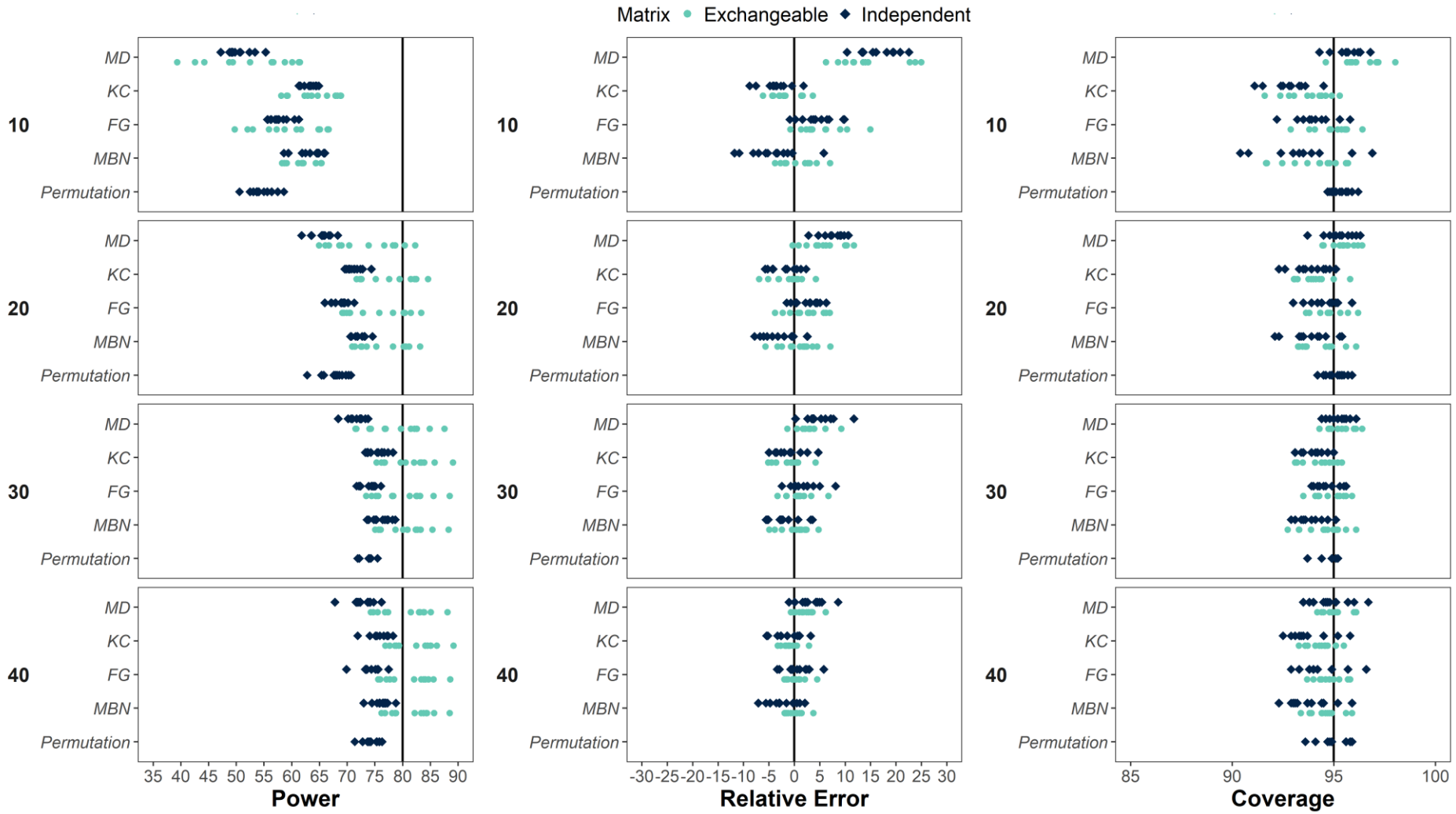


Figure 3: Type I error rate, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in presence of intervention effect by number of clusters (in rows) according to the matrix of correlation (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios with an intervention effect are displayed.

Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

for both correlation matrices. When $K = 20$ or 30, this pattern still applies only with an independent correlation matrix. For $K = 10$, the power of all correction methods is much lower than the 80% nominal rate, but increased with increasing total number of clusters. In addition, for $K = 10$, the power increased as the Kendall's tau increased (Web Figures E3). Otherwise, the values of the mean cluster size, coefficient of variation and Kendall's tau did not seem to affect the power (Web Figures E1, E2 and E3).

Overall, the results for the relative error and the coverage rate were similar than the one observed under scenarios with the null hypothesis.

The power of the small-sample bias correction methods with an independent matrix was similar to the power of the permutation test, but tented to be higher with an exchangeable matrix, especially when $K \geq 20$. The permutation-based confidence intervals produced coverage rate close to 95%. The performance measures of the permutation test were insensitive to the simulation parameters (Web Figures E1, E2 and E3).

For more details on the results of the simulation study, the values of the performance measures of all scenarios are available in the tables in Appendix F.

4. Illustrative example

We illustrated the proposed methods with the cluster-crossover DEMETER trial (clinicaltrials.gov identifier: NCT02515617)[43]. The main objective of this medico-economic study was to assess whether incorporating the subglottic secretion drainage, in addition to the usual measures for the prevention of ventilator-associated pneumonia, was associated with a cost-utility ratio at 1 year, that would make its implementation acceptable in intensive care and emergency departments. Patients in the intervention period were intubated with a specific endotracheal tube allowing subglottic secretion drainage while those in the control period were intubated with a standard endotracheal tube that did not allow such drainage.

For demonstration purposes, we used only the data from the first period to mimic a two parallel-group CRT. Our analysis included 1287 patients from 26 intensive cares from French hospitals, of which 14 were randomised to the control arm (722 patients) and 12 to the intervention arm in the first period (565 patients). The mean cluster size was 49.5 with a coefficient of variation of the cluster size of 0.6.

To illustrate the proposed methods, we focus on death at 90 days after the admission to intensive care, a secondary outcome. There was no loss to follow-up and the administrative censoring rate was 63%. The Kendall's tau was 0.03. Figure 4 shows the Kaplan-Meier

curves. The proportional hazards assumption does not hold, considering the log(-log(S(t)) curves (Web Figure G1). Therefore, proportional hazards models such as the Cox model, could not be performed.

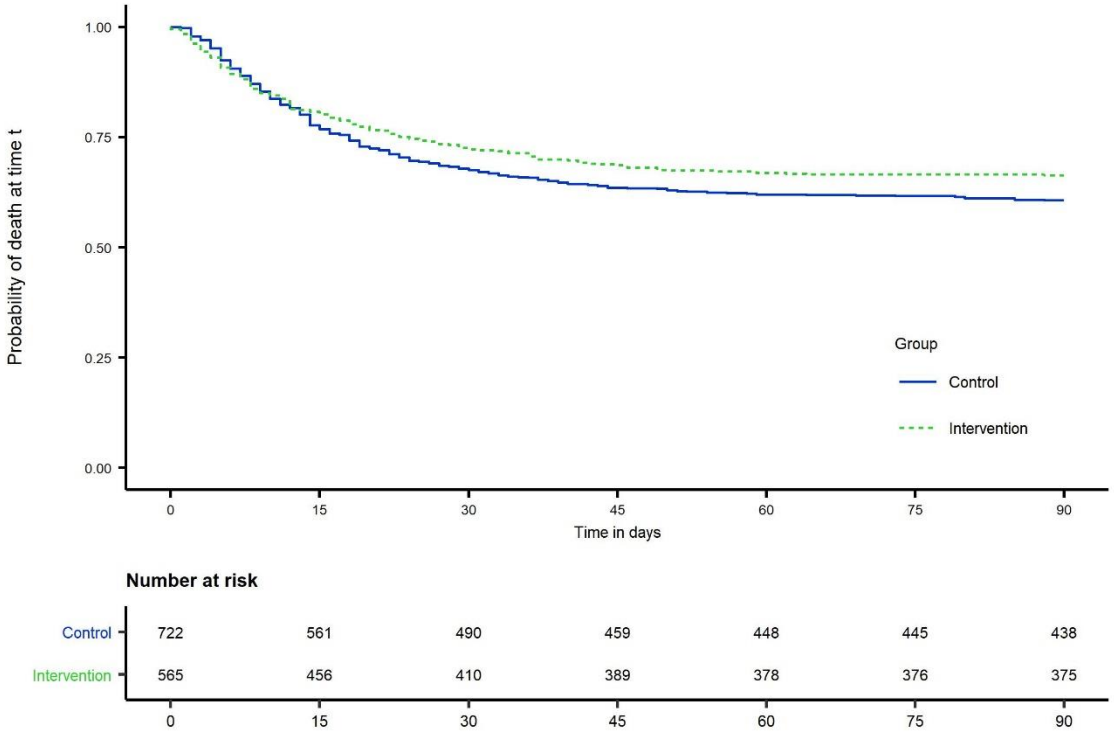


Figure 4: Kaplan-Meier curves for the time to death over the 90 days after the admission in intensive cares

We estimated the $\Delta RMST(t^*)$ at t^*= 90 days using pseudo-values regression. The associated 95% confidence interval was estimated with the MD, KC, FG and MBN variance estimators with a Student distribution with 24 degrees of freedom and both independent and exchangeable matrices and by inverting the permutation test (Table 2). The R code used can be found in Web Appendix G2.

We found a 3.55-days difference between the intervention and control groups for death in the first 90 days after the admission in intensive care. The 4 bias-corrected variance estimators and the permutation test produced very similar confidence intervals. The confidence intervals based on an exchangeable matrix were slightly wider than those obtained from an independent matrix. Of note, the confidence intervals derived from the MD estimator were slightly wider than the confidence intervals estimated with the three other corrections. All the methods led to the same conclusion that adding the subglottic secretion drainage to the usual measures to

prevent ventilator-associated pneumonia had not significant impact on death at 90 days after the admission to intensive care.

Of note, the permutation-based confidence interval was computed in 4 minutes and 10 seconds and the confidence intervals based on the small-sample corrections were computed in less than 1 second for each correction method.

Table 2: *Estimations of the 95% confidence intervals for the time to death in the DEMETER trial.*

Method	Independant matrix	Exchangeable matrix
MD	[-3.20 ;10.35]	[-3.40 ;10.50]
KC	[-2.85 ;10.00]	[-3.11 ;10.21]
FG	[-2.96 ;10.11]	[-3.21 ;10.31]
MBN	[-2.90 ;10.05]	[-3.22 ;10.32]
Permutation	[-2.53 ;10.60]	

5. Discussion

We aimed to evaluate and compare different bias-corrected sandwich estimators, combined with a normal or a Student distribution, and the permutation test in pseudo-values regression for $\Delta RMST(t^*)$ in CRTs with a limited number of clusters. Based on a simulation study we showed that (i) the Student distribution outperforms the normal distribution; (ii) the standard sandwich variance estimator combined with either the normal or the Student distribution does not lead to controlled type I error rate and must be avoided; (iii) the bias-corrected variance estimators alone do not control the type I error and must be combined with a Student distribution; (iv) when the total number of clusters is very limited ($K = 10$), the Fay and Graubard estimator controls the type I error rate, while the Mancl and DeRouen estimator is too conservative and the Kauermann and Carroll and Morel et al. estimators occasionally produce an inflated type I error rate; (v) when $K = 20$, all the estimators, except the Mancl and DeRouen variance estimator, had good performance (vi) the 4 small-sample corrections lead to appropriate type I error when $K \geq 30$; (vii) the exchangeable matrix tends to produce a type I error rate closer to 5% and, when $K \geq 20$, higher power compared to the independent correlation matrix; (viii) the permutation test exhibits an appropriate type I error rate even for a very limited number of clusters, a similar power to the variance corrections with an independent correlation matrix and a lower power than the variance corrections with an exchangeable correlation matrix.

The results of our simulations are similar to those observed in other works evaluating small-sample corrections for GEE in CRTs with continuous outcomes, a situation that could be considered closed to the pseudo-values regression context. Wang et al.[20] also concluded that the Student distribution outperformed the normal distribution across all their considered scenarios. They also evaluated the Mancl and DeRouen, Kauermann and Carroll, Fay and Graubard and Morel et al. variance estimators but found different results. The 4 correction methods produced conservative type I error rate, especially as the number of clusters decreased. The differences could be explained, at least partly, by the assumption of the fixed cluster size (corresponding to a coefficient of variation of the cluster size of 0). Ford and Wesgate[37] studied type I error rates for the Mancl and DeRouen, Kauermann and Carroll and Fay and Graubard estimators with a Student distribution when $K \leq 20$. They found comparable results to ours for the configuration similar to those we evaluated. In addition, they varied the coefficient of variation over a wider range of values (between 0 and 1 with increments of 0.1). Contrary to us, they found the coefficient of variation had an effect on the type I error for the Kauermann and Carroll correction. Type I error increased for increasing coefficients of variation when $K = 10$ and $CV \geq 0.5$ and reached around 8% when $CV = 1$. In our simulation study, we selected parameters to mimic real-world CRTs and therefore limited coefficients of variation of the cluster size to 0.6 and 0.8. We may have missed the potential impact of the coefficient of variation of the cluster size on the performance measures, especially for low values of the coefficient of variation. In addition to this limitation, we did not adjust for covariates in the simulation study. The results could be impacted by the adjustment of covariates and further research is needed.

Several small-sample corrections have been suggested and we have only selected some of them for this simulation study. For example, Ford and Westgate[37] suggested using the average of the Mancl and DeRouen and Kauermman and Carroll estimators, based on the observation that the former tends to be conservative and the latter liberal[45]. However, this suggestion does not rely on real theorical developments. Pan[46] suggested a bias-corrected variance estimator, initially proposed by Liang and Zeger[35], based on a correlation matrix $\boldsymbol{R}(\rho)$ estimated without a parametric specification of its structure. Several extensions have been proposed. Gosho et al.[47] suggested to add a multiplicative correction to the Pan estimator and Wang and Long[48] to combine Pan and Mancl and DeRouen estimators. We decided to evaluate only the estimator proposed by Mancl and DeRouen[22], Kauermann and Carroll[23], Fay and Graubard[24] and Morel, Bokossa and Neerchal[30] because these are the most commonly used corrections for the analysis of CRTs.

Based on our results, we recommend always using a Student distribution with, when $K = 10$, the Fay and Graubard variance estimators, when $K = 20$, either the Kauermann and Carroll, Fay and Graubard or Morel et al. variance estimators, and when $K \geq 30$, any of the 4 corrections. The exchangeable correlation matrix should be preferred when it is possible, that is when there is no convergence problem. Of note, in this work, we assumed non-informative cluster sizes. Kahan et al.[44] have shown that GEE with an exchangeable correlation matrix can produce biased estimates of participant-average and cluster-average intervention effects in CRTs when the cluster size is informative and advised to use an independent matrix in this situation. Therefore, if there is a reason to believe that the cluster size is informative, the independent matrix should be preferred. Finally, when the use of a small-sample correction with an exchangeable matrix is not feasible, the permutation test could also be used regardless of the number of clusters. Of note, the performance measure of the permutation test is equivalent to those of the variance corrections with an independent correlation matrix, but its computation time is much longer.

To conclude, this work has shown that the bias-corrected estimators usually used in CRTs with small number of clusters and the permutation test can be used in practice with the specific method of the pseudo-values regression to estimate a $\Delta RMST(t^*)$. The 4 evaluated small-sample bias-corrected variance estimators, and especially the Fay and Graubard estimator, led to controlled type I error rate and, with an exchangeable correlation matrix, a higher power than the permutation test.

Acknowledgments

The authors benefitted from the use of the cluster at the Centre de Calcul Scientifique en région Centre-Val de Loire (CaSciModOT).

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

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].

Supplemental Material

Web Appendix for this article is available online.

References

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

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

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

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

9. 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.

10. 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.

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

12. 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.

13. Le Vilain-Abraham F, Tavernier E, Dantan E, Desmée S, Caille A.· Restricted mean survival time to estimate an intervention effect in a cluster randomized trial. *Stat Methods Med Res 2023*: 9622802231192960.

14. Kaplan EL and Meier P. Nonparametric Estimation From Incomplete Observations. *J Am*

Stat Assoc 1958; 53: 457-481.
15. 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.
16. Andersen PK and Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010; 19:71-99.
17. 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.
18. Eldridge S and Kerry SM. *A practical guide to cluster randomised trials in health services research*. Chichester, West Sussex: John Wiley & Son, 2011.
19. Wang M, Kong L, Li Z, et al. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Stat Med* 2016; 35: 1706-21.
20. Li P and Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat Med* 2015; 34: 281–96.
21. Pan W and Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med* 2002; 21: 1429-41.
22. Mancl LA and DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; 57: 126-34.
23. Kauermann G and Carroll RJ. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc* 2001; 96: 1387–1396.
24. Fay MP and Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; 57: 1198–1206.
25. 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.
26. 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.
27. 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.
28. Rabideau DJ and Wang R. Randomization-based confidence intervals for cluster randomized trials. *Biostatistics* 2021; 22: 913-927.
29. Garthwaite, PH. Confidence Intervals from Randomization Tests. *Biometrics* 1996; 52: 1387–1393

30. Morel J, Bokossa M, and Neerchal N. Small sample correction for the variance of GEE estimators. *Biom J* 2003; 45: 395–409.
31. Ford WP and Westgate PM. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biom J* 2017; 59: 478-495.
32. 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.
33. Wang X, Turner EL, and Li F. Improving sandwich variance estimation for marginal Cox analysis of cluster randomized trials. *Biom J* 2022: e2200113.
34. Chen X and Li F. Finite-sample adjustments in variance estimators for clustered competing risks regression. *Stat Med* 2022; 41: 2645-2664.
35. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.
36. Zeger SL and Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42: 124-130.
37. Gallis JA, Li F, and Turner EL. xtgeebcv: A command for bias-corrected sandwich variance estimation for GEE analyses of cluster randomized trials. *Stata J* 2020; 20: 363-381.
38. Fai AHT, Cornelius PL. Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation* 1996; 54: 363–378.
39. Thompson JA, Leyrat C, Fielding KL, Hayes RJ. Cluster randomised trials with a binary outcome and a small number of clusters: comparison of individual and cluster level analysis method. *BMC Med Res Methodol* 2022; 22: 222.
40. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med* 2013; 32: 4118-34.
41. 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.
42. 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.
43. Lachérade JC. Drainage des sécrétions sous-glottiques et pneumonies acquises sous ventilation mécanique. *Réanimation* 2013; 22: 257-264.

44. Kahan BC, Li F, Copas AJ, et al. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *Int J Epidemiol* 2023; 52:107-118.

45. Ford WP, Westgate PM. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biom J* 2017; 59: 478-495.

46. Pan W. On the robust variance estimator in generalised estimating equations. *Biometrika* 2001; 88: 901–906.

47. Gosho M, Sato Y, and Takeuchi H. Robust covariance estimator for small-sample adjustment in the generalized estimating equations: A simulation study. *Science Journal of Applied Mathematics and Statistics* 2014; 2: 20-25.

48. Wang M, Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Stat Med* 2011; 30:1278–1291.

Web Appendix for “Restricted mean survival time in cluster randomized trials with a small number of clusters: Improving variance estimation of the pseudo-values regression”

Le Vilain-Abraham F, Desmée S, Thompson JA, Tavernier E, Dantan E, Caille A

Table of contents

Web appendix A: Permutation-based confidence intervals 3
Web Appendix B: Additional details for the simulation study 6
B.1. Censoring times generating mechanism 6
B.2. Selection of the simulated intervention effect 6
B.3. Performance measures of the simulation study 6
B.3.1. True difference in RMST 6
B.3.2. Definition of the performance measures 7
Web Appendix C: Simulation results for the convergence 9
Web Appendix D: Simulation results in absence of intervention effect 10
Web Appendix E: Simulation results in presence of intervention effect 13
Web Appendix F: Complementary results for the simulation study 16
Web Appendix G: Checking of the proportional hazards assumption and R code for the illustrative example DEMETER 32
F.1. Checking of the proportional hazards assumption 32
G.2. R code to estimate the difference in RMST 32
G.2.1. Estimation of the difference in RMST 33
G.2.2. Confidence interval with the MD estimator 33
G.2.3. Confidence interval with the KC estimator 33
G.2.4. Confidence interval with the FG estimator 33
G.2.5. Confidence interval with the MBN estimator 34
G.2.6. Permutation-based confidence interval 34

Web appendix A: Permutation-based confidence intervals

We consider the model (1) described in the main text:

$$E\left[\min(T_{kl}, t^*) \mid \widetilde{\boldsymbol{Z}}_{kl}\right] = \boldsymbol{\beta}' \widetilde{\boldsymbol{Z}}_{kl} = \beta_0 + \beta_1 Z_{kl} \quad (1)$$

We computed a 100(1-α)% permutation-based confidence interval for the intervention effect β_1 by inverting the permutation test. It requires to determine the set of values β^* such as the non-zero null hypothesis $H_0: \beta_1 = \beta^*$ is not rejected by the permutation test. For this, the model (1) is rewritten as:

$$\begin{aligned} E[\min(T_{kl}, t^*) \mid \tilde{\mathbf{Z}}_l] &= \beta_0 + \beta^* Z_{kl} + (\beta_1 - \beta^*) Z_{kl} \\ &= \beta_0 + \beta^* Z_{kl} + \delta_{offset} Z_{kl} \end{aligned} \quad (2)$$

This model (2) lead to test an equivalent zero null hypothesis $H_0: \delta_{offset} = (\beta_1 - \beta^*) = 0$. To do so, model (2) is fitting with the pseudo-values regression with the observed data set to obtain the observed test statistic. Then, the observed test statistic is compared to permuted test statistics estimated by fitting the following model with the pseudo-values regression with several permuted intervention allocations:

$$E[\min(T_l, t^*) \mid \tilde{\mathbf{Z}}_l] = \beta_0 + \beta^* Z_l^{obs} + \delta_{offset} Z_l^{perm} \quad (3)$$

where Z_l^{obs} is the intervention allocation in the observed dataset, $\beta^* Z_l^{obs}$ is the fixed offset term, and Z_l^{perm} is a permuted intervention allocation.

This method requires to carry out a large number of permutation tests, for each considered β^*, and each permutation test is performed for a large number of permuted intervention allocations. Therefore, inverting the permutation test could be computationally intensive. To address this, Rabideau and Wang[1] introduced an efficient technique, inspired by a method proposed by Garthwaite[2], to reduce computational time. This method uses a step-by-step search, performed independently for both lower and upper bounds, and considers a single permuted intervention allocation for each permutation test. We used this method to compute

permutation-based confidence intervals. We briefly describe the method below. See Rabideau and Wang[1] and Garthmaite[2], for more details.

A.1. Search procedure

For the upper bound of the 100(1-α)% confidence interval of the intervention effect β_1, denoted U, we considered a sequential I-step search. In the *i*th step, $\beta^* = U_i$, the current value of U. The allocations of the clusters are randomly allocated between the control and intervention groups and the model (3) is fitted with this permutated allocation to estimate the permuted test statistic T_i^{perm}. The observed test statistic T_i^{obs} is computed by fitting the model (2) with the observed data. The upper bound is updated comparing the permuted and the observed test statistics.

$$U_{i+1} = \begin{cases} U_i - \dfrac{c(\alpha/2)}{i}, & \text{if } T_i^{perm} > T_i^{obs} \\ U_i + \dfrac{c(1-\alpha/2)}{i}, & \text{if } T_i^{perm} \leq T_i^{obs} \end{cases}$$

with c the step length constant that is a positive constant.

Separately, the lower bound L is computed with the similar I-step search, using

$$L_{i+1} = \begin{cases} L_i + \dfrac{c(\alpha/2)}{i}, & \text{if } T_i^{perm} < T_i^{obs} \\ L_i - \dfrac{c(1-\alpha/2)}{i}, & \text{if } T_i^{perm} \geq T_i^{obs} \end{cases}$$

The updated values at the *I*th step, are the estimated bounds of the 100(1-α)% permutation-based confidence interval.

A.2. Implementation

We followed the recommendations of Rabideau and Wang[4] and Garthwaite[5] to set the parameters. At the *i*th step, we used $c = \kappa(U_i - \hat{\beta}_1)$ and $c = \kappa(\hat{\beta}_1 - L_i)$ for the upper and lower bound searchs, respectively, with $\kappa = 2/(z_{\alpha/2}(2\pi)^{-\frac{1}{2}}\exp(-z_{\alpha/2})/2)$, α the

significance level of the permutation test, $z_{\alpha/2}$ the $\alpha/2$-quantile of the standard normal distribution and $\hat{\beta}_1$ the estimation of β_1 obtained by fitting the model (1) with the observed dataset. Let t_1 and t_2 be the second smallest and second largest permuted estimations of δ_{offset} from a permutation test testing $H_0: \beta_1 = \hat{\beta}_1$, with $\lceil (4-\alpha)/\alpha \rceil$ permutated allocations. We initialized the lower bounds at $\hat{\beta}_1 - \frac{t_2 - t_1}{2}$ and the upper bound at $\hat{\beta}_1 + \frac{t_2 - t_1}{2}$. At the beginning, $i = \min(\lceil 0.3(4-\alpha)/\alpha \rceil, 50)$, to prevent large change in the early part of the search. Finally, we set $I = 5000$ iterations to achieve an asymptotic variance of the coverage of the 95% confidence interval of 1.3×10^{-5}.

References

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

2. Garthwaite, PH. Confidence Intervals from Randomization Tests. *Biometrics* 1996; 52: 1387–13936. Srivastava DK, George EO, Lu Z, et al. Impact of unequal censoring and insufficient follow-up on comparing survival outcomes: Applications to clinical studies. *Stat Methods Med Res* 2021; 30: 2057-2074.

Web Appendix B: Additional details for the simulation study

B.1. Censoring times generating mechanism

We assumed that the censoring times and the event times were independent and that the censoring times within a same cluster were independent. We considered the censoring rate ψ. For the individual l from the cluster k ($l = 1, \dots, m_k, k = 1, \dots, K$), we censored the observation if $R_{kl} < \psi$, with R_{kl} drawn from a continuous uniform distribution on 0 to 1. If the observation was censored, we simulated the censoring time using a uniform distribution on 0 to T_{kl}, with T_{kl} the time-to-event of the individual.[1]

B.2. Selection of the simulated intervention effect

For each scenario, we fixed the number of clusters K, the mean cluster size m, the coefficient of variation of the cluster size CV and the degree of clustering τ. The intervention effect to achieve 80% power was obtained iteratively. For each iteration, 20,000 CRTs were simulated. For each CRT, a 95% confidence interval for $\Delta RMST(t^*)$ was computed using the pseudo-value regression with the standard sandwich estimators and the normal distribution of the test statistic. We used an independent correlation matrix, which is not subject to non-convergence unlike the exchangeable correlation matrix. Power was estimated as the proportion of the 20,000 95% confidence intervals for $\Delta RMST(t^*)$ that did not contain 0. The value of the intervention effect, β_G, in the next step was selected regarding to the power obtained. For example, if the power was greater than 80%, a value closer to 0 was chosen, and vice versa. The intervention effect in the first iteration was arbitrarily initialised at $\log(0.6)$. In the second iteration, β_G was fixed at $\log(0.7)$ or $\log(0.5)$, depending on whether the power of the first iteration was higher or lower than 80%. For the next steps, the intervention effect was selected as follows. If the power was at least once higher and once lower than 80% in the previous steps, the intervention effect was selected using the bisection method. Otherwise, the exponential of the intervention effect was increased or decreased by 0.1. The iterations were stopped when the rounded power was 80% and the power at least 80%. The value of β_G obtained in the last iteration was chosen for the simulation study.

B.3. Performance measures of the simulation study

B.3.1. True difference in RMST

Under the null hypothesis ($\beta_G = 0$), the true $\Delta RMST(t^*) = 0$. For the scenarios with a non-null intervention effect, the true difference in RMST up to t^* was calculated as the difference

between the integrals of the marginal survival functions in the intervention and the control groups, denoted $S(t, Z = 1)$ and $S(t, Z = 0)$, respectively, from 0 to t^*.

$$\int_0^{t^*} S(t, Z = 1)dt - \int_0^{t^*} S(t, Z = 0)dt \quad (1)$$

Let $S(t, Z = j \mid u_k)$ be the conditional survival function in intervention group j ($j = 0,1$) and $f(\cdot)$ the density function of the distribution gamma $\Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$ with a mean of 1 and variance θ. The marginal survival function in group j ($j = 0,1$) was obtained from the conditional survival function by integrating out the frailty.[2]

$$S(t, Z = j) = \int_0^{+\infty} S(t, Z = j|u_k) f(u_k) du_k \quad (2)$$

The Laplace transform, $\mathcal{L}(.)$, led to the following analytic form:[2]

$$S(t, Z = j) = \mathcal{L}(H(t)) = \frac{1}{\left(1 + \frac{\lambda t^{\rho} \exp(Z\beta_G)}{1/\theta}\right)^{1/\theta}}$$

where $H(.)$ is the cumulative hazard function.

Then the numerical integration method of the right Riemann sum was used to estimate the integrals in (1).

B.3.2. Definition of the performance measures

Let's denote $\Delta\widehat{RMST}(t^*)_d$ and $\hat{V}(\Delta\widehat{RMST}(t^*)_d)$ the estimations of the $\Delta RMST(t^*)$ and its variance in the simulated dataset d (d = 1,…, 1000), respectively. The performance of the methods was assessed with the following criteria:

(1) the type I error rate that corresponds to the proportion of 95% confidence intervals for $\Delta RMST(t^*)$ that do not contain 0 when there is no intervention effect (i.e., when $\beta_G = 0$)

(2) the relative error of the asymptotic standard errors (ASE) that corresponds to $\frac{ASE - ESE}{ESE} \times 100$, where $ASE = \sqrt{\frac{1}{1000}\sum_{d=1}^{1000} \hat{V}(\Delta\widehat{RMST}(t^*)_d)}$ and $ESE = \sqrt{\frac{1}{999}\sum_{d=1}^{1000}\left(\Delta\widehat{RMST}(t^*)_d - \overline{\Delta RMST}(t^*)\right)^2}$, with ESE the empirical standard error and $\overline{\Delta RMST}(t^*) = \frac{1}{1000}\sum_{d=1}^{1000} \Delta\widehat{RMST}(t^*)_d$. The relative error represents the bias in estimating the true standard error of the $\Delta RMST(t^*)$. A relative error near to 0 indicates a good estimation of the asymptotic standard error. If the relative error

deviates from 0, it reflects overestimation of the variance if relative error is positive and an underestimation otherwise.

(3) the empirical coverage rate of the nominal 95% confidence intervals for $\Delta RMST(t^*)$ defined as the proportion of 95% confidence intervals that contain the true RMST difference (defined in Appendix A.3.1).

(4) the power that corresponds to the proportion of 95% confidence intervals for $\Delta RMST(t^*)$ that do not contain 0 when there is an intervention effect (i.e., when $\beta_G \neq 0$ thus $\Delta RMST(t^*) \neq 0$).

References

1. Srivastava DK, George EO, Lu Z, et al. Impact of unequal censoring and insufficient follow-up on comparing survival outcomes: Applications to clinical studies. *Stat Methods Med Res* 2021; 30: 2057-2074.

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

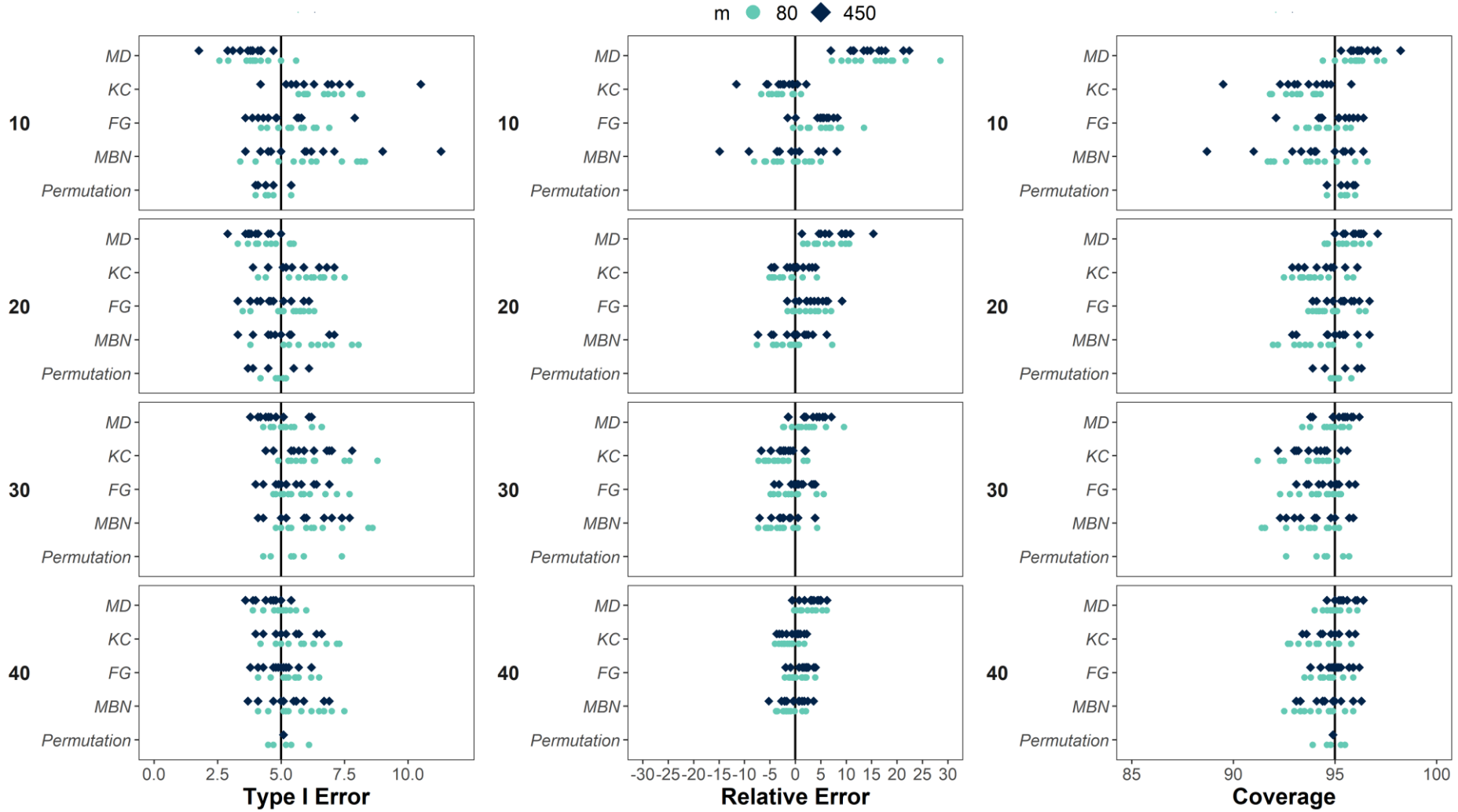
Web Appendix C: Simulation results for the convergence

Table C1: Scenario settings and percentage of non-convergence for the pseudo-values regression with an exchangeable correlation matrix. There were no convergence issues for the scenarios not reported in the Table.

Total number of clusters	Mean cluster size	Coefficient of variation	Kendall's tau τ	Percentage of non-convergences
Null hypothesis ($\beta_G = 0$)				
10	80	0.6	0.001	14.7
10	80	0.6	0.01	4.1
10	80	0.8	0.001	21.5
10	80	0.8	0.01	6.7
10	80	0.8	0.1	0.1
10	450	0.6	0.001	6.9
10	450	0.6	0.01	0.1
10	450	0.8	0.001	9.6
10	450	0.8	0.01	0.6
20	80	0.6	0.001	5.1
20	80	0.6	0.01	0.3
20	80	0.8	0.001	8.1
20	80	0.8	0.01	0.9
20	450	0.6	0.001	0.7
20	450	0.8	0.001	1.5
30	80	0.6	0.001	2.2
30	80	0.6	0.01	0.1
30	80	0.8	0.001	5.2
30	450	0.6	0.001	0.1
30	450	0.8	0.001	0.2
40	80	0.6	0.001	1.4
40	80	0.8	0.001	2.8
40	450	0.6	0.001	0.1
Intervention effect ($\beta_G \neq 0$)				
10	80	0.6	0.001	13.6
10	80	0.6	0.01	3.6
10	80	0.8	0.001	19.0
10	80	0.8	0.01	5.9
10	80	0.8	0.1	0.1
10	450	0.6	0.001	5.7
10	450	0.6	0.01	0.2
10	450	0.8	0.001	12.3
20	80	0.6	0.001	5.4
20	80	0.6	0.01	0.1
20	80	0.8	0.001	8.1
20	80	0.8	0.01	0.9
20	450	0.6	0.001	1.2
20	450	0.8	0.001	1.7
30	80	0.6	0.001	2.0
30	80	0.8	0.001	5.0
30	80	0.8	0.01	0.2
30	450	0.6	0.001	0.2
30	450	0.8	0.001	0.3
40	80	0.6	0.001	1.4
40	80	0.8	0.001	2.7
40	450	0.8	0.001	0.2

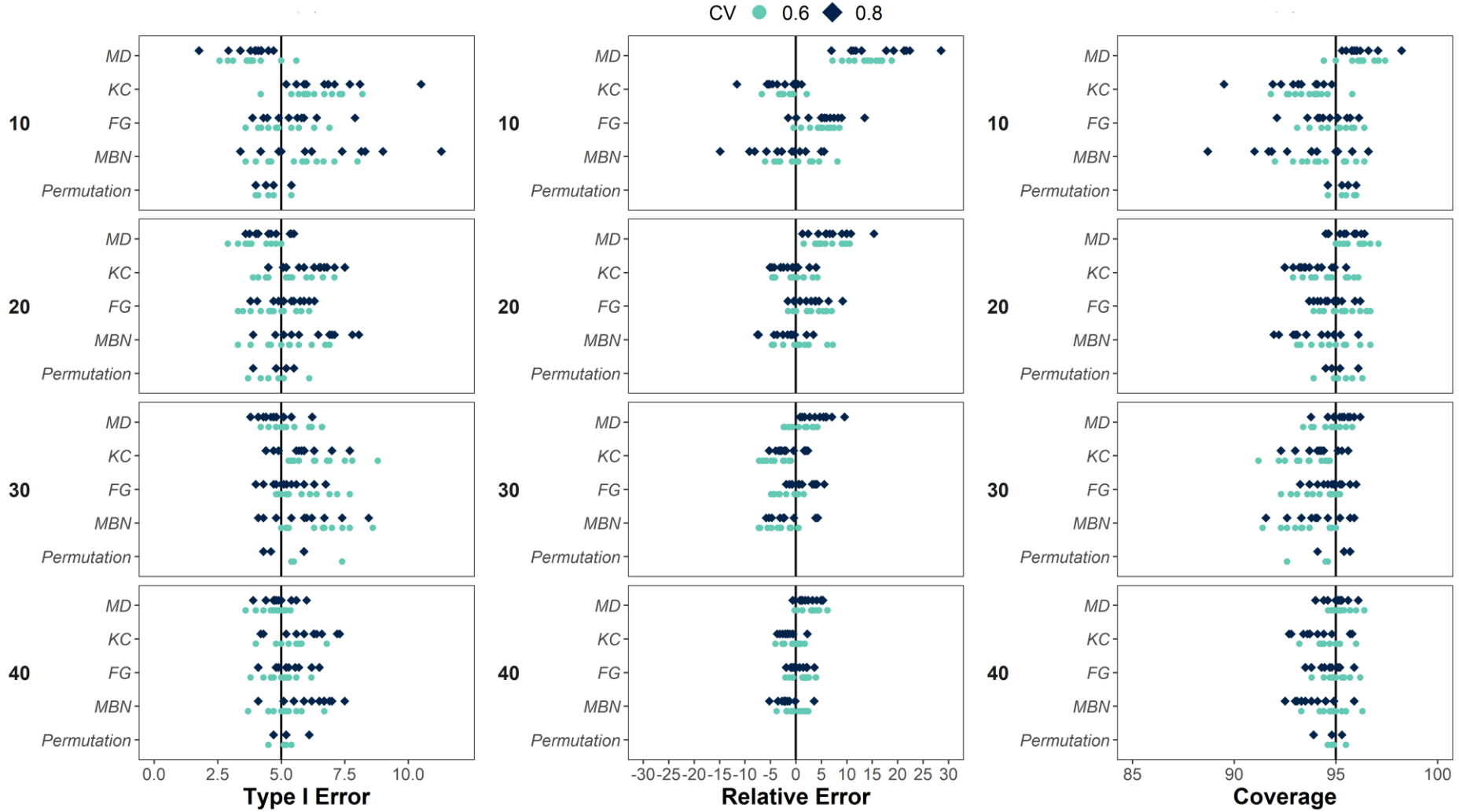
Web Appendix D: Simulation results in absence of intervention effect

Figure D1: Type I error rate, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) according to the mean cluster size m (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios without intervention effect are displayed for each performance measure.



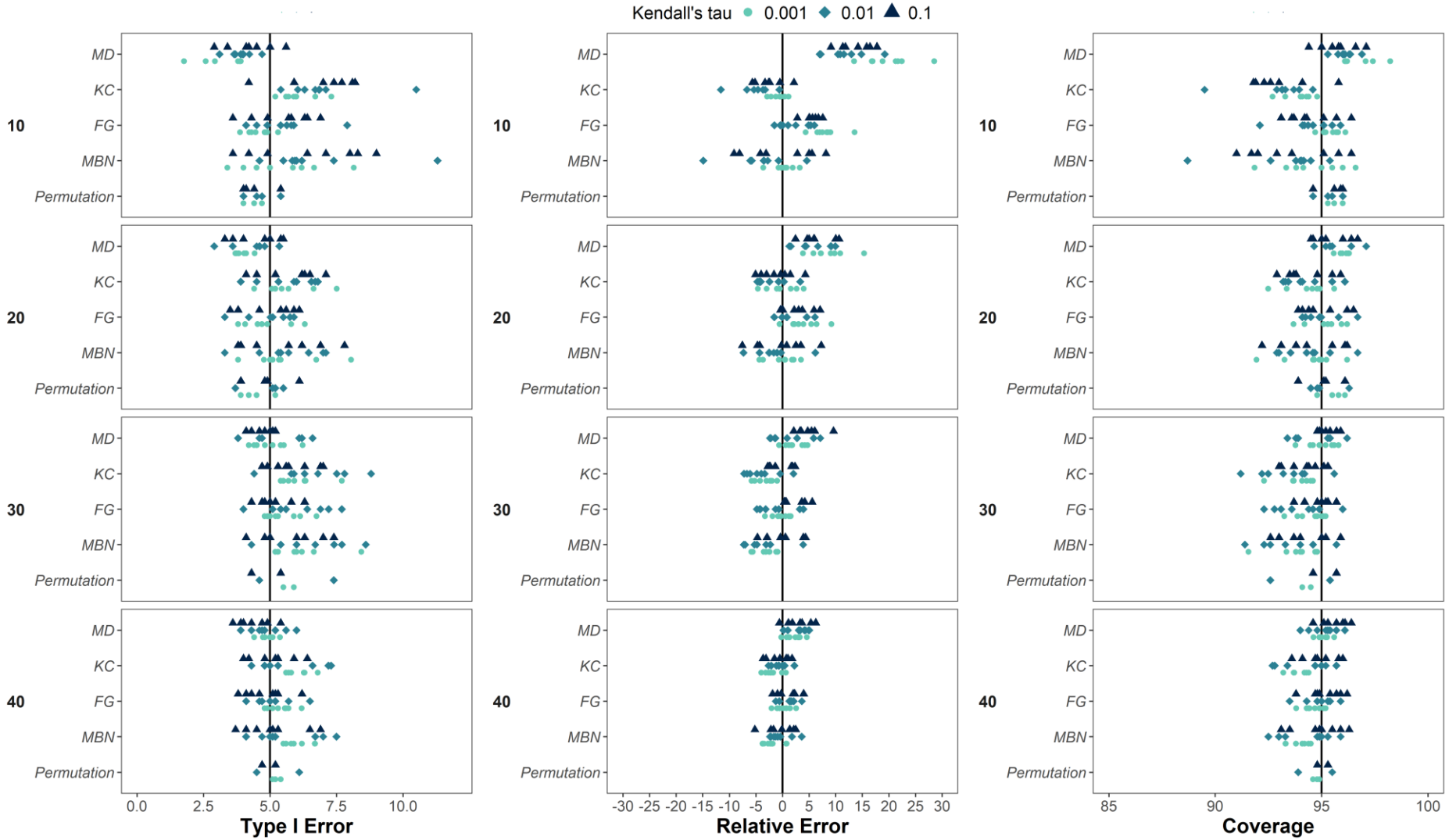
Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

Figure D2: Type I error rate, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the coefficient of variation of the cluster size CV (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios without intervention effect are displayed for each performance measure.



Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

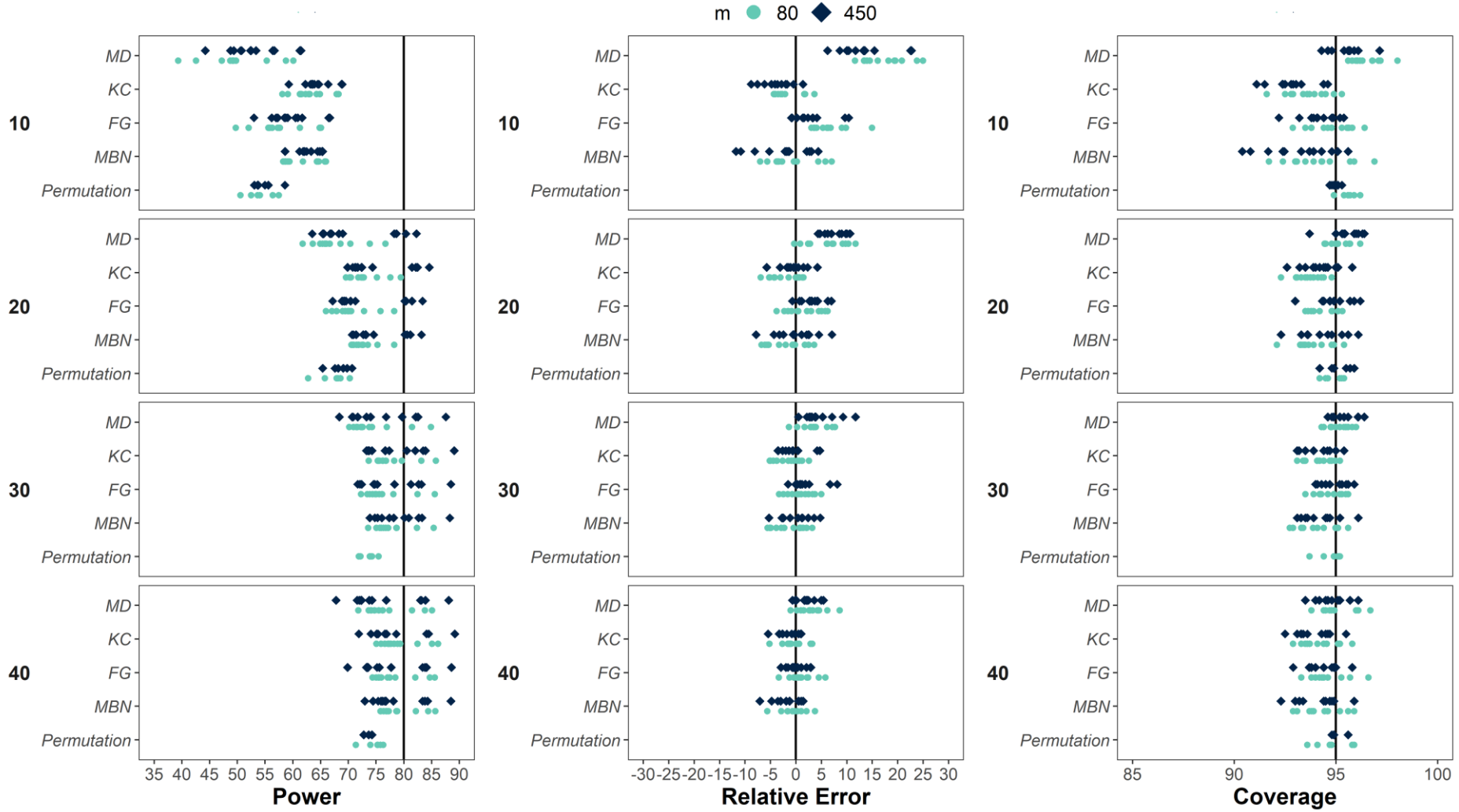
Figure D3: Type I error rate, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the Kendall's tau (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios without intervention effect are displayed for each performance measure.



Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

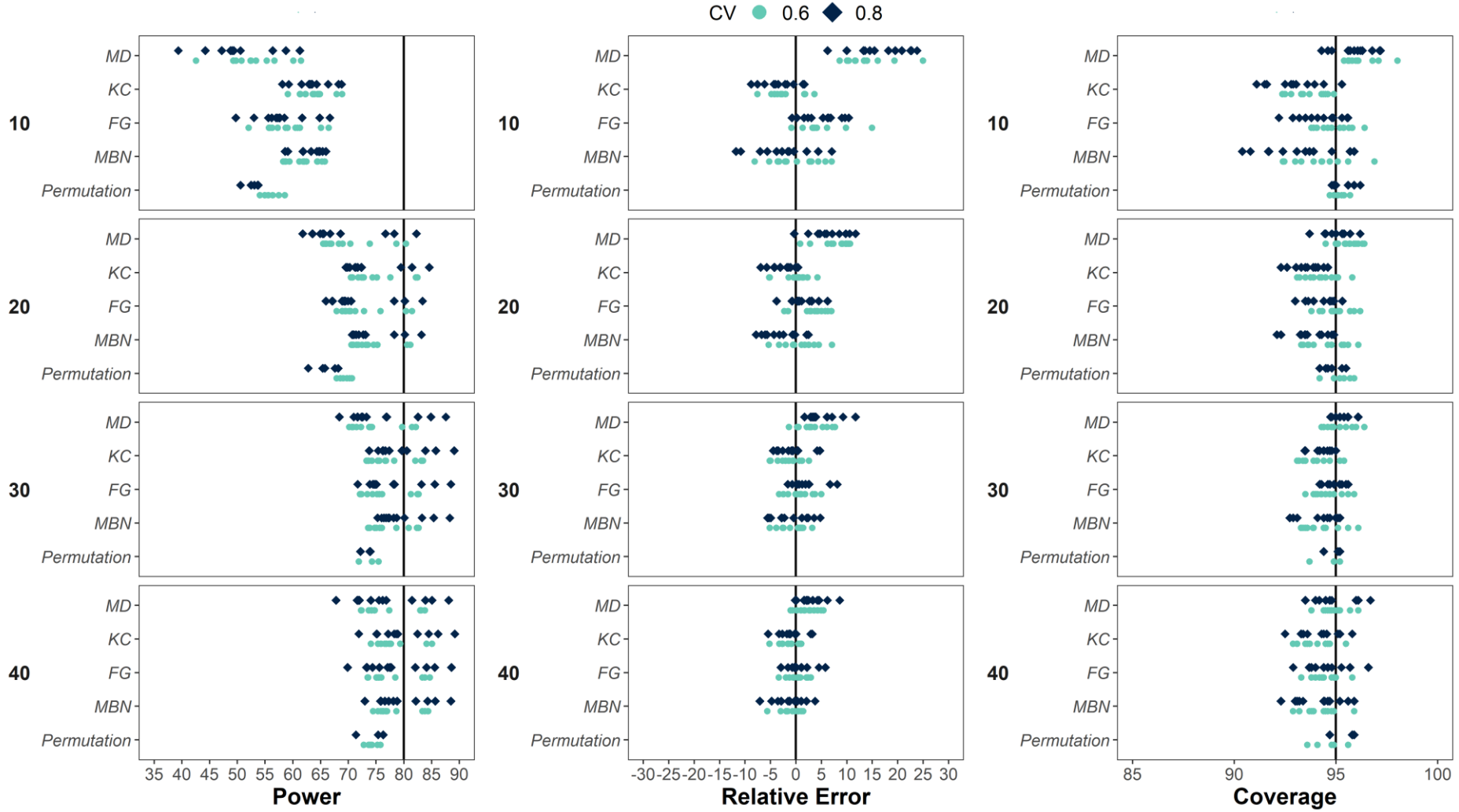
Web Appendix E: Simulation results in presence of intervention effect

Figure E1: Power, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the mean cluster size m (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios with an intervention effect are displayed for each performance measure.



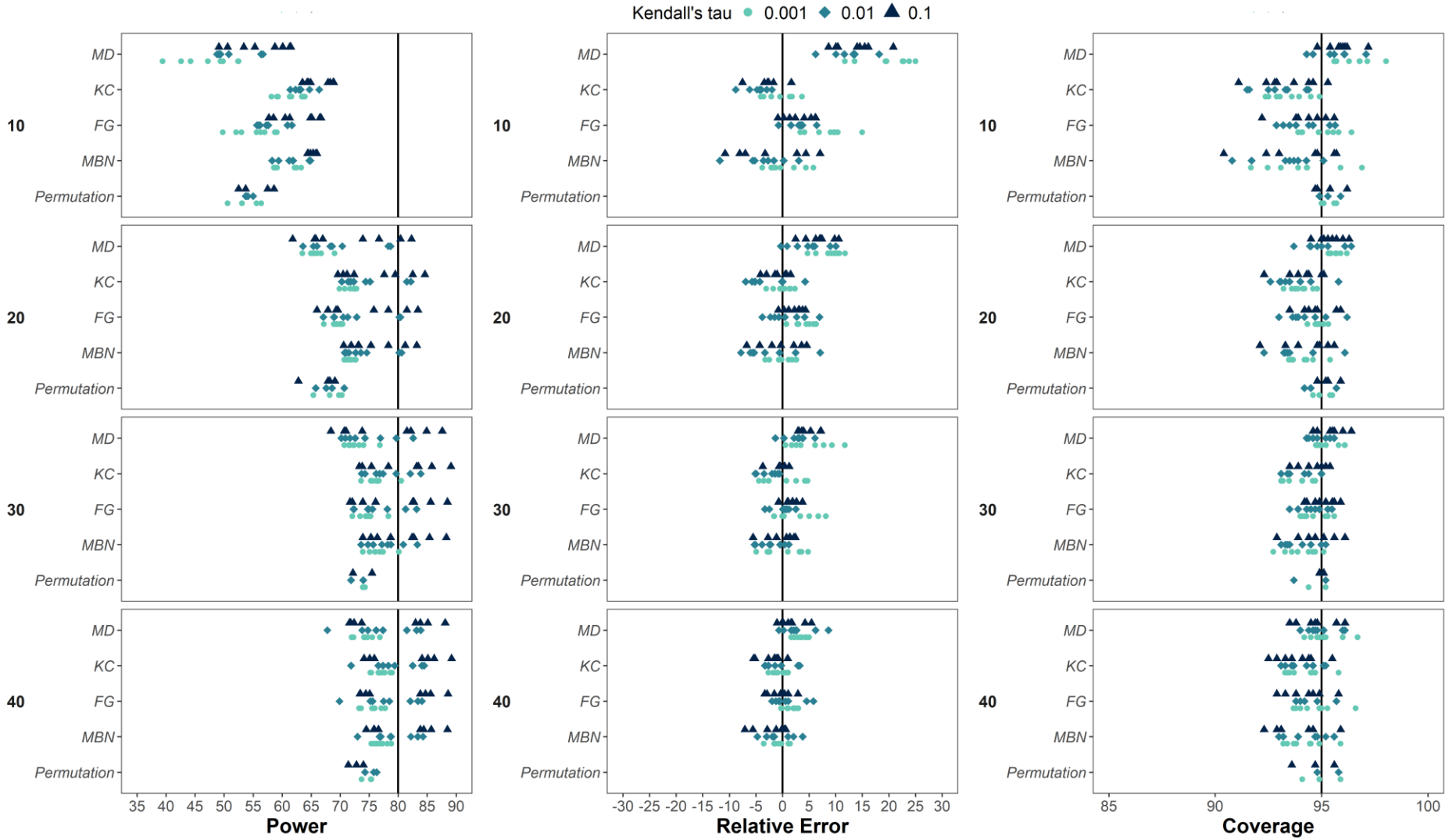
Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

Figure E2: Power, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the coefficient of variation of the cluster size CV (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios with an intervention effect are displayed for each performance measure.



Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

Figure E3: Power, relative error and coverage rate of the four correction estimators with a Student distribution and permutation test in absence of intervention effect by number of clusters (in rows) and according to the Kendall's tau (in colour) for all scenarios. Each dot corresponds to a scenario, summarised over the iterations that converged. All 48 scenarios with an intervention effect are displayed for each performance measure.



Note: No relative error is reported for the permutation test because the inversion of the permutation test directly provides a confidence interval without estimating the variance of the intervention effect.

Web Appendix F: Complementary results for the simulation study

Table F1: Type I error rates (%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an independent matrix. The distribution of the Wald test statistic is either a normal distribution (z) or a Student distribution (t).

K	m	CV	τ	V		V_{MD}		V_{KC}		V_{FG}		V_{MBN}		*Permutation test*
				z	t	z	t	z	t	z	t	z	t	
10	80	0.6	0.001	13.4	8.8	6.0	3.8	9.3	5.7	7.9	4.9	7.2	4.0	4.7
			0.01	15.5	10.0	6.7	3.9	10.5	6.7	8.2	5.4	9.8	5.5	4.5
			0.1	16.7	11.4	8.8	5.6	11.9	8.2	10.5	6.9	11.5	8.0	5.4
		0.8	0.001	15.2	10.2	6.8	3.8	10.3	6.7	8.9	5.3	6.8	3.4	4.4
			0.01	17.1	11.3	6.8	4.0	11.5	7.1	9.3	5.8	10.5	6.2	4.7
			0.1	17.8	11.5	7.5	4.5	11.4	8.1	9.6	6.4	11.9	8.3	4.0
	450	0.6	0.001	13.9	9.8	6.3	3.9	10.1	5.9	8.2	4.8	8.2	4.5	4.0
			0.01	14.2	10.1	6.4	3.7	10.3	6.3	8.6	4.5	10.0	6.0	4.0
			0.1	16.7	11.0	7.5	4.2	11.7	7.0	9.3	5.7	11.4	7.1	4.1
		0.8	0.001	15.6	10.7	5.5	3.8	10.4	5.6	7.7	4.3	9.5	5.0	4.7
			0.01	20.1	16.3	9.6	4.7	15.8	10.5	12.5	7.9	16.2	11.3	5.4
			0.1	18.6	13.4	7.4	4.1	12.9	7.7	9.9	5.8	13.4	9.0	4.4
20	80	0.6	0.001	9.4	6.6	4.6	3.7	6.6	4.4	5.6	3.8	5.8	3.8	4.2
			0.01	10.8	7.6	5.8	4.8	7.6	6.0	6.7	5.1	7.5	5.7	5.1
			0.1	9.3	7.8	6.1	4.8	7.4	6.2	6.3	5.6	7.5	6.2	4.9
		0.8	0.001	9.6	7.8	5.6	4.1	7.6	5.7	6.6	4.9	7.2	5.1	5.2
			0.01	11.2	9.0	6.0	4.8	8.7	6.7	7.1	5.5	8.6	7.0	5.2
			0.1	12.0	10.0	6.8	5.5	9.2	7.1	7.7	6.1	9.6	7.8	4.8
	450	0.6	0.001	7.8	6.5	5.3	3.7	6.7	5.2	5.9	4.7	6.2	5.0	4.5
			0.01	9.5	6.7	4.4	3.6	6.6	4.5	5.5	4.2	6.4	4.6	3.7
			0.1	11.6	9.2	6.6	5.0	9.1	7.1	7.8	6.1	9.1	6.9	6.1
		0.8	0.001	8.5	6.9	4.7	4.1	6.5	5.2	5.5	4.7	6.7	5.4	3.9
			0.01	11.7	9.5	6.6	4.5	9.1	6.8	7.4	5.9	9.5	7.1	5.5
			0.1	11.4	9.5	5.6	4.0	8.4	6.5	7.1	5.4	9.3	6.9	3.9
30	80	0.6	0.001	8.1	7.0	5.6	5.1	7.0	5.4	6.1	5.3	6.7	5.3	5.5
			0.01	11.0	10.0	8.5	6.6	9.8	8.8	9.1	7.7	9.6	8.6	7.4
			0.1	8.6	7.5	6.0	5.2	7.4	6.3	6.9	5.8	7.3	6.3	5.4
		0.8	0.001	9.2	7.9	6.2	5.4	7.7	6.3	6.8	5.9	7.4	6.2	5.9
			0.01	9.4	7.8	5.6	4.7	7.0	5.9	6.2	5.4	7.6	6.0	4.6
			0.1	8.5	6.9	5.1	4.3	6.7	5.6	5.8	4.7	6.8	6.0	4.3
	450	0.6	0.001	7.0	6.1	5.3	4.2	5.9	5.4	5.6	4.8	6.0	5.2	
			0.01	10.5	9.1	7.5	6.2	8.6	7.8	8.0	6.9	8.6	7.7	

			0.1	9.0	7.9	6.4	5.1	7.8	6.9	7.3	5.8	7.9	7.0	
		0.8	0.001	8.4	7.3	4.9	4.4	6.7	5.7	6.2	5.0	7.0	6.0	
			0.01	9.1	8.1	5.5	4.6	7.6	6.3	6.4	5.6	7.9	6.7	
			0.1	9.1	8.3	6.6	5.1	7.6	7.0	7.1	6.3	8.1	7.4	
40	80	0.6	0.001	8.1	7.4	5.8	5.1	7.4	5.8	6.8	5.6	6.9	5.8	5.4
			0.01	6.8	6.2	4.7	4.3	6.2	5.0	5.5	4.6	6.1	5.1	4.5
			0.1	7.1	6.3	5.3	4.9	6.1	5.3	5.6	5.1	6.2	5.3	5.2
		0.8	0.001	7.8	7.4	6.1	5.0	7.1	6.3	6.5	5.7	7.0	6.2	5.2
			0.01	8.8	8.6	6.8	6.0	8.4	7.3	7.4	6.5	8.5	7.5	6.1
			0.1	8.2	7.5	5.4	4.9	6.8	5.9	6.3	5.3	7.3	6.5	4.7
	450	0.6	0.001	7.0	6.5	5.6	5.0	6.3	5.6	5.8	5.1	6.1	5.6	5.1
			0.01	6.8	5.8	5.0	4.7	5.5	5.0	5.1	5.0	5.4	5.0	
			0.1	7.3	5.9	4.7	4.0	5.9	4.8	5.2	4.3	5.9	5.0	
		0.8	0.001	7.5	6.6	5.1	4.4	6.3	5.6	5.7	4.8	6.4	5.9	
			0.01	8.6	7.4	6.0	4.8	7.1	6.6	6.8	5.7	7.3	6.7	
			0.1	8.4	7.8	6.3	5.4	7.3	6.4	6.9	6.2	7.6	6.9	

Table F2: Type I error rates (%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an exchangeable matrix. The distribution of the Wald test statistic is either a normal distribution (z) or a Student distribution (t).

K	m	CV	τ	V		V_{MD}		V_{KC}		V_{FG}		V_{MBN}	
				z	t	z	t	z	t	z	t	z	t
10	80	0.6	0.001	15.0	9.7	4.6	2.6	8.7	6.0	7.4	4.2	9.3	5.9
			0.01	14.4	9.1	6.0	3.6	9.4	6.0	7.8	4.9	8.8	5.8
			0.1	13.2	9.3	8.2	5.0	10.3	7.4	9.1	6.3	9.4	6.4
		0.8	0.001	16.4	12.6	5.1	2.9	10.1	6.0	7.4	4.5	11.8	8.2
			0.01	15.8	10.1	5.8	4.0	10.0	6.9	8.3	5.9	10.3	7.4
			0.1	12.5	8.2	6.4	4.2	9.4	5.9	7.6	4.9	8.0	4.9
	450	0.6	0.001	15.9	11.4	5.9	3.9	11.0	7.3	8.7	4.8	11.0	6.7
			0.01	11.8	7.8	6.4	3.1	8.5	5.4	7.4	4.1	7.7	4.6
			0.1	10.7	6.6	5.4	2.9	7.9	4.2	6.4	3.6	6.3	3.6
		0.8	0.001	16.2	10.6	4.2	1.8	9.4	5.2	7.0	3.9	10.2	6.2
			0.01	13.9	9.9	7.8	4.2	10.3	6.8	9.2	5.6	9.9	5.9
			0.1	12.2	8.5	6.6	3.4	9.2	5.9	8.1	4.3	8.0	4.2
20	80	0.6	0.001	11.4	9.2	5.7	4.4	8.5	6.6	7.3	5.8	9.0	6.7
			0.01	9.5	7.0	5.4	4.6	7.3	5.3	6.0	5.0	6.5	5.3
			0.1	7.0	5.3	4.3	3.3	6.0	4.1	5.0	3.5	4.9	3.8
		0.8	0.001	12.4	10.4	6.0	4.0	9.9	7.5	8.5	6.3	10.0	8.1
			0.01	10.4	8.7	6.6	5.3	8.7	6.6	7.6	5.8	8.4	6.5
			0.1	9.1	7.7	6.6	5.4	7.8	6.3	7.4	5.9	7.4	5.7
	450	0.6	0.001	9.4	7.2	5.5	3.8	7.0	5.4	6.1	4.5	6.7	5.3
			0.01	5.9	4.9	4.6	2.9	5.2	3.9	4.9	3.3	4.8	3.3
			0.1	9.1	7.2	5.8	3.6	7.8	5.2	7.1	4.6	7.0	4.5
		0.8	0.001	8.8	6.9	4.7	3.8	6.5	5.1	5.7	4.1	6.7	4.8
			0.01	9.8	7.2	6.4	4.6	7.8	5.9	7.0	5.1	6.8	5.4
			0.1	7.3	5.5	4.9	3.6	6.1	4.5	5.3	3.8	5.4	3.9
30	80	0.6	0.001	9.1	7.7	6.0	5.5	7.4	6.3	6.5	6.1	7.6	6.6
			0.01	9.9	8.6	7.6	6.6	8.7	7.5	8.0	7.2	8.3	7.4
			0.1	7.4	6.1	5.4	5.0	6.3	5.3	6.0	5.0	5.8	5.0
		0.8	0.001	10.7	9.7	7.1	6.2	8.8	7.7	7.7	6.8	9.3	8.4
			0.01	8.0	6.9	5.8	4.6	7.2	5.8	6.4	5.1	6.7	5.4
			0.1	6.6	5.6	5.0	4.6	5.7	4.9	5.4	4.8	5.4	4.8

	450	0.6	0.001	7.1	6.2	5.4	4.5	6.3	5.5	5.7	4.9	6.0	5.2
			0.01	9.0	7.5	7.0	6.1	8.0	6.8	7.2	6.4	7.4	6.7
			0.1	7.1	6.5	6.2	4.8	6.5	5.7	6.3	5.2	6.3	5.0
		0.8	0.001	7.8	7.1	5.6	4.8	6.9	5.9	6.2	5.2	6.8	5.9
			0.01	6.6	5.1	4.7	3.8	5.3	4.4	5.0	4.0	5.1	4.3
			0.1	7.2	5.5	4.8	4.1	5.9	4.7	5.2	4.3	5.2	4.1
40	80	0.6	0.001	8.3	7.7	6.4	5.4	7.7	6.8	7.4	6.2	7.6	6.7
			0.01	7.0	6.0	5.4	5.2	6.1	5.3	5.6	5.2	5.7	5.2
			0.1	6.2	5.3	4.8	4.3	5.4	4.8	5.0	4.6	5.1	4.5
		0.8	0.001	8.1	7.4	5.9	4.7	7.2	6.3	6.3	5.6	7.1	6.7
			0.01	8.7	7.9	7.2	5.6	7.8	7.2	7.5	6.5	7.7	7.0
			0.1	5.5	5.0	4.5	3.9	5.1	4.2	4.8	4.1	5.0	4.1
	450	0.6	0.001	7.3	6.2	5.5	4.8	6.3	5.7	5.9	5.3	6.1	5.6
			0.01	5.8	5.4	5.0	4.6	5.4	4.8	5.3	4.7	5.3	4.7
			0.1	5.2	4.7	4.1	3.6	4.8	4.0	4.5	3.8	4.5	3.7
		0.8	0.001	8.0	6.5	5.4	4.4	6.4	5.6	5.8	4.9	6.3	5.5
			0.01	5.7	5.0	4.3	3.9	5.1	4.3	4.7	4.1	4.7	4.1
			0.1	6.9	6.0	5.2	4.7	6.2	5.2	5.6	5.2	5.8	5.1

Table F3: Relative errors (RE) (%) and coverage rates (CR)(%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an independent matrix and a Student distribution of the Wald test statistic for $\beta_G = 0$.

K	m	CV	τ	V_{MD}		V_{KC}		V_{FG}		V_{MBN}		*Permutation test*
				RE	*CR*	*RE*	*CR*	*RE*	*CR*	*RE*	*CR*	*CR*
10	80	0.6	0.001	16.8	96.2	-0.5	94.3	7.0	95.1	3.3	96.0	95.3
			0.01	10.4	96.1	-6.7	93.3	1.0	94.6	-6.1	94.5	95.5
			0.1	15.8	94.4	-3.3	91.8	5.1	93.1	-4.2	92.0	94.6
		0.8	0.001	21.7	96.2	-0.3	93.3	9.0	94.7	1.9	96.6	95.6
			0.01	19.2	96.0	-3.7	92.9	6.0	94.2	-5.8	93.8	95.3
			0.1	17.8	95.5	-5.2	91.9	6.8	93.6	-8.1	91.7	96.0
	450	0.6	0.001	17.0	96.1	-1.2	94.1	6.6	95.2	0.0	95.5	96.0
			0.01	14.8	96.3	-3.3	93.7	5.3	95.5	-3.5	94.0	96.0
			0.1	16.5	95.8	-2.4	93.0	6.2	94.3	-3.1	92.9	95.9
		0.8	0.001	21.3	96.2	-2.1	94.4	7.5	95.7	-3.6	95.0	95.3
			0.01	10.9	95.3	-11.6	89.5	-1.5	92.1	-14.9	88.7	94.6
			0.1	17.7	95.9	-5.7	92.3	5.6	94.2	-9.1	91.0	95.6
20	80	0.6	0.001	7.2	96.3	-0.6	95.6	3.0	96.2	0.5	96.2	95.8
			0.01	4.2	95.2	-4.1	94.0	-0.1	94.9	-4.3	94.3	94.9
			0.1	10.6	95.2	1.4	93.8	5.9	94.4	0.8	93.8	95.1
		0.8	0.001	9.1	95.9	-0.5	94.3	3.9	95.1	-0.7	94.9	94.8
			0.01	9.9	95.2	-0.7	93.3	4.6	94.5	-2.5	93.0	94.8
			0.1	6.0	94.5	-5.1	92.9	-0.2	93.9	-7.6	92.2	95.2
	450	0.6	0.001	9.8	96.3	1.5	94.8	5.4	95.3	1.8	95.0	95.5
			0.01	9.2	96.4	0.3	95.5	4.5	95.8	-0.3	95.4	96.3
			0.1	4.7	95.0	-4.0	92.9	0.1	93.9	-4.7	93.1	93.9
		0.8	0.001	15.4	95.9	4.0	94.8	9.2	95.3	2.1	94.6	96.1
			0.01	6.7	95.5	-4.7	93.2	0.8	94.1	-7.3	92.9	94.5
			0.1	10.0	96.0	-1.6	93.5	3.7	94.6	-4.3	93.1	96.1
30	80	0.6	0.001	0.6	94.9	-4.2	94.6	-1.9	94.7	-3.5	94.7	94.5
			0.01	-2.2	93.4	-7.2	91.2	-4.8	92.3	-7.3	91.4	92.6
			0.1	3.3	94.8	-2.5	93.7	0.4	94.2	-2.9	93.7	94.6
		0.8	0.001	3.7	94.6	-2.3	93.7	0.5	94.1	-2.5	93.8	94.1
			0.01	2.7	95.3	-4.0	94.1	-0.7	94.6	-5.3	94.0	95.4
			0.1	9.6	95.7	1.7	94.4	5.6	95.3	-0.4	94.0	95.7

	450	0.6	0.001	4.2	95.8	-1.0	94.6	1.5	95.2	-0.9	94.8	
			0.01	-1.3	93.8	-6.7	92.2	-4.1	93.1	-7.0	92.3	
			0.1	3.5	94.9	-2.3	93.1	0.5	94.2	-2.9	93.0	
		0.8	0.001	4.7	95.6	-1.9	94.3	1.2	95.0	-3.1	94.0	
			0.01	7.1	95.4	-0.4	93.7	3.2	94.4	-2.3	93.3	
			0.1	4.7	94.9	-2.8	93.0	0.7	93.7	-4.7	92.6	
40	80	0.6	0.001	1.3	94.9	-2.3	94.2	-0.5	94.4	-1.8	94.2	94.6
			0.01	3.4	95.7	-0.7	95.0	1.3	95.4	-0.9	94.9	95.5
			0.1	6.2	95.1	1.8	94.7	3.9	94.9	1.4	94.7	94.8
		0.8	0.001	2.4	95.0	-2.0	93.7	0.1	94.3	-2.3	93.8	94.8
			0.01	4.1	94.0	-1.2	92.7	1.5	93.5	-2.3	92.5	93.9
			0.1	5.3	95.1	-0.5	94.1	2.2	94.7	-2.2	93.5	95.3
	450	0.6	0.001	4.6	95.0	0.7	94.4	2.6	94.9	0.7	94.4	94.5
			0.01	4.3	95.3	0.0	95.0	2.1	95.0	-0.4	95.0	
			0.1	6.3	96.0	1.8	95.2	4.0	95.7	1.3	95.0	
		0.8	0.001	3.3	95.6	-1.7	94.4	0.7	95.2	-2.7	94.1	
			0.01	4.9	95.2	-0.7	93.4	2.1	94.3	-2.1	93.3	
			0.1	1.8	94.6	-3.7	93.6	-0.9	93.8	-5.2	93.1	

Table F4: Relative errors (RE) (%) and coverage rates (CR)(%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an exchangeable matrix and a Student distribution of the Wald test statistic for $\beta_G = 0$.

K	m	CV	τ	V_{MD}		V_{KC}		V_{FG}		V_{MBN}	
				RE	***CR***	***RE***	***CR***	***RE***	***CR***	***RE***	***CR***
10	80	0.6	0.001	18.8	97.4	-0.5	94.0	8.5	95.8	0.5	94.1
			0.01	7.2	96.4	-6.7	94.0	-0.4	95.1	-3.5	94.2
			0.1	9.1	95.0	-2.6	92.6	2.8	93.7	2.8	93.6
		0.8	0.001	28.6	97.1	1.2	94.0	13.5	95.5	-0.3	91.8
			0.01	12.9	96.0	-4.6	93.1	2.5	94.1	-2.8	92.6
			0.1	11.8	95.8	-0.5	94.1	5.1	95.1	5.0	95.1
	450	0.6	0.001	13.5	96.1	-2.9	92.7	4.4	95.2	-0.7	93.3
			0.01	11.5	96.9	-0.6	94.6	4.9	95.9	4.6	95.4
			0.1	14.2	97.1	2.1	95.8	7.6	96.4	8.2	96.4
		0.8	0.001	22.4	98.2	0.4	94.8	8.3	96.1	0.8	93.8
			0.01	7.0	95.8	-5.4	93.2	0.0	94.4	-0.7	94.1
			0.1	11.3	96.6	-0.5	94.1	5.0	95.7	5.5	95.8
20	80	0.6	0.001	3.8	95.6	-4.6	93.4	-0.5	94.2	-4.3	93.3
			0.01	1.6	95.4	-4.4	94.7	-1.5	95.0	-2.5	94.7
			0.1	10.0	96.7	4.3	95.9	7.1	96.5	7.3	96.2
		0.8	0.001	7.3	96.0	-3.0	92.5	2.0	93.7	-3.7	91.9
			0.01	4.4	94.7	-2.4	93.4	0.9	94.2	-1.0	93.5
			0.1	2.4	94.6	-3.0	93.7	-0.4	94.1	-0.2	94.3
	450	0.6	0.001	5.8	96.2	-1.0	94.6	2.3	95.5	0.4	94.7
			0.01	9.1	97.1	3.3	96.1	6.1	96.7	6.2	96.7
			0.1	5.0	96.4	-0.4	94.8	2.2	95.4	2.5	95.5
		0.8	0.001	10.8	96.2	2.7	94.9	6.4	95.9	3.5	95.2
			0.01	1.3	95.4	-4.2	94.1	-1.6	94.9	-1.6	94.6
			0.1	5.9	96.4	0.4	95.5	3.0	96.2	3.4	96.1
30	80	0.6	0.001	-0.6	94.5	-5.8	93.7	-3.3	93.9	-5.6	93.4
			0.01	-2.3	93.4	-6.1	92.5	-4.3	92.8	-4.8	92.6
			0.1	2.1	95.0	-1.4	94.7	0.3	95.0	0.5	95.0
		0.8	0.001	1.2	93.8	-5.2	92.3	-1.9	93.2	-5.8	91.6
			0.01	0.9	95.4	-3.3	94.2	-1.3	94.9	-2.3	94.6
			0.1	6.1	95.4	2.4	95.1	4.2	95.2	4.3	95.2

	450	0.6	0.001	1.9	95.5	-2.3	94.5	-0.3	95.1	-1.2	94.8
			0.01	-1.4	93.9	-4.8	93.2	-3.2	93.6	-3.1	93.3
			0.1	2.1	95.2	-1.4	94.3	0.3	94.8	0.6	95.0
		0.8	0.001	1.7	95.2	-3.0	94.1	-0.8	94.8	-2.4	94.1
			0.01	5.8	96.2	2.1	95.6	3.9	96.0	3.9	95.7
			0.1	5.4	95.9	1.9	95.3	3.6	95.7	3.9	95.9
40	80	0.6	0.001	-0.2	94.6	-4.0	93.2	-2.1	93.8	-3.8	93.3
			0.01	0.2	94.8	-2.6	94.7	-1.3	94.8	-1.6	94.8
			0.1	3.3	95.7	0.7	95.2	2.0	95.4	2.1	95.5
		0.8	0.001	1.3	95.3	-3.2	93.7	-1.0	94.4	-3.5	93.3
			0.01	1.0	94.4	-2.1	92.8	-0.6	93.5	-1.3	93.0
			0.1	1.1	96.1	-1.5	95.8	-0.3	95.9	-0.1	95.9
	450	0.6	0.001	3.0	95.2	-0.1	94.3	1.4	94.7	0.8	94.4
			0.01	3.1	95.4	0.4	95.2	1.7	95.3	1.8	95.3
			0.1	3.6	96.4	1.0	96.0	2.3	96.2	2.5	96.3
		0.8	0.001	0.8	95.6	-2.6	94.4	-1.0	95.1	-2.0	94.5
			0.01	5.1	96.1	2.3	95.7	3.6	95.9	3.6	95.9
			0.1	-0.6	95.3	-3.1	94.8	-1.9	94.8	-1.7	94.9

Table F5: Power (%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an independent matrix and a Student distribution of the Wald test statistic.

K	m	CV	τ	V_{MD}	V_{KC}	V_{FG}	V_{MBN}	*Permutation test*
10	80	0.6	0.001	49.8	61.3	56.3	58.6	56.4
			0.01	49.3	61.4	55.7	59.4	54.1
			0.1	55.3	64.9	61.3	65.8	57.5
		0.8	0.001	47.2	61.6	55.6	58.6	50.6
			0.01	49.1	63.1	57.6	64.7	53.7
			0.1	49.1	64.3	57.7	66.0	52.5
	450	0.6	0.001	52.4	63.9	59.1	62.5	55.6
			0.01	50.8	62.3	57.3	61.9	55.0
			0.1	53.4	64.5	60.5	64.5	58.6
		0.8	0.001	49.4	63.4	57.0	63.3	53.1
			0.01	48.8	63.2	56.2	64.9	53.8
			0.1	50.6	63.5	58.5	64.9	53.7
20	80	0.6	0.001	66.7	72.8	70.2	71.8	70.3
			0.01	66.0	71.8	69.0	71.5	68.6
			0.1	65.8	70.5	67.9	70.6	67.9
		0.8	0.001	65.4	72.4	69.5	71.9	68.2
			0.01	63.6	70.3	67.1	70.8	65.8
			0.1	61.8	69.6	66.0	71.9	62.8
	450	0.6	0.001	65.5	70.8	68.9	70.7	69.8
			0.01	68.3	74.4	71.3	74.6	70.7
			0.1	67.0	72.4	69.6	73.2	69.1
		0.8	0.001	63.5	69.9	67.2	70.7	65.4
			0.01	65.4	71.4	68.9	72.7	67.6
			0.1	65.6	71.2	69.3	73.1	68.2
30	80	0.6	0.001	72.3	75.6	74.4	75.1	74.3
			0.01	70.2	73.7	72.3	73.6	71.9
			0.1	73.8	78.3	76.1	78.7	75.5
		0.8	0.001	72.4	76.3	75.1	76.8	73.9
			0.01	72.6	76.2	74.9	77.2	74
			0.1	71.0	75.4	73.9	76.4	72.2
	450	0.6	0.001	70.7	73.6	72.1	73.9	
			0.01	70.9	74.3	72.4	74.8	

			0.1	70.7	73.3	72.1	73.9	
		0.8	0.001	73.3	76.7	74.6	77.4	
			0.01	71.7	77.4	74.8	78.2	
			0.1	68.4	73.8	71.7	75.3	
40	80	0.6	0.001	74.2	76.6	75.5	76.4	75.3
			0.01	74.8	77.4	75.6	77.0	75.8
			0.1	73.7	75.9	75.1	75.9	74.0
		0.8	0.001	74.1	77.2	75.6	77.3	75.4
			0.01	76.2	78.3	77.5	78.8	76.3
			0.1	71.8	75.1	74.4	75.8	71.4
	450	0.6	0.001	72.3	75.4	73.6	75.4	73.7
			0.01	73.8	76.6	75.2	76.8	74.3
			0.1	72.4	74.1	73.5	74.5	72.8
		0.8	0.001	72.0	75.2	73.3	75.9	
			0.01	67.8	71.9	69.9	73.0	
			0.1	71.6	75.1	73.4	76.6	

Note: The intervention effect, β_G, have been selected by simulation to achieve 80% power in each scenario.

Table F6: Power (%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an exchangeable matrix and a Student distribution of the Wald test statistic.

K	m	CV	τ	V_{MD}	V_{KC}	V_{FG}	V_{MBN}
10	80	0.6	0.001	42.6	59.1	52.1	58.8
			0.01	49.4	62.3	55.9	58.3
			0.1	60.1	67.9	65.1	64.4
		0.8	0.001	39.4	58.1	49.8	59.1
			0.01	48.8	62.9	57.3	61.8
			0.1	58.8	68.3	64.9	64.4
	450	0.6	0.001	52.5	63.6	58.7	62.1
			0.01	56.7	64.7	60.9	61.2
			0.1	61.5	68.9	66.5	65.3
		0.8	0.001	44.2	59.3	53.0	58.6
			0.01	56.4	66.4	61.7	61.9
			0.1	61.3	68.8	66.7	65.4
20	80	0.6	0.001	66.1	72.6	69.5	72.5
			0.01	70.4	75.2	72.9	73.6
			0.1	73.9	77.6	75.8	75.3
		0.8	0.001	65.0	72.4	70.0	72.7
			0.01	68.6	72.4	70.5	71.0
			0.1	76.7	79.5	78.3	78.3
	450	0.6	0.001	69.0	72.5	70.4	71.0
			0.01	78.7	82.2	80.4	80.6
			0.1	80.4	82.5	81.5	81.2
		0.8	0.001	66.7	71.7	69.2	71.4
			0.01	78.3	81.5	80.2	80.2
			0.1	82.3	84.6	83.4	83.2
30	80	0.6	0.001	71.5	75.3	73.5	75.1
			0.01	74.3	76.8	75.6	75.7
			0.1	81.5	83.2	82.5	82.4
		0.8	0.001	71.7	76.2	74.4	76.0
			0.01	77.0	79.7	78.2	78.8
			0.1	84.9	85.8	85.6	85.4
	450	0.6	0.001	74.0	76.7	75.3	76.1
			0.01	79.7	82.1	81.3	80.9

			0.1	82.2	83.5	82.7	82.7
		0.8	0.001	76.8	80.5	78.3	80.1
			0.01	82.6	83.9	83.2	83.3
			0.1	87.6	89.1	88.5	88.3
40	80	0.6	0.001	74.7	77.7	76.0	76.9
			0.01	77.4	79.4	78.5	78.7
			0.1	83.8	85.1	84.7	84.4
		0.8	0.001	75.5	78.9	77.1	78.8
			0.01	81.5	82.5	82.1	82.2
			0.1	85.1	86.2	85.6	85.7
	450	0.6	0.001	74.3	76.9	75.6	76.2
			0.01	83.2	84.1	83.4	83.4
			0.1	83.0	84.1	83.8	83.8
		0.8	0.001	76.9	78.7	77.8	78.2
			0.01	83.9	84.5	84.1	84.3
			0.1	88.1	89.2	88.6	88.5

Note: The intervention effect, β_G, have been selected by simulation to achieve 80% power in each scenario

Table F7: Relative errors (RE) (%) and coverage rates (CR)(%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an independent matrix and a Student distribution of the Wald test statistic for $\beta_G \neq 0$.

K	m	CV	τ	V_{MD}		V_{KC}		V_{FG}		V_{MBN}		*Permutation test*
				RE	*CR*	*RE*	*CR*	*RE*	*CR*	*RE*	*CR*	*CR*
10	80	0.6	0.001	19.4	96.8	1.8	94.5	9.9	95.8	5.8	96.9	95.7
			0.01	13.4	95.6	-4.2	93.4	3.7	94.6	-3.6	93.9	94.9
			0.1	16.1	96.0	-2.6	92.8	4.1	93.8	-3.3	93.0	95.4
		0.8	0.001	19.6	96.3	-2.1	93.6	6.9	95.3	-0.4	95.9	95.6
			0.01	18.2	95.6	-4.0	92.5	6.4	93.5	-5.6	93.5	95.9
			0.1	20.8	96.2	-3.5	92.9	5.3	94.4	-7.0	92.4	96.2
	450	0.6	0.001	13.6	95.6	-3.6	92.5	4.1	93.9	-2.1	94.3	95.1
			0.01	13.5	95.4	-4.8	93.3	3.3	94.4	-5.2	93.3	95.3
			0.1	10.4	95.4	-7.5	92.4	-0.9	93.9	-8.1	92.4	94.7
		0.8	0.001	22.6	95.7	-0.5	92.9	9.7	94.1	-1.3	93.9	95.0
			0.01	13.3	94.3	-8.8	91.5	1.6	93.2	-11.8	90.8	94.9
			0.1	15.5	94.8	-7.5	91.1	0.2	92.2	-10.8	90.4	94.8
20	80	0.6	0.001	9.2	95.7	1.3	94.8	5.1	95.1	2.5	95.4	95.4
			0.01	2.8	95.0	-5.2	93.5	-1.5	94.2	-5.3	93.4	94.2
			0.1	7.4	95.1	-1.4	93.9	2.2	94.2	-2.0	93.9	95.2
		0.8	0.001	9.8	95.7	0.0	93.6	4.5	94.8	-0.2	94.3	94.6
			0.01	6.1	94.8	-4.3	93.3	0.5	93.9	-6.0	93.5	94.5
			0.1	7.1	94.5	-4.1	92.3	0.3	93.5	-6.7	92.1	95.3
	450	0.6	0.001	10.7	95.9	2.3	94.2	6.3	94.9	2.6	94.6	94.9
			0.01	9.0	96.1	0.0	94.5	4.2	95.2	-0.6	94.6	95.7
			0.1	9.8	96.3	0.5	95.1	4.3	95.9	-0.3	95.3	95.9
		0.8	0.001	8.5	95.4	-1.7	94.6	3.1	94.9	-3.2	94.2	95.5
			0.01	4.8	93.7	-5.7	92.6	-0.7	93.0	-7.8	92.3	94.2
			0.1	10.6	95.3	-1.4	93.5	3.2	94.4	-4.3	93.3	94.8
30	80	0.6	0.001	7.7	95.8	2.6	94.7	5.0	95.3	3.2	95.1	95.2
			0.01	0.3	94.4	-5.0	93.4	-2.4	93.9	-5.1	93.4	93.7
			0.1	7.2	95.5	1.3	94.4	3.8	94.9	0.8	94.4	94.9
		0.8	0.001	3.5	95.0	-2.6	94.1	0.3	94.6	-2.8	94.4	94.4
			0.01	6.1	95.4	-0.9	94.2	2.5	94.9	-2.2	94.1	95.2
			0.1	3.6	94.8	-3.7	93.5	-0.7	94.2	-5.5	92.9	95.1

	450	0.6	0.001	2.6	94.8	-2.6	93.1	0.0	94.0	-2.6	93.6	
			0.01	3.8	94.6	-2.0	93.5	0.8	94.5	-2.5	93.5	
			0.1	5.3	94.6	-0.6	93.9	1.8	94.3	-1.1	93.9	
		0.8	0.001	11.7	96.1	4.7	94.7	8.1	95.6	3.6	94.7	
			0.01	3.9	95.2	-3.5	93.5	0.1	94.3	-5.3	93.1	
			0.1	7.1	95.6	-0.6	95.0	2.6	95.5	-2.7	94.7	
40	80	0.6	0.001	4.3	94.5	0.6	93.5	2.4	93.8	1.1	93.7	94.1
			0.01	2.6	94.7	-1.4	93.7	0.6	94.2	-1.5	93.9	94.8
			0.1	-1.0	93.8	-5.2	92.9	-3.4	93.3	-5.6	92.9	93.6
		0.8	0.001	4.6	96.7	0.0	95.8	2.2	96.6	-0.2	95.9	95.9
			0.01	8.6	96.0	3.2	95.2	5.8	95.7	2.1	95.2	95.8
			0.1	4.2	94.7	-1.4	93.3	1.1	93.8	-2.9	93.1	94.7
	450	0.6	0.001	5.0	95.1	1.1	94.5	3.0	94.9	1.1	94.5	94.9
			0.01	1.7	94.6	-2.6	93.1	-0.5	94.0	-3.0	93.2	94.8
			0.1	5.4	95.7	1.0	94.5	2.9	94.9	0.5	94.4	95.6
		0.8	0.001	2.2	94.8	-2.6	93.4	-0.3	94.0	-3.5	93.2	
			0.01	2.1	94.0	-3.3	93.3	-0.7	93.8	-4.7	93.0	
			0.1	0.1	93.5	-5.5	92.5	-2.9	92.9	-7.1	92.3	

Note: The intervention effect, β_G, have been selected by simulation to achieve 80% power in each scenario.

Table F8: Relative errors (RE) (%) and coverage rates (CR)(%) under different combinations of number of clusters (K = 10, 20, 30, 40), mean cluster size (m = 80, 450), coefficient of variation of the cluster size (CV = 0.6, 0.8) and Kendall's tau (τ = 0.001, 0.01, 0.1) for pseudo-values regression with an exchangeable matrix and a Student distribution of the Wald test statistic for $\beta_G \neq 0$.

K	m	CV	τ	V_{MD}		V_{KC}		V_{FG}		V_{MBN}	
				RE	***CR***	***RE***	***CR***	***RE***	***CR***	***RE***	***CR***
10	80	0.6	0.001	25.0	98.0	3.7	94.9	15.0	96.4	4.4	94.3
			0.01	11.6	97.1	-3.0	94.3	3.6	95.6	0.2	94.3
			0.1	14.0	95.8	1.7	93.7	6.2	94.8	7.0	94.7
		0.8	0.001	23.8	96.8	-2.1	94.0	9.0	95.6	-3.8	93.1
			0.01	13.6	96.1	-4.3	91.6	3.0	92.9	-2.7	91.7
			0.1	14.5	97.2	1.7	95.3	6.2	95.6	7.1	95.7
	450	0.6	0.001	11.7	95.7	-4.1	92.4	3.3	94.1	-1.9	92.5
			0.01	10.1	96.1	-2.0	94.4	3.4	95.4	3.1	95.1
			0.1	8.6	96.1	-2.8	94.6	1.3	95.2	2.7	95.6
		0.8	0.001	22.7	97.1	1.4	93.0	10.4	94.9	2.2	91.7
			0.01	6.2	94.6	-6.2	92.8	-0.7	93.8	-1.6	93.7
			0.1	10.0	95.9	-1.7	94.4	2.4	94.8	4.4	94.8
20	80	0.6	0.001	10.4	95.7	1.5	93.8	5.8	94.8	1.8	93.7
			0.01	0.9	94.5	-5.1	93.1	-2.3	93.8	-3.3	93.3
			0.1	6.3	95.5	0.7	94.3	3.0	94.8	3.6	94.8
		0.8	0.001	11.8	96.2	0.4	94.1	6.2	95.3	-0.7	93.5
			0.01	-0.3	94.5	-6.9	93.0	-3.8	93.6	-5.7	93.2
			0.1	2.5	95.7	-3.0	94.4	-0.8	94.8	-0.3	94.9
	450	0.6	0.001	6.3	95.4	-0.4	93.2	2.8	94.3	1.1	93.6
			0.01	10.0	96.4	4.3	95.8	7.0	96.2	7.1	96.1
			0.1	7.0	96.0	1.5	95.0	3.8	95.7	4.6	95.6
		0.8	0.001	4.7	95.3	-3.1	93.9	0.7	94.7	-2.4	93.6
			0.01	5.7	95.3	-0.1	94.0	2.7	94.7	2.5	94.6
			0.1	4.4	95.0	-1.0	94.4	1.2	94.7	2.1	94.8
30	80	0.6	0.001	6.2	95.2	0.8	94.1	3.4	94.3	1.0	93.9
			0.01	-1.3	94.3	-5.1	93.1	-3.3	93.5	-3.8	93.3
			0.1	3.1	96.0	-0.5	95.2	1.1	95.6	1.4	95.6
		0.8	0.001	1.7	94.7	-4.4	93.5	-1.6	94.2	-4.9	92.7
			0.01	2.9	95.6	-1.5	95.0	0.7	95.5	-0.5	95.0
			0.1	4.0	95.6	0.3	94.8	1.9	95.2	2.2	95.1

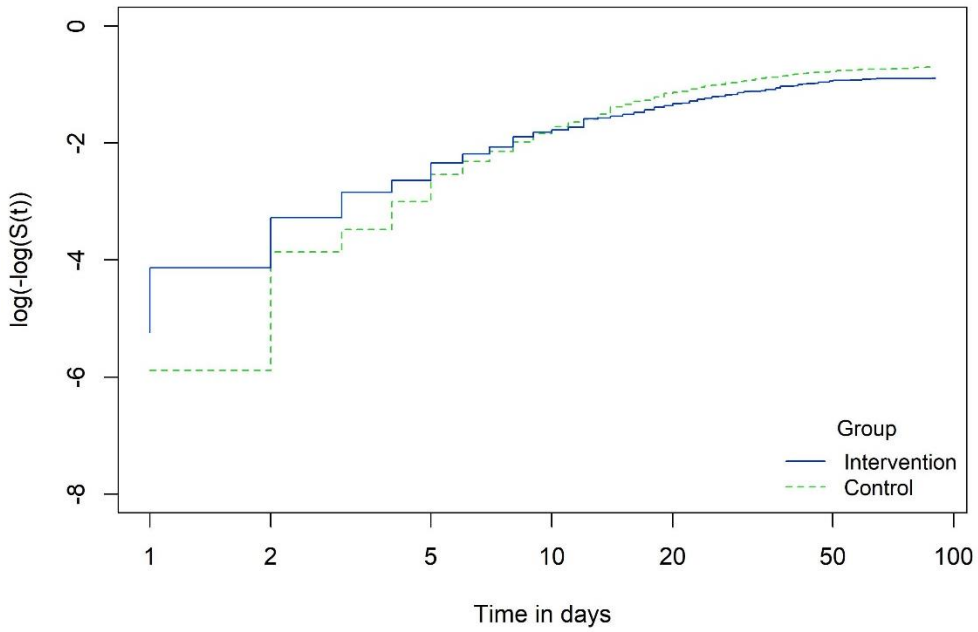
	450	0.6	0.001	0.6	94.9	-3.5	93.2	-1.5	94.1	-2.4	93.3
			0.01	2.2	94.8	-1.4	94.4	0.3	94.7	0.4	94.5
			0.1	2.9	96.4	-0.6	95.4	0.9	95.9	1.3	96.1
		0.8	0.001	9.3	96.1	4.2	94.6	6.8	95.2	4.8	94.6
			0.01	3.2	95.6	-0.5	95.0	1.3	95.3	1.2	95.2
			0.1	4.0	95.4	0.4	94.4	1.9	94.7	2.4	94.7
40	80	0.6	0.001	2.8	94.9	-0.9	93.7	0.9	94.3	-0.6	93.8
			0.01	0.2	94.4	-2.6	93.6	-1.2	94.0	-1.6	93.9
			0.1	1.1	94.8	-1.5	94.1	-0.4	94.4	-0.2	94.6
		0.8	0.001	3.3	96.0	-1.2	94.6	1.1	95.3	-1.5	94.5
			0.01	6.2	96.1	2.9	95.1	4.5	95.7	3.8	95.6
			0.1	1.6	94.8	-1.1	94.4	0.1	94.6	0.3	94.6
	450	0.6	0.001	3.6	95.2	0.6	94.7	2.1	95.0	1.4	94.9
			0.01	-0.7	95.1	-3.2	94.6	-2.0	94.8	-1.9	94.8
			0.1	1.8	96.1	-0.8	95.5	0.4	95.8	0.6	95.9
		0.8	0.001	1.7	94.2	-1.8	93.3	-0.1	93.7	-1.2	93.4
			0.01	2.5	94.8	-0.2	94.3	1.1	94.8	1.1	94.7
			0.1	-0.1	94.5	-2.7	93.6	-1.5	94.4	-1.2	94.4

Note: The intervention effect, β_G, have been selected by simulation to achieve 80% power in each scenario.

Web Appendix G: Checking of the proportional hazards assumption and R code for the illustrative example DEMETER

G.1. Checking of the proportional hazards assumption

Figure G1: The $log(-log(S(t)))$ curves



G.2. R code to estimate the difference in RMST

The R code used to compute the difference in RMST for the death at 90 days in the DEMETER trial and its confidence intervals with the four correction methods and the permutation test in Table 2 is detailed in the following section.

We consider a dataset, denoted *dataset*, with the following variables: the observed survival time (denoted *time)*, the event indicator = 1 if the observed time corresponds to the time between the origin and the occurrence of the survival event and 0 otherwise (denoted *indicator*), the trial group = 0 for the control group and 1 for the intervention group (denoted *group*) and the cluster ID (denoted *ID)*. The horizon time t^* is fixed to 90. The *GEE.var.kc.approx*, *GEE.var.fg.cor* and *GEE.var.fg.cor* functions and the R code for the permutation-based confidence interval are available on https://github.com/Le-Vilain-Abraham/RMST-estimation-with-pseudo-values-regression-in-small-sample-CRTs.

The "pseudo", "gee", "geesmv" and "matrixcalc" packages are needed to compute the code.

G.2.1. Estimation of the difference in RMST

```
#Computation of the pseudo-values
dataset <- dataset[order(dataset[,"ID"], decreasing = T), ]
dataset$pv <- pseudomean(dataset$time,
                         dataset$indicator,
                         tmax = 90))

#Regression with an exchangeable matrix
fit_standard <- summary(gee(pv ~ group,
                data = dataset,
                id = ID,
                family = gaussian,
                corstr = "exchangeable"))

#Difference in RMST
delta.rmst <- fit_standard$coefficients["group", "Estimate"]
```

G.2.2. Confidence interval with the MD estimator

```
fit_md <- GEE.var.md(pv ~ group,
                     data = dataset,
                     id = ID,
                     family = gaussian,
                     corstr = " exchangeable")

# Estimation of the variance
var.md <- fit_md$cov.beta["group"]

# Estimation of the confidence interval
ci.low <- delta.rmst - qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.md)
ci.up <- delta.rmst + qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.md)
```

G.2.3. Confidence interval with the KC estimator

```
fit_kc <- GEE.var.kc.approx(pv ~ group,
                     data = dataset,
                     id = ID,
                     family = gaussian,
                     corstr = " exchangeable")

# Estimation of the variance
var.kc <- fit_kc$cov.beta["group"]

# Estimation of the confidence interval
ci.low <- delta.rmst - qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.kc)
ci.up <- delta.rmst + qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.kc)
```

G.2.4. Confidence interval with the FG estimator

```
fit_fg <- GEE.var.fg.cor(pv ~ group,
                         data = dataset,
```

```
                    id = ID,
                    family = gaussian,
                    corstr = " exchangeable")

# Estimation of the variance
var.fg <- fit_fg$cov.beta["group"]

# Estimation of the confidence interval
ci.low <- delta.rmst - qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.fg)
ci.up <- delta.rmst + qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.fg)
```

G.2.5. Confidence interval with the MBN estimator

```
fit_mbn <- GEE.var.mbn.corr(pv ~ group,
                    data = dataset,
                    id = ID,
                    family = gaussian,
                    corstr = " exchangeable")

# Estimation of the variance
var.mbn <- fit_mbn$cov.beta["group"]

# Estimation of the confidence interval
ci.low <- delta.rmst - qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.mbn)
ci.up <- delta.rmst + qt(0.975, df = length(unique(dataset$ID))-2)*sqrt(var.mbn)
```

G.2.6. Permutation-based confidence interval

```
dataset$arm <- dataset$group
dataset$cluster <- dataset$ID

dataset$arm.obs <- dataset$arm

ci <- ci(dataset, "independence")

results <- data.frame(matrix = "ind",
                ci_low =ci$ci_low,
                ci_up =ci$ci_up)
```

Chapitre IV

Travaux annexes liés aux ERC

En parallèle de mes travaux de thèse, j'ai également travaillé sur deux projets, ne portant pas sur les critères de jugement de type survie mais en lien avec les ERCs. Pour le premier projet, j'ai été en charge de l'analyse statistique d'un ERC. Pour le second projet, j'ai participé en tant que lectrice à l'extraction des données d'une revue méthodologique évaluant les difficultés de recrutement et d'implémentation de l'intervention dans les ERC en *stepped-wedge*. Ces deux travaux ont chacun conduit à la publication d'un article et sont présentés dans cette section.

1 Analyse de l'essai randomisé en clusters FADEMELA

1.1 Rationnel et objectif de l'étude

Le mélanome est un cancer fréquent dont l'incidence a triplé en 30 ans et qui était responsable de 3,5 décès sur 100.000 personnes en Europe en 2020 [99]. Un diagnostic tardif et une épaisseur de la tumeur plus grande sont associés à une mortalité plus élevée. Un examen médical cutané effectué par un dermatologue ou un médecin généraliste et l'auto-examen de la peau sont significativement associés à des mélanomes plus fins [100, 101, 102]. Des actions de prévention et de détection précoce dans la population générale sont efficaces mais coûteuses à grande échelle.

Le mélanome est plus fréquent (risque relatif entre 2 et 5) chez les apparentés au premier degré (frères, soeurs, parents, enfants) des patients ayant déjà eu un mélanome [103, 104, 105]. Le dépistage des personnes à risque est recommandé dans de nombreux pays, dont la France [106]. Par ailleurs, une étude américaine a montré qu'une consultation bisannuelle serait justifiée pour les apparentés au premier degré [107].

Dans cet objectif de développement de la prévention, des études pilotes ont été réalisées au CHU de Tours, en 2005 et 2012, afin d'évaluer l'utilité d'une fiche écrite remise aux patients atteints d'un mélanome, en plus des conseils oraux, à destination de leurs apparentés au premier degré, contenant des conseils de photoprotection et de détection précoce. Il a pu être observé que l'adhésion au dépistage était plus importante lorsque la fiche conseil avait été remise (87,5% versus 50,5% dans l'étude pilote de 2012). Cependant ces études pilotes n'étaient pas randomisées et reposaient sur les déclarations des patients sans contact avec les apparentés.

L'essai FADEMELA a donc été mis en place afin d'évaluer l'efficacité de la remise d'une fiche conseil associée à un conseil oral à des patients ayant un mélanome sur le comportement de détection précoce chez leur apparentés au premier degré.

1.2 Méthodologie de l'étude

1.2.1 *Schéma d'étude*

Cette étude était un essai de supériorité, multicentrique, randomisé en clusters en deux groupes parallèles. La justification du schéma en cluster reposait sur le souhait de limiter la contamination inter-groupes. L'unité de randomisation était le centre hospitalier. Les patients ayant un mélanome, nommés cas-index par la suite, ainsi que leurs apparentés au premier degré étaient tous inclus dans le même groupe. L'unité d'analyse était les apparentés au premier degré.

1.2.2 *Déroulement de l'étude*

L'étude s'est déroulée entre janvier 2014 et mai 2019.

Au cours d'une des premières visites de prise en charge initiale (exérèse ou reprise élargie, consultation d'annonce ou programme personnalisé de soins), les cas-index recevaient une information sur l'utilité de la détection précoce et la photoprotection de leurs apparentés au premier degré. Dans le groupe intervention, l'information était transmise de manière orale et écrite, à l'aide d'une fiche conseil, et dans le groupe contrôle uniquement oralement (intervention usuelle).

Un an après, pendant la visite de suivi, les cas-index étaient informés de l'étude. Leur non-opposition était recueillie ainsi que les coordonnées de leur apparentés au premier degré, pour les cas-index ne s'étant pas opposés à participer à l'étude. Des questionnaires étaient remplis par les cas-index pour évaluer, en particulier, si la fiche conseil avait été distribuée à leurs apparentés au premier degré. Les apparentés au premier degré étaient ensuite contactés par téléphone. Après avoir exprimé oralement leur non-opposition, les critères permettant d'évaluer la détection précoce et la photoprotection étaient recueillis.

1.2.3 *Critère de jugement principal*

Le critère de jugement principal était la consultation effective d'un dermatologue et/ou d'un médecin généraliste pour un examen de détection précoce d'un mélanome malin chez les apparentés au premier degré des cas-index, entre la première visite du cas-index (environ 1 an avant) et le jour de l'appel téléphonique.

1.2.4 *Nombre de sujets nécessaires*

Le nombre de sujets nécessaires a été estimé en considérant une puissance de 90% et un risque alpha de 5%. Dans le calcul, la corrélation au sein des familles, et non au sein des hôpitaux, a été prise en compte, car on pouvait s'attendre à ce que la corrélation intra-familiale soit plus importante. En utilisant les données de l'étude pilote, on a considéré un coefficient de corrélation intra-classe pour le dépistage au sein d'une famille de 0.40 et que 60% des apparentés au premier degré du groupe contrôle feraient une consultation de détection précoce. Afin de mettre en évidence une augmentation de 15% de la détection précoce du mélanome dans le groupe ayant reçu une information écrite en

plus de l'information orale, 900 apparentés au premier degré devaient être recrutés. En anticipant une moyenne de 4 apparentés au premier degré par cas-index, 225 cas-index devaient être inclus.

1.2.5 *Analyses statistiques*

Le critère de jugement principal a été décrit en termes de pourcentages de succès, le succès correspondant à la réalisation d'une consultation avec un examen de détection précoce du mélanome. Il a ensuite été analysé avec un modèle marginal GEE qui prenait en compte uniquement la corrélation intra-classe de la famille. En effet, cette dernière était supposée plus importante que la corrélation au sein des centres hospitaliers. Une matrice de corrélation échangeable et un estimateur de variance sandwich ont été utilisés. Les résultats étaient fournis sous forme d'odds-ratio avec leur intervalle de confiance à 95%.

Les CCI pour le critère de jugement principal ont été calculés pour chaque groupe avec la méthode d'analyse de la variance. Les CCI ont été calculés au niveau de la famille mais aussi du centre hospitalier afin de confirmer l'hypothèse selon laquelle la corrélation serait plus forte au sein de la famille que du centre hospitalier.

L'analyse a été réalisée selon le principe de l'intention de traiter.

Les analyses ont été réalisées sous SAS (SAS Institute Inc., Cary, NC, USA).

1.3 Résultats

Au total, 9 centres ont été randomisés dont 4 dans le groupe contrôle et 5 dans le groupe intervention. Au final, 60 cas-index avec 166 apparentés au premier degré ont été inclus dans le groupe intervention et 48 cas-index avec 114 apparentés au premier degré dans le groupe contrôle.

Dans le groupe intervention, 36,1% (60/166) des apparentés au premier degré ont rapporté avoir effectué une consultation avec un dermatologue et/ou d'un médecin généraliste contre 39,5% (45/114) des apparentés au premier degré dans le groupe contrôle (Figure IV.1). L'odds ratio entre les deux groupes était estimé à 0,9 (IC : [0,5 ;1,5], p = 0,63).

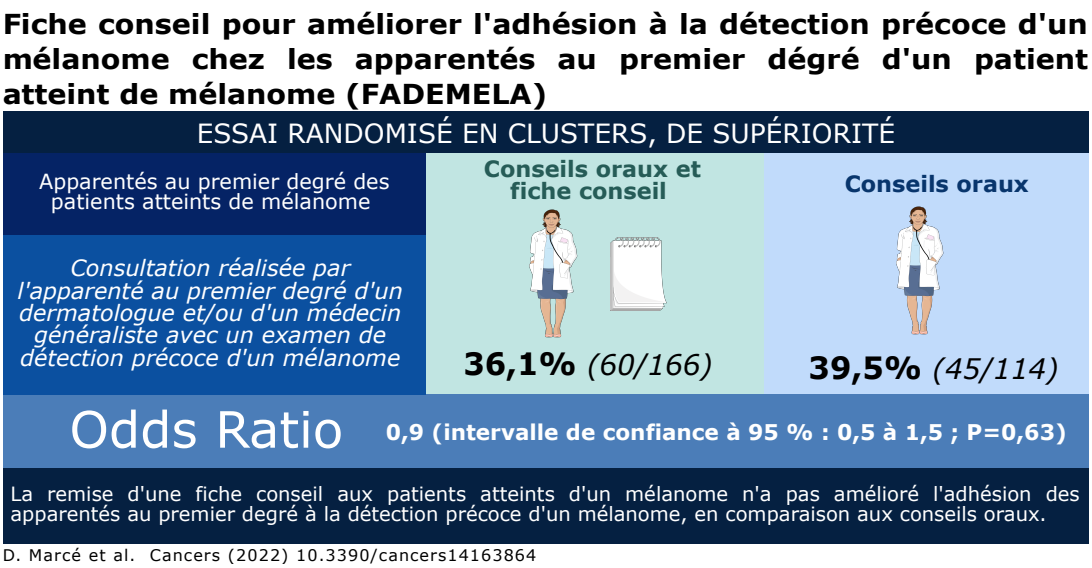


FIGURE IV.1 – Résumé de l'étude FADEMELA

Les CCI pour le critère de jugement principal dans les groupes d'intervention et de contrôle étaient de 0,009 et 0,06 au niveau du centre hospitalier et de 0,12 et 0,15 au niveau de la famille.

1.4 Conclusion

La remise d'une fiche conseil aux patients atteints d'un mélanome n'a pas amélioré significativement l'adhésion des apparentés au premier degré à la détection précoce d'un mélanome, en comparaison aux conseils oraux. A noter que la puissance était moindre qu'escomptée du fait de la difficulté de recrutement et donc de la non atteinte du nombre de sujets nécessaires.

Ce travail a donné lieu à un article publié en 2022 dans *Cancers* [108], disponible en Annexe B.

2 Difficultés de recrutement et d'implémentation de l'intervention dans les ERC en *"stepped-wedge"*

2.1 Rationnel et objectif

L'ERC en *stepped-wedge* (ERC-SW) est un schéma d'étude alternatif aux ERC en deux bras parallèles. Dans les ERC-SW, ce ne sont pas les clusters qui sont randomisés, mais le moment où l'intervention sera implémentée dans les clusters. Les clusters sont randomisés dans des séquences pour lesquelles la date de mise en place de l'intervention est différente. Par exemple, dans la Figure IV.2, l'intervention sera implémentée à partir de la période 2 dans le ou les clusters randomisés dans la séquence 1, alors que le ou les clusters de la séquence 4 ne bénéficieront de l'intervention qu'à la dernière période. Ainsi, à la fin de l'étude, tous les clusters bénéficient de l'intervention.

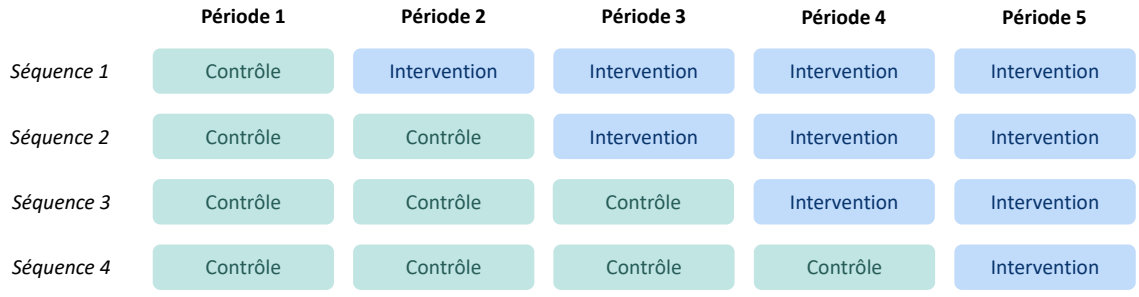


FIGURE IV.2 – Essai randomisé en clusters en *stepped-wedge*

Ce type de schéma d'étude présente plusieurs avantages : (1) il facilite la logistique, grâce à la mise en place de l'intervention échelonnée dans le temps, (2) il permet de motiver les clusters à participer, puisqu'ils bénéficieront tous de l'intervention avant la fin de l'étude et (3) il permet d'obtenir une meilleure efficacité statistique, dans certaines conditions [109].

Néanmoins, les ERC-SW ont un risque plus élevé de biais que les ERC en deux bras parallèles [110] et des méthodes d'analyses statistiques spécifiques sont nécessaires [111]. Par ailleurs, ce schéma d'étude est complexe et nécessite la spécification de plusieurs éléments lors de la planification de l'essai, comme le nombre de séquences, de clusters et de périodes. Dès le début de l'étude, les clusters doivent suivre

et respecter le schéma d'étude planifié que ce soit la date à laquelle implémenter l'intervention ou le nombre de participants à recruter à chaque période. Cependant, des difficultés d'implémentation de l'intervention et de recrutement ont été rapportées dans le cadre des ERC-SW [112].

L'objectif de ce travail était d'étudier et de décrire les difficultés d'implémentation et de recrutement dans les ERC-SW, ainsi que d'évaluer si et comment ces difficultés avaient été prises en compte dans le schéma d'étude et la stratégie d'analyse.

2.2 Méthodes

Sur *PubMed*, une recherche a été effectuée afin d'identifier les ERC-SW éligibles. Pour être inclus, les articles rapportant les ERC-SW devaient être publiés en anglais entre janvier 2019 et septembre 2020, période postérieure à la publication de l'extension du CONSORT pour les ERC-SW [113]. Les ERC-SW devaient inclure au moins 2 séquences, 3 périodes et 5 clusters. Les études pilotes et de faisabilité ainsi que les études évaluant plus d'une intervention étaient exclues.

Pour chaque ERC-SW, les caractéristiques suivantes étaient extraites par deux lecteurs indépendants : les caractéristiques générales (type de clusters, justification du choix d'un ERC-SW, etc), caractéristiques du schéma d'étude planifié et effectivement réalisé, l'intervention évaluée et l'intervention dans le groupe contrôle. Les informations étaient collectées à partir des articles mais également des fichiers supplémentaires, des articles connexes et des protocoles s'ils étaient disponibles.

L'ERC-SW était considéré comme ayant eu une difficulté de recrutement si le nombre de participants ou de clusters à inclure n'était pas atteint ou si le schéma d'étude avait dû être modifié pour atteindre l'effectif planifié. Une difficulté d'implémentation de l'intervention était définie comme une implémentation avancée ou retardée par rapport à la date planifiée ou une absence d'implémentation dans au moins un des clusters.

2.3 Résultats

Au total, 55 ERC-SW ont été sélectionnés parmi lesquels 64% (35/55) présentaient une difficulté de recrutement ou d'implémentation de l'intervention. Certaines études, 7/55 (13%), cumulaient les deux difficultés.

Difficultés de recrutement

Parmi les 55 ERC-SW inclus, 18 (33%) ont présenté une difficulté de recrutement, 28 (42%) n'ont présenté aucun problème de recrutement et pour 23 (26%) essais il n'y avait pas assez d'informations pour juger de l'existence d'un problème de recrutement.

Parmi les 18 ERC-SW présentant une difficulté de recrutement, 1 étude n'avait pas recruté assez de clusters et 13 n'avait pas recruté assez de participants. Quatre ERC-SW ont modifié leur schéma d'étude pour répondre aux difficultés de recrutement, en allongeant la durée de l'étude ou en ajoutant des périodes ou des clusters. Les raisons des difficultés de recrutement étaient indiquées pour seulement 7 essais. La principale raison avancée était qu'il y avait moins de participants éligibles que prévu (n=3). Parmi les 18 articles présentant des difficultés de recrutement, 4 ont adapté leur stratégie d'analyse,

soit en incluant une analyse de sensibilité (n=3), soit en excluant de l'analyse principale la période affectée (n=1).

Difficultés d'implémentation de l'intervention

Des difficultés d'implémentation de l'intervention ont été identifiées pour 24 ERC-SW (44%) et 8 (15%) ERC-SW n'ont pas présenté de difficultés d'implémentation de l'intervention. Pour 23 (42%) ERC-SW il était impossible de juger de l'existence d'un problème d'implémentation.

Parmi les 24 ERC-SW avec un problème d'implémentation, 8 ont avancé l'implémentation pour au moins un cluster, 18 ont retardé l'implémentation et 2 n'ont pas implémenté l'intervention (certains cumulant plusieurs problèmes). Les justifications des problèmes d'implémentation ont été données dans 18 des 24 articles : problèmes logistiques ou techniques (n=6), manque de personnel, grève ou implémentation durant les vacances (n=5), recrutement plus faible que prévu (n=4), problèmes avec les autorisations réglementaires (n=2) ou intervention devenue l'intervention standard au cours de l'étude (n=2). Les raisons étaient donc assez hétérogènes et les difficultés d'implémentation pouvaient être plus ou moins anticipées. Parmi les 24 études, 5 ont adapté leur stratégie d'analyse, soit en ajoutant une analyse de sensibilité (n=3), soit en adaptant l'analyse principale en prenant en compte le schéma effectivement réalisé (n=2).

2.4 Conclusion

Pour conclure, les difficultés d'implémentation de l'intervention et de recrutement dans les ERC-SW ne sont pas rares. Nous avons proposé des recommandations pour aider les futurs chercheurs planifiant un ERC-SW.

Ce travail a donné lieu à un article publié en 2022 dans le *Journal of Clinical Epidemiology* [114], disponible en Annexe C.

Chapitre V

Discussion générale

1 Synthèse des principaux travaux

L'objectif principal de ce travail était de proposer des méthodes adaptées pour estimer l'effet d'une intervention dans un ERC avec un critère de jugement de type survie à l'aide d'une différence de temps de survie moyen restreint.

Dans un premier travail, nous avons étendu aux ERC deux méthodes initialement développées pour des données indépendantes : l'intégration directe des courbes de Kaplan-Meier et la régression sur des pseudo-valeurs. À l'aide d'une étude de simulation, nous avons montré que ces méthodes permettent d'estimer correctement la variance de l'effet intervention et de contrôler l'erreur de type I, s'il y a au moins 50 clusters au total dans l'essai. Pour un nombre plus faible de clusters, la régression sur des pseudo-valeurs avec un test de permutation peut être utilisée pour corriger l'erreur de type I et son inversion permet d'obtenir à la fois une p-valeur et un intervalle de confiance.

Dans un second travail, nous nous sommes concentré sur les cas d'ERC avec un petit nombre de clusters. Nous avons étudié des méthodes alternatives au test de permutation pour la régression sur des pseudo-valeurs, reposant sur des méthodes de correction du biais de l'estimateur de la variance sandwich. À l'aide d'une étude de simulation, nous avons comparé plusieurs de ces méthodes et évalué l'utilisation d'une distribution de Student, à la place d'une distribution normale, pour estimer l'intervalle de confiance et la p-valeur. Nous avons montré que seul l'estimateur de Fay et Graubard a de bonnes performances dans tous les scénarios étudiés et que la distribution de Student doit être utilisée. Cette étude a également montré que les méthodes de correction permettent d'obtenir d'aussi bons résultats que le test de permutation, voire de meilleurs résultats avec une matrice de corrélation échangeable.

2 Recommandations

Ces deux études de simulation mènent aux recommandations suivantes :

1/ Pour estimer une différence de RMST dans un ERC, la méthode de régression sur des pseudo-valeurs

semble plus intéressante que la méthode reposant sur l'intégration des courbes de Kaplan-Meier. En effet, la méthode de régression sur des pseudo-valeurs a de nombreux avantages. L'ajustement sur les covariables est plus simple et elle est déjà disponible dans de nombreux logiciels de statistiques. Le dernier avantage est qu'il est possible d'utiliser cette méthode pour des ERC avec un faible nombre de clusters (< 50), soit avec un test de permutation ou une correction de la variance. Ce dernier avantage n'est pas négligeable, puisqu'il n'est pas rare que les ERC présentent un nombre limité de clusters.

2/ Pour des ERC avec un faible nombre de clusters (< 50), l'intervalle de confiance de l'effet intervention et la p-valeur associée peuvent être estimés avec la régression sur des pseudo-valeurs en utilisant l'estimateur de la variance de Fay et Graubard et une distribution de Student à $K - 2$ degré de liberté, avec K le nombre de clusters au total. Concernant la matrice de corrélation, il est préférable d'utiliser une matrice de corrélation échangeable, quel que soit le nombre de clusters, et quand cela est possible, c'est-à-dire lorsqu'il n'y a pas de problème de convergence. Sinon, une matrice de corrélation indépendante peut être utilisée.

3 Limites et perspectives

3.1 Ajustement sur des covariables

Les deux études de simulation ont évalué la régression sur des pseudo-valeurs en considérant uniquement le groupe d'intervention comme covariable, tant dans la simulation des données que dans l'analyse statistique, sans ajuster sur d'autres covariables. Pourtant, l'ajustement sur les covariables peut s'avérer utile dans les ERC. La randomisation d'un nombre parfois limité de clusters ne permet pas toujours d'équilibrer les variables au niveau du cluster et de l'individu [115, 116], comparé aux essais randomisés individuels. Les performances de la régression sur les pseudo-valeurs pourraient être affectées par l'ajout de covariables supplémentaires dans le modèle (2.7). Des recherches supplémentaires sont donc nécessaires pour étudier l'effet de l'ajustement.

3.2 Autre méthode d'estimation du RMST dans les ERC

Une autre méthode permettant d'estimer une différence de RSMT pour des ERC a récemment été proposée par Chen *et al.* [97], ne reposant pas sur les pseudo-valeurs. Elle modélise le RMST comme une fonction continue du temps d'intérêt t^* dans un modèle de régression et estime les coefficients de régression à l'aide de GEE pondérées par une méthode IPCW, qui permet de prendre en compte des censures dépendantes des covariables. Chen *et al.* ont également développé quatre méthodes de correction de la variance dérivées des estimateurs proposés par Mancl et DeRouen [53], Kauermann et Carroll [54], Fay et Graubard [45] et Morel *et al.* [46].

Cette méthode a l'avantage de pouvoir aider à sélectionner un temps d'horizon t^* pendant la phase exploratoire, contrairement à notre méthode qui nécessite la définition d'un unique temps d'intérêt. Ce travail de thèse se concentrait sur l'analyse primaire ou secondaire d'un ERC et nous avons considéré qu'un temps d'intérêt unique était prédéfini en amont. La régression du RMST en fonction du temps n'était donc pas nécessaire. L'approche de Chen *et al.* permet également de prendre en compte la censure dépendante des covariables. Dans notre travail, nous avons supposé une censure complètement

aléatoire, une hypothèse qui peut ne pas être satisfaite dans la pratique. De plus, Chen *et al.* ont conduit une étude de simulation qui a montré que leur méthode a des bonnes performances, y compris avec un ajustement sur les covariables. Ils ont également étudié deux types de fonction de lien dans l'équation (2.7) (identité et log), qui ont toutes les deux donné des résultats satisfaisants. Enfin, leur méthode peut être utilisée pour des ERC avec un faible nombre de clusters ($K = 20, 30, 40$), grâce aux quatre méthodes de correction qu'ils ont développées. Leur étude de simulation a montré que la meilleure méthode de correction dépendait du niveau de la covariable (niveau individuel ou cluster). Pour des covariables au niveau cluster (comme le groupe dans nos simulations), l'estimateur basé sur Morel *et al.* est à privilégier et pour des covariables au niveau individuel, l'estimateur dérivé de Mancl et DeRouen est à préférer. Ces résultats confirment la nécessité de mener des recherches supplémentaires pour la régression sur des pseudo-valeurs afin d'évaluer l'effet d'un ajustement sur des covariables.

Néanmoins, Chen *et al.* n'ont pas étudié les performances de leurs approches en terme d'erreur de type I dans leur étude de simulation. Les auteurs n'ont pas fait varier le degré de corrélation intra-cluster ou exploré des scénarios avec des risques non-proportionnels. De plus, leur méthode peut paraître complexe à mettre en œuvre par des statisticiens non-experts. En effet, elle nécessite d'abord d'estimer des poids et, si les effets des covariables dépendant du temps sont modélisées par des splines comme dans l'application de l'article, de sélectionner le nombre et l'emplacement des nœuds de ces splines.

3.3 Utilisation des méthodes proposées en pratique

Même si nous avons pu montrer que les méthodes que nous proposons ont de bonnes performances statistiques dans un certain cadre, un frein important à leur utilisation en pratique subsiste : il n'existe pas de formule pour calculer un nombre de sujets nécessaires pour mettre en évidence une différence de RMST dans un ERC.

Une première idée pourrait être de déterminer le nombre de sujets nécessaires pour un essai à randomisation individuelle puis de le multiplier par le facteur d'inflation (1.2) afin d'obtenir le nombre de sujets nécessaires pour un ERC. Cependant, le facteur d'inflation dépend du coefficient de corrélation intra-classe (CCI). Le CCI permet de quantifier la dépendance entre les réponses des individus d'un même cluster pour des critères de jugement continus ou binaires. Or, pour des critères de jugement de type survie, il n'existe pas de définition claire d'une mesure de cette dépendance. Pour les deux études de simulation présentées dans ce manuscrit, la ressemblance des individus au sein d'un même cluster a été quantifiée par le tau de Kendall [117]. Cette mesure est facilement calculable pour un modèle à fragilité gamma [66], qui a été utilisé pour simuler les données de survie corrélées dans nos travaux. Cependant, d'autres mesures existent comme par exemple le coefficient de corrélation de Spearman ou la *median concordance* [117]. Cette thèse se focalise sur l'analyse d'un ERC avec un critère de jugement de type survie et non sur la quantification de la corrélation intra-cluster. Le choix du tau de Kendall a été fait afin de pouvoir donner un indicateur du niveau de corrélation intra-cluster pour chaque scénario dans les études de simulation et, par conséquent, des recherches approfondies n'ont pas été menées. Cependant, des recherches supplémentaires sont nécessaires afin de fournir une définition claire et une mesure de corrélation intra-cluster pour des critères de jugement de type survie dans les ERC afin de pouvoir la quantifier pour la prendre en compte dans le calcul du nombre de sujets nécessaires et de la rapporter dans chaque étude.

Pour finir, une alternative pour calculer le nombre de sujets nécessaires pour être une approche par simulation.

Les questions concernant le nombre de sujets nécessaires et la quantification de la corrélation intra-cluster sont actuellement en cours de recherche au sein de l'unité Inserm SPHERE dans le cadre du projet QUARTET (QUantification of clustering and Analysis of cluster Randomised Trials with time-to-EvenT outcomes), dans lequel s'inscrit cette thèse.

3.4 Extensions des méthodes

Dans ce travail de thèse, nous nous sommes concentrés sur le cadre spécifique des données de survie censurées à droite. Des recherches pour des données censurées par intervalle ou tronquées à gauche pourraient également être menées. Pour des données indépendantes, une extension de la régression sur des pseudo-valeurs aux données censurées par intervalle a été proposée dans le cadre d'un modèle *illness-death*, ou "sain-malade-mort", pour estimer l'effet de facteurs de risques sur différents indicateurs de la démence, dont le temps de survie moyen restreint sans démence [118]. Une extension pour les données tronquées à gauche a également été développée, pour des censures indépendantes ou non-indépendantes (dépendantes des covariables ou des temps d'événement) [119].

Des extensions des méthodes proposées dans cette thèse aux modèles multi-états pourraient également être étudiées. La régression de pseudo-valeurs a d'ailleurs été initialement proposée pour des données indépendantes dans le cadre de modèle multi-états pour estimer la probabilité d'être dans un état [120]. Pour des données corrélées en présence de risques compétitifs, la régression sur des pseudo-valeurs a été étendue pour étudier la fonction d'incidence cumulée pour des censures indépendantes [121] ou corrélées au sein des clusters [122].

Par ailleurs, dans le cadre des risques compétitifs, une autre mesure d'intérêt découlant du RMST peut être intéressante : le nombre d'années perdues. Il correspond à la différence entre le temps d'intérêt t^* et le RMST [123]. Cette mesure est particulièrement intéressante dans le cas des risques compétitifs car le nombre total d'années perdues peut être décomposé en une somme de termes qui peuvent être interprétés comme le nombre d'années perdues "dues à chaque cause" [123, 124]. Dans le cadre des données indépendantes, la régression sur des pseudo-valeurs a été proposée pour calculer un nombre d'années perdues dans le cadre de risques compétitifs [123]. L'extension des méthodes proposées au nombre d'années perdues, pour des risques compétitifs ou non, dans le cadre des ERC pourrait faire l'objet de recherches.

Les méthodes présentées dans les travaux de cette thèse ont été proposées pour le cas particulier des données corrélées des ERC. Il pourrait être intéressant de voir si ces méthodes peuvent être utilisées pour d'autres types de données corrélées, comme pour des événements récurrents ou pour des méta-analyses.

De plus, nous nous sommes intéressés au cas de données hiérarchiques à deux "niveaux" : les individus et les clusters. Cependant, dans certains cas, les données peuvent être structurées selon plus de niveaux. Par exemple, l'essai FADEMELA (section 1 du chapitre IV) présentaient trois niveaux hiérarchiques : les individus, les familles et les hôpitaux. L'estimation d'une différence de RMST pourrait également être étudiée dans ce cadre.

Enfin, ce travail pourrait aussi être étendu à d'autres schémas d'étude que celui des ERC en deux bras parallèles, comme les ERC en cross-over ou en stepped-weddge, qui nécessite, notamment, la prise en compte d'une corrélation longitudinale.

3.5 Revue de la littérature des méthodes d'analyses de survie dans les ERC

Une dernière perspective plus générale est de mener une revue de la littérature concernant l'ensemble des méthodes existantes, pas uniquement pour l'estimation d'une différence de RMST, qui pourraient être utilisées pour analyser un critère de jugement de type survie dans un ERC. L'objectif final serait de rédiger un guide synthétisant l'ensemble des méthodes disponibles, que ce soit pour tester l'existence d'un effet intervention ou de le quantifier, et donnant des recommandations afin de faciliter en pratique l'analyse des données de survie dans le cadre des ERC.

4 Conclusion générale

La popularité des ERC a fortement augmenté au cours des dernières années. L'utilisation d'un critère de jugement de type survie dans ces essais au design spécifique n'est pas rare, mais est souvent mal analysé, sans prendre en compte de la corrélation intra-cluster dans l'analyse statistique. Dans ce contexte, ces travaux de thèse ouvrent la voie à une meilleure analyse des critères de jugement de type survie dans les ERC. Ils permettent l'utilisation dans le cadre des ERC de la différence de RMST, qui présente de nombreux avantages, notamment de fournir une interprétation simple de l'effet d'une intervention.

Références

[1] Donner, A. et Klar, N. *Design and Analysis of Cluster Randomization Trials in Health Research.* Wiley, 2000.

[2] Moberg, J. et Kramer, M. A brief history of the cluster randomised trial design. *Journal of the Royal Society of Medicine*, 108(5) :192–198, 2015.

[3] Murray, D. M., Varnell, S. P. et Blitstein, J. L. Design and analysis of group-randomized trials : a review of recent methodological developments. *American Journal of Public Health*, 94(3) :423–432, 2004.

[4] McRae, A. D., Taljaard, M. et Weijer, C. Cluster-randomized trials : A closer look. *Clinical Trials*, 13(3) :294–300, 2016.

[5] Lorenz, E., Köpke, S., Pfaff, H. et Blettner, M. Cluster-Randomized Studies. *Deutsches Arzteblatt International*, 115(10) :163–168, 2018.

[6] Campbell, M. K., Elbourne, D. R., Altman, D. G. et CONSORT group. CONSORT statement : extension to cluster randomised trials. *BMJ*, 328(7441) :702–708, 2004.

[7] Campbell, M. K., Piaggio, G., Elbourne, D. R., Altman, D. G. et CONSORT Group. Consort 2010 statement : extension to cluster randomised trials. *BMJ*, 345 :e5661, 2012.

[8] Eldridge, S. M. et Kerry, S. *Practical Guide to Cluster Randomised Trials in Health Services Research.* Wiley, 2012.

[9] Drahota, A. K., Ward, D., Udell, J. E., Soilemezi, D., Ogollah, R., Higgins, B., Dean, T. P. et Severs, M. Pilot cluster randomised controlled trial of flooring to reduce injuries from falls in wards for older people. *Age and Ageing*, 42(5) :633–640, 2013.

[10] Griffiths, C., Bremner, S., Islam, K., Sohanpal, R., Vidal, D.-L., Dawson, C., Foster, G., Ramsay, J., Feder, G., Taylor, S., Barnes, N., Choudhury, A., Packe, G., Bayliss, E., Trathen, D., Moss, P., Cook, V., Livingstone, A. E. et Eldridge, S. Effect of an Education Programme for South Asians with Asthma and Their Clinicians : A Cluster Randomised Controlled Trial (OEDIPUS). *PloS One*, 11(12) :e0158783, 2016.

[11] Naldi, L., Chatenoud, L., Bertuccio, P., Zinetti, C., Di Landro, A., Scotti, L., La Vecchia, C. et Oncology Cooperative Group of the Italian Group for Epidemiologic Research in Dermatology (GISED). Improving sun-protection behavior among children : results of a cluster-randomized trial in Italian elementary schools. The "SoleSi SoleNo-GISED" Project. *The Journal of Investigative Dermatology*, 127(8) :1871–1877, 2007.

[12] Kelly, P., Katubulushi, M., Todd, J., Banda, R., Yambayamba, V., Fwoloshi, M., Zulu, I., Kafwembe, E., Yavwa, F., Sanderson, I. R. et Tomkins, A. Micronutrient supplementation has

limited effects on intestinal infectious disease and mortality in a Zambian population of mixed HIV status : a cluster randomized trial. *The American Journal of Clinical Nutrition*, 88(4) :1010–1017, 2008.

[13] Lemaitre, M., Meret, T., Rothan-Tondeur, M., Belmin, J., Lejonc, J.-L., Luquel, L., Piette, F., Salom, M., Verny, M., Vetel, J.-M., Veyssier, P. et Carrat, F. Effect of influenza vaccination of nursing home staff on mortality of residents : a cluster-randomized trial. *Journal of the American Geriatrics Society*, 57(9) :1580–1586, 2009.

[14] Murira, N., Munjanja, S. P., Zhanda, I., Nystrom, L. et Lindmark, G. Effect of a new antenatal care programme on the attitudes of pregnant women and midwives towards antenatal care in Harare. *The Central African Journal of Medicine*, 43(5) :131–135, 1997.

[15] Giraudeau, B. Negative values of the intraclass correlation coefficient are not theoretically possible. *Journal of Clinical Epidemiology*, 49(10) :1205–1206, 1996.

[16] Eldridge, S. M., Ukoumunne, O. C. et Carlin, J. B. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials : A Review of Definitions on JSTOR. 77(3) :378–394, 2009.

[17] Campbell, M. K., Grimshaw, J. M. et Elbourne, D. R. Intracluster correlation coefficients in cluster randomized trials : empirical insights into how should they be reported. *BMC medical research methodology*, 4 :9, 2004.

[18] Gulliford, M. C., Adams, G., Ukoumunne, O. C., Latinovic, R., Chinn, S. et Campbell, M. J. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58(3) :246–251, 2005.

[19] Mbekwe Yepnang, A. M., Caille, A., Eldridge, S. M. et Giraudeau, B. Is the R coefficient of interest in cluster randomized trials with a binary outcome ? *Statistical Methods in Medical Research*, 29(9) :2470–2480, 2020.

[20] Mbekwe Yepnang, A. M., Caille, A., Eldridge, S. M. et Giraudeau, B. Association of intracluster correlation measures with outcome prevalence for binary outcomes in cluster randomised trials. *Statistical Methods in Medical Research*, 30(8) :1988–2003, 2021.

[21] Campbell, M. K., Fayers, P. M. et Grimshaw, J. M. Determinants of the intracluster correlation coefficient in cluster randomized trials : the case of implementation research. *Clinical Trials*, 2(2) :99–107, 2005.

[22] Donner, A. A review of Inference Procedures for the Intraclass COrrelation Coefficient in the One-Way Random Effect Model. 54 :67–82, 1986.

[23] Ridout, M. S., Demétrio, C. G. et Firth, D. Estimating intraclass correlation for binary data. *Biometrics*, 55(1) :137–148, 1999.

[24] Ukoumunne, O. C. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine*, 21(24) :3757–3774, 2002.

[25] Froud, R., Eldridge, S., Diaz Ordaz, K., Marinho, V. C. C. et Donner, A. Quality of cluster randomized controlled trials in oral health : a systematic review of reports published between 2005 and 2009. *Community Dentistry and Oral Epidemiology*, 40 Suppl 1 :3–14, 2012.

[26] Korevaar, E., Kasza, J., Taljaard, M., Hemming, K., Haines, T., Turner, E. L., Thompson, J. A., Hughes, J. P. et Forbes, A. B. Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials. *Clinical Trials*, 18(5) :529–540, 2021.

[27] CLOUD (CLustered OUtcome Dataset) Bank. https://monash-biostat.shinyapps.io/CLOUDbank/ [Consulté le 02/08/2023].

[28] Rutterford, C., Copas, A. et Eldridge, S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3) :1051–1067, 2015.

[29] Guittet, L., Giraudeau, B. et Ravaud, P. A priori postulated and real power in cluster randomized trials : mind the gap. *BMC medical research methodology*, 5 :25, 2005.

[30] Eldridge, S. M., Ashby, D. et Kerry, S. Sample size for cluster randomized trials : effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5) :1292–1300, 2006.

[31] Hayes, R. J. et Moulton, L. H. *Cluster Randomised Trials*. Chapman and Hall/CRC, New York, 2 edition, 2017.

[32] Taljaard, M., McRae, A. D., Weijer, C., Bennett, C., Dixon, S., Taleban, J., Skea, Z., Eccles, M. P., Brehaut, J. C., Donner, A., Saginur, R., Boruch, R. F. et Grimshaw, J. M. Inadequate reporting of research ethics review and informed consent in cluster randomised trials : review of random sample of published trials. *BMJ*, 342 :d2496, 2011.

[33] Turner, E. L., Prague, M., Gallis, J. A., Li, F. et Murray, D. M. Review of Recent Methodological Developments in Group-Randomized Trials : Part 2-Analysis. *American Journal of Public Health*, 107(7) :1078–1086, 2017.

[34] Noh, M. et Lee, Y. REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98(5) :896–915, 2007.

[35] Liang, K.-Y. et Zeger, S. L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1) :13–22, 1986.

[36] Zeger, S. L. et Liang, K.-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42(1) :121–130, 1986.

[37] Leyrat, C., Morgan, K. E., Leurent, B. et Kahan, B. C. Cluster randomized trials with a small number of clusters : which analyses should be used ? *International Journal of Epidemiology*, 47(1) :321–331, 2018.

[38] Hemming, K. et Taljaard, M. Key considerations for designing, conducting and analysing a cluster randomized trial. *International Journal of Epidemiology*, page dyad064, 2023.

[39] Ivers, N., Taljaard, M., Dixon, S., Bennett, C., Mcrae, A., Taleban, J., Skea, Z., Brehaut, J., Boruch, R., Eccles, M., Grimshaw, J., Weijer, C., Zwarenstein, M. et Donner, A. 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*, 343 :d5886, 2011.

[40] Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R. et Ukoumunne, O. C. Lessons for cluster randomized trials in the twenty-first century : a systematic review of trials in primary care. *Clinical Trials*, 1(1) :80–90, 2004.

[41] Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., Hooper, R., Wright, N., Eldridge, S. M. et Leyrat, C. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters : a review, reanalysis, and simulation study. *Trials*, 17 :438, 2016.

[42] Li, P. et Redden, D. T. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC medical research methodology*, 15 :38, 2015.

[43] Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics*, 2(6) :110–114, 1946.

[44] Kenward, M. G. et Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3) :983–997, 1997.

[45] Fay, M. P. et Graubard, B. I. Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. *Biometrics*, 57(4) :1198–1206, 2001.

[46] Morel, J., Bokossa, M. et Neerchal, N. Small sample correction for the variance of GEE estimators. *Biometrical Journal*, 45(4) :395–409, 2003.

[47] Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I. et Wolfson, M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63(3) :935–941, 2007.

[48] Blaha, O., Esserman, D. et Li, F. Design and analysis of cluster randomized trials with time-to-event outcomes under the additive hazards mixed model. *Statistics in Medicine*, 41(24) :4860–4885, 2022.

[49] Wang, X., Turner, E. et Li, F. Improving sandwich variance estimation for marginal Cox analysis of cluster randomized trials. *Biometrical Journal*, 65, 2022.

[50] Wang, M., Kong, L., Li, Z. et Zhang, L. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine*, 35(10) :1706–1721, 2016.

[51] Ford, W. P. et Westgate, P. M. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biometrical Journal*, 59(3) :478–495, 2017.

[52] Li, P. et Redden, D. T. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34(2) :281–296, 2015.

[53] Mancl, L. A. et DeRouen, T. A. A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1) :126–134, 2001.

[54] Kauermann, G. et Carroll, R. J. A Note on the Efficiency of Sandwich Covariance Matrix Estimation. *Journal of the American Statistical Association*, 96(456) :1387–1396, 2001.

[55] Pan, W. et Wall, M. M. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, 21(10) :1429–1441, 2002.

[56] Kaplan, E. L. et Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481, 1958.

[57] Greenwood, M. A Report on the Natural Duration of Cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926. Publisher : London : H.M.S.O.

[58] Nelson, W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14(4) :945–966, 1972.

[59] Aalen, O. O. Nonparametric Inference for a Family of Counting Processes. *Annals of Statistics*, 6(4) :701–726, 1978.

[60] Breslow, N. E. Contribution to the discussion of paper by D.R. Cox. 34 :216–217, 1972.

[61] Fleming, T. R. et Harrington, D. P. *Counting Processes and Survival Analysis*. Wiley & Sons, Chichester, 1991.

[62] Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220, 1972.

[63] Cox, D. R. Partial likelihood. *Biometrika*, 62(2) :269–276, 1975.

[64] Breslow, N. E. Covariance analysis of censored survival data. *Biometrics*, 30(1) :89–99, 1974.

[65] Efron, B. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 72(359) :557–565, 1977.

[66] Duchateau, L. et Janssen, P. *The frailty model*. Springer Verlag, 2008. ISSN : 1431-8776 Publication Title : Statistics for Biology and Health series.

[67] Therneau, T. et Grambsch, P. *Modeling survival data : extending the cox model*. Statistics for biology and health. Springer, New York, USA, 2000.

[68] Aalen, O. O. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8) :907–925, 1989.

[69] Lin, D. Y. et Ying, Z. Semiparametric Analysis of the Additive Risk Model. *Biometrika*, 81(1) :61–71, 1994.

[70] McKeague, I. W. et Sasieni, P. D. A Partly Parametric Additive Risk Model. *Biometrika*, 81(3) :501–514, 1994.

[71] Messori, A., Bartoli, L. et 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 failure*, 8(3) :2345–2348, 2021.

[72] Trinquart, L., Jacot, J., Conner, S. C. et Porcher, R. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 34(15) :1813–1819, 2016.

[73] Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. et Wei, L.-J. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 32(22) :2380–2385, 2014.

[74] Uno, H., Wittes, J., Fu, H., Solomon, S. D., Claggett, B., Tian, L., Cai, T., Pfeffer, M. A., Evans, S. R. et Wei, L.-J. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. *Annals of Internal Medicine*, 163(2) :127–134, 2015.

[75] Royston, P. et Parmar, M. K. B. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19) :2409–2421, 2011.

[76] Hernán, M. A. The Hazards of Hazard Ratios. *Epidemiology (Cambridge, Mass.)*, 21(1) :13–15, 2010.

[77] Stensrud, M. J., Aalen, J. M., Aalen, O. O. et Valberg, M. Limitations of hazard ratios in clinical trials. *European Heart Journal*, 40(17) :1378–1383, 2019.

[78] Hasegawa, T., Misawa, S., Nakagawa, S., Tanaka, S., Tanase, T., Ugai, H., Wakana, A., Yodo, Y., Tsuchiya, S. et Suganami, H. Restricted mean survival time as a summary measure of time-to-event outcome. *Pharmaceutical Statistics*, 19(4) :436–453, 2020.

[79] Tian, L., Zhao, L. et Wei, L. J. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2) :222–233, 2014.

[80] Vilain-Abraham, F. L., Tavernier, E., Dantan, E., Desmée, S. et Caille, A. Restricted mean survival time to estimate an intervention effect in a cluster randomized trial. *Statistical Methods in Medical Research*, page 9622802231192960, 2023.

[81] Conner, S. C., Sullivan, L. M., Benjamin, E. J., LaValley, M. P., Galea, S. et Trinquart, L. Adjusted restricted mean survival times in observational studies. *Statistics in Medicine*, 38(20) :3832–3860, 2019.

[82] Karrison, T. Restricted Mean-Life with Adjustment for Covariates. *Journal of the American Statistical Association*, 82(400) :1169–1176, 1987.

[83] Zucker, D. M. Restricted mean life with covariates : Modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442) :702–709, 1998.

[84] Chen, P.-Y. et Tsiatis, A. A. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4) :1030–1038, 2001.

[85] Zhang, M. et Schaubel, D. E. Estimating Differences in Restricted Mean Lifetime Using Observational Data Subject to Dependent Censoring. *Biometrics*, 67(3) :740–749, 2011.

[86] Royston, P. et Parmar, M. K. B. 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 medical research methodology*, 13 :152, 2013.

[87] Andersen, P. K., Hansen, M. G. et Klein, J. P. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10(4) :335–350, 2004.

[88] Andersen, P. K. et Perme, M. P. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1) :71–99, 2010.

[89] Wang, X. et Schaubel, D. E. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime Data Analysis*, 24(1) :176–199, 2018.

[90] Zhong, Y. et Schaubel, D. E. Restricted mean survival time as a function of restriction time. *Biometrics*, 78(1) :192–201, 2022.

[91] Caille, A., Tavernier, E., Taljaard, M. et Desmée, S. Methodological review showed that time-to-event outcomes are often inadequately handled in cluster randomized trials. *Journal of Clinical Epidemiology*, 134 :125–137, 2021.

[92] Wienke, A. *Frailty Models in Survival Analysis*. Chapman and Hall/CRC, New York, 2010.

[93] Lin, D. et Wei, L. J. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, 84(408) :1074–1078, 1989.

[94] Lin, D. Y. Cox regression analysis of multivariate failure time data : the marginal approach. *Statistics in Medicine*, 13(21) :2233–2247, 1994.

[95] Wei, L., Lin, D. et Weissfeld, L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of The American Statistical Association*, 84 :1065–1073, 1989.

[96] Lee, E. W., Wei, L. J., Amato, D. A. et Leurgans, S. Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations. In Klein, J. P. et Goel, P. K., editors, *Survival Analysis : State of the Art*, Nato Science, pages 237–247. Springer Netherlands, 1992.

[97] Chen, X., Harhay, M. O. et Li, F. Clustered restricted mean survival time regression. *Biometrical Journal. Biometrische Zeitschrift*, 2022.

[98] CaSciModOT (Calcul Scientifique et Modélisation Orléans-Tours). https://cascimodot.fr/ [Consulté le 07/08/2023].

[99] Dyba, T., Randi, G., Bray, F., Martos, C., Giusti, F., Nicholson, N., Gavin, A., Flego, M., Neamtiu, L., Dimitrova, N., Negrão Carvalho, R., Ferlay, J. et Bettio, M. The European cancer burden in 2020 : Incidence and mortality estimates for 40 countries and 25 major cancers. *European Journal of Cancer*, 157 :308–347, 2021.

[100] Kovalyshyn, I., Dusza, S. W., Siamas, K., Halpern, A. C., Argenziano, G. et Marghoob, A. A. The Impact of Physician Screening on Melanoma Detection. *Archives of Dermatology*, 147(11) :1269–1275, 2011.

[101] Swetter, S. M., Pollitt, R. A., Johnson, T. M., Brooks, D. R. et Geller, A. C. Behavioral determinants of successful early melanoma detection. *Cancer*, 118(15) :3725–3734, 2012.

[102] Dessinioti, C., Geller, A. C., Stergiopoulou, A., Swetter, S. M., Baltas, E., Mayer, J. E., Johnson, T. M., Talaganis, J., Trakatelli, M., Tsoutsos, D., Tsourouflis, G. et Stratigos, A. J. Association of Skin Examination Behaviors and Thinner Nodular vs Superficial Spreading Melanoma at Diagnosis. *JAMA Dermatology*, 154(5) :544–553, 2018.

[103] Olsen, C. M., Carroll, H. J. et Whiteman, D. C. Familial Melanoma : A Meta-analysis and Estimates of Attributable Fraction. *Cancer Epidemiology, Biomarkers & Prevention*, 19(1) :65–73, 2010.

[104] Ford, D., Bliss, J. M., Swerdlow, A. J., Armstrong, B. K., Franceschi, S., Green, A., Holly, E. A., Mack, T., Mackie, R. M., Østerlind, A., Walter, S. D., Peto, J. et Easton, D. F. Risk of cutaneous melanoma associated with a family history of the disease. *International Journal of Cancer*, 62(4) :377–381, 1995.

[105] Rhodes, A. R., Weinstock, M. A., Fitzpatrick, T. B., Mihm, Jr, M. C. et Sober, A. J. Risk Factors for Cutaneous Melanoma : A Practical Method of Recognizing Predisposed Individuals. *JAMA*, 258(21) :3146–3154, 1987.

[106] ALD n° 30. Guide Médecin sur le Mélanome Cutané. https://www.has-sante.fr/upload/docs/application/pdf/201203/ald_30_guide_melanome_web.pdf [Consulté le 20/04/2022].

[107] Losina, E., Walensky, R. P., Geller, A., Beddingfield, III, F. C., Wolf, L. L., Gilchrest, B. A. et Freedberg, K. A. Visual Screening for Malignant Melanoma : A Cost-effectiveness Analysis. *Archives of Dermatology*, 143(1) :21–28, 2007.

[108] Marcé, D., Le Vilain-Abraham, F., Bridou, M., Quéreux, G., Dupuy, A., Lesimple, T., Le Corre, Y., Wierzbicka-Hainaut, E., Legoupil, D., Célérier, P., Maillard, H., Machet, L. et Caille, A. Impact of Distribution of a Tip Sheet to Increase Early Detection and Prevention Behavior among First-Degree Relatives of Melanoma Patients : A Randomized Cluster Trial. *Cancers*, 14(16) :3864, 2022.

[109] Campbell, M. J., Hemming, K. et Taljaard, M. The stepped wedge cluster randomised trial : what it is and when it should be used. *Medical Journal of Australia*, 210(6) :253–254.e1, 2019.

[110] Hemming, K. et Taljaard, M. Reflection on modern methods : when is a stepped-wedge cluster randomized trial a good study design choice ? *International Journal of Epidemiology*, 49(3) :1043–1052, 2020.

[111] Hussey, M. A. et Hughes, J. P. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2) :182–191, 2007.

[112] Prost, A., Binik, A., Abubakar, I., Roy, A., De Allegri, M., Mouchoux, C., Dreischulte, T., Ayles, H., Lewis, J. J. et Osrin, D. Logistic, ethical, and political dimensions of stepped wedge trials : critical review and case studies. *Trials*, 16(1) :351, 2015.

[113] Hemming, K., Taljaard, M., McKenzie, J. E., Hooper, R., Copas, A., Thompson, J. A., Dixon-Woods, M., Aldcroft, A., Doussau, A., Grayling, M., Kristunas, C., Goldstein, C. E., Campbell, M. K., Girling, A., Eldridge, S., Campbell, M. J., Lilford, R. J., Weijer, C., Forbes, A. B. et Grimshaw, J. M. Reporting of stepped wedge cluster randomised trials : extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*, 363 :k1614, 2018.

[114] Caille, A., Taljaard, M., Le Vilain-Abraham, F., Le Moigne, A., Copas, A. J., Tubach, F. et Dechartres, A. Recruitment and implementation challenges were common in stepped-wedge cluster randomized trials : Results from a methodological review. *Journal of Clinical Epidemiology*, 148 :93–103, 2022.

[115] Wright, N., Ivers, N., Eldridge, S., Taljaard, M. et Bremner, S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *Journal of Clinical Epidemiology*, 68(6) :603–609, 2015.

[116] Dron, L., Taljaard, M., Cheung, Y. B., Grais, R., Ford, N., Thorlund, K., Jahan, F., Nakimuli-Mpungu, E., Xavier, D., Bhutta, Z. A., Park, J. J. H. et Mills, E. J. The role and challenges of cluster randomised trials for global health. *The Lancet. Global Health*, 9(5) :e701–e710, 2021.

[117] Hougaard, P. *Analysis of Multivariate Survival Data.* Springer Science & Business Media, 2000.

[118] Sabathé, C., Andersen, P. K., Helmer, C., Gerds, T. A., Jacqmin-Gadda, H. et Joly, P. Regression analysis in an illness-death model with interval-censored data : A pseudo-value approach. *Statistical Methods in Medical Research*, 29(3) :752–764, 2020.

[119] Rong, R., Ning, J. et Zhu, H. Regression modeling of restricted mean survival time for left-truncated right-censored data. *Statistics in Medicine*, 41(16) :3003–3021, 2022.

[120] Andersen, P., Klein, J. et Rosthøj, S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 1, 2003.

[121] Logan, B. R., Zhang, M.-J. et Klein, J. P. Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics*, 67(1) :1–7, 2011.

[122] Ahn, K. W. et Logan, B. R. Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data. *The Annals of Applied Statistics*, 10(2) :618–637, 2016.

[123] Andersen, P. K. Decomposition of number of life years lost according to causes of death. *Statistics in Medicine*, 32(30) :5278–5285, 2013.

[124] Andersen, P. K., Canudas-Romo, V. et Keiding, N. Cause-specific measures of life years lost. *Demographic Research*, 29 :1127–1152, 2013.

Annexe A

Algorithme d'estimation des paramètres avec les équations d'estimation généralisées (GEE)

L'algorithme suivant est utilisé pour estimer les paramètres du modèle (1.4) à l'aide des GEE.

1. Les estimations initiales des coefficients de régression $\boldsymbol{\beta}$ sont obtenues avec un modèle de régression linéaire généralisé en supposant que les individus sont indépendants

2. La matrice de corrélation $\mathbf{R}$ est estimée en fonction des résidus, des valeurs actuelles de $\boldsymbol{\beta}$ et de sa structure supposée. Dans le cas d'une matrice de corrélation échangeable, définie par l'équation (1.7), le paramètre ρ est estimé de la manière suivante :

$$\hat{\rho} = \frac{1}{(n^* - p)\hat{\phi}} \sum_{k=1}^{K} \sum_{i \neq j} \hat{r}_{ki} \hat{r}_{kj} \tag{0.1}$$

   avec $n^* = \sum_{k=1}^{K} m_k(m_k - 1)$, p la taille du vecteur $\boldsymbol{\beta}$, $\hat{\phi} = \frac{1}{n-p} \sum_{k=1}^{K} \sum_{i=1}^{m_k} \hat{r}_{ki}^2$, n le nombre total d'individus où $n = \sum_{k=1}^{K} m_k$ et $\hat{r}_{ki}$ les résidus de Pearson défini comme $\frac{(Y_{ki} - \mu_{ki})}{\sqrt{Var(\mu_{ki})}}$

3. La matrice de covariance est ensuite estimée :

$$\hat{\mathbf{V}}_k = \hat{\phi} \mathbf{A}_k^{\frac{1}{2}} \hat{\mathbf{R}}(\rho) \mathbf{A}_k^{\frac{1}{2}} \tag{0.2}$$

4. Le vecteur des paramètres $\boldsymbol{\beta}$ est mis à jour :

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j + \left[\sum_{k=1}^{K} \left(\frac{\partial \mu_k}{\partial \beta} \right)' \mathbf{V}_k^{-1} \left(\frac{\partial \mu_k}{\partial \beta} \right) \right]^{-1} \left[\sum_{k=1}^{K} \left(\frac{\partial \mu_k}{\partial \beta} \right)' \mathbf{V}_k^{-1} (\mathbf{Y}_k - \mu_k) \right] \tag{0.3}$$

5. Les étapes 2 à 4 sont répétées jusqu'à convergence.

Annexe B

Analyse de l'essai randomisé en clusters FADEMELA : Article

cancers

MDPI

Article

Impact of Distribution of a Tip Sheet to Increase Early Detection and Prevention Behavior among First-Degree Relatives of Melanoma Patients: A Randomized Cluster Trial

Diane Marcé [1], Floriane Le Vilain-Abraham [2], Morgiane Bridou [3], Gaelle Quéreux [4], Alain Dupuy [5], Thierry Lesimple [6], Yannick Le Corre [7], Ewa Wierzbicka-Hainaut [8], Delphine Legoupil [9], Philippe Célérier [10], Hervé Maillard [11], Laurent Machet [1,12,*] and Agnès Caille [2,13,*]

1 Service de Dermatologie, Centre Hospitalier Régional Universitaire de Tours, 37044 Tours, France
2 INSERM, Université de Tours, Université de Nantes, SPHERE U1246, 37044 Tours, France
3 Laboratoire de Psychopathologie et Neuropsychologie, Université Paris 8, 93200 Paris, France
4 Dermatology Department, Nantes University Hospital, Université de Nantes, CIC 1413, Inserm UMR 1302/EMR6001 INCIT, 44000 Nantes, France
5 Service de Dermatologie, Centre Hospitalier Universitaire de Rennes, 35033 Rennes, France
6 Service d'Oncologie Médicale, Centre Anticancéreux de Rennes, 35033 Rennes, France
7 Service de Dermatologie, Centre Hospitalier Universitaire d'Angers, 49100 Angers, France
8 Service de Dermatologie, Centre Hospitalier Universitaire de Poitiers, 86000 Poitiers, France
9 Service de Dermatologie, Centre Hospitalier Universitaire de Brest, 29200 Brest, France
10 Service de Dermatologie, Centre Hospitalier de La Rochelle, 17000 La Rochelle, France
11 Service de Dermatologie, Centre Hospitalier du Mans, 72037 Le Mans, France
12 INSERM, Université de Tours, iBrain U1253, 37044 Tours, France
13 INSERM CIC1415, Centre Hospitalier Régional Universitaire de Tours, 37044 Tours, France
* Correspondence: machet@univ-tours.fr (L.M.); agnes.caille@univ-tours.fr (A.C.)

check for updates

Citation: Marcé, D.; Le Vilain-Abraham, F.; Bridou, M.; Querreux, G.; Dupuy, A.; Lesimple, T.; Le Corre, Y.; Wierzbicka-Hainaut, E.; Legoupil, D.; Celerier, P.; et al. Impact of Distribution of a Tip Sheet to Increase Early Detection and Prevention Behavior among First-Degree Relatives of Melanoma Patients: A Randomized Cluster Trial. *Cancers* **2022**, *14*, 3864. https://doi.org/10.3390/cancers14163864

Received: 20 June 2022
Accepted: 8 August 2022
Published: 10 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Simple Summary: The risk of melanoma is higher in first-degree relatives (FDRs, i.e., brother, sister, father, mother, or children) of a patient with melanoma than in the general population. FDRs are advised to undergo annual screening to detect any melanoma earlier and to adopt sun-protective behavior by seeking shade, wearing a hat and long-sleeved clothing, staying indoors between 12 noon and 4 p.m., and applying sunscreen with SPF > 50. We know that these general instructions, usually given orally to the patients, are inconsistently followed by FDRs. Our goal was to determine whether written support intended for FDRs of patients would improve early detection and photoprotection as compared with usual oral advice. We developed and evaluated the use of a tip sheet given by patients to their FDRs. The adherence of FDRs to early detection by medical examination and to sun protection was not improved by delivery of the tip sheet as compared with the usual oral advice.

Abstract: Background: First-degree relatives (FDRs, defined as parents, children, and siblings) of melanoma patients are at a two-to-fivefold increased risk of developing melanoma themselves. FDRs are advised to perform self-skin examination (SSE) and annual medical total cutaneous examination (TCE) performed either by a dermatologist or a general practitioner, and to change their sun-related behavior. This advice is given orally to melanoma patients who are asked to relay the information to their FDRs. Objective: Our aim was to determine the impact of providing a tip sheet to melanoma patients intended to their first-degree relatives (FDRs) on early detection and sun-related behaviors in this group at increased risk of melanoma. Methods: A superiority, cluster-randomized trial was conducted at nine hospital centers. In the intervention group, dermatologists were asked to deliver to melanoma patients (index cases) the tip sheet and oral advice intended to their FDRs. The control group were asked to deliver the usual oral advice alone. The primary outcome was early detection of melanoma in FDRs with a medical TCE performed within one year after the first visit of the index case. Secondary outcomes were SSE and sun-related behaviors in FDRs. Results: A total of 48 index cases and 114 FDRS in the control group, 60 index cases and 166 FDRS in the intervention group were recruited. In the intervention group, 36.1% of FDRs performed a medical TCE as compared to 39.5% of FDRs in the control group (OR 0.9 [95% CI 0.5 to 1.5], $p = 0.63$). We did not find a

between-group difference in SSE and sun-related behaviors. Conclusion: A tip sheet added to the usual oral advice did not increase medical TCE among FDRs of melanoma patients. Overall, the rate of TCE among FDRs was low. Research on other strategies is needed to increase melanoma detection in this population.

Keywords: cluster randomized trial; melanoma; prevention; screening; intervention; medical skin examination; sun-related behaviors; first-degree relatives; high-risk population; familial

1. Introduction

The incidence of cutaneous melanoma (CM) still continues to increase worldwide [1] and results in 3.5 deaths per 100,000 people per year in Europe [2]. Greater tumor thickness and delayed diagnosis are associated with increased mortality. Medical total cutaneous examination (TCE) performed by a dermatologist or a general practitioner (GP), and skin self-examination (SSE) are significantly associated with thinner melanomas [3–5]. However, prevention and early detection on a large scale in the general population are expensive, and the potential benefit of skin cancer screening on CM mortality has not been demonstrated [6,7]. The cost of modern treatments in advanced stages has markedly increased with modern therapeutic options [8]; therefore, prevention programs may become cost-effective or even cost-saving.

First-degree relatives (FDRs; parents, siblings, children) of CM patients are at increased risk of developing CM [9–11]. Personalized screening of at-increased-risk individuals is recommended in many countries, including France [12]. A computer simulation study performed in the United States before the modern era of effective but costly medical treatments for advanced cancer stages found that a biennial consultation with medical TCE was cost-effective in the FDRs of CM patients [13]. Therefore, patients with newly diagnosed melanoma are advised to inform their FDRs of the increased risk to access counselling about the usefulness of protecting their skin against sun exposure and proceeding to SSE, and to consult a dermatologist or a GP annually for TCE.

The best way to deliver these messages to induce behavioral changes from FDRs remains uncertain, as well as the psychological determinants of adherence to prevention messages [14–17]. In 2012, 49% of dermatologists surveyed in the United States considered the lack of written support for FDRs a communication barrier [18]. A pilot study performed in our center suggested that a tip sheet delivered to CM patients for their FDRs which promoted sun protection behavior and medical TCE, could increase medical TCE in FDRs. We found a significant increase in FDRs' medical TCE (50.5% without vs. 87.5% with the tip sheet, $p = 0.006$), but the TCE of FDRs was patient-reported and thus subject to misclassification and social desirability bias [19].

Here, we report the results of a cluster-randomized trial aiming to assess the impact on FDRs of the delivery to CM patients of a tip sheet in addition to usual oral advice as compared to usual oral advice.

2. Materials and Methods

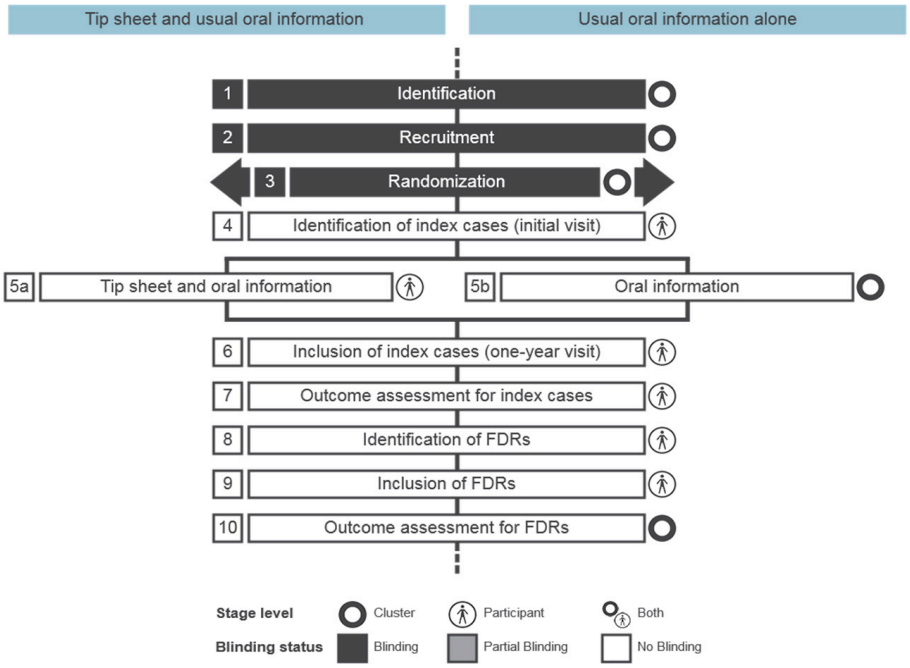
2.1. Design

This was a superiority, two-parallel groups, cluster randomized controlled trial involving nine French hospital centers (six based in a university and three general) used as randomization units to limit contamination bias.

2.2. Setting, Participants and Data Collection

The study was conducted between January 2014 and May 2019; participants were enrolled from 13 December 2017 to 26 April 2019. Patients with CM (index cases) were informed of the study one year after the initial diagnosis of their melanoma and were asked to participate.

Figure 1 outlines the trial processes depicting order of cluster recruitment, randomization, information delivery to index cases, then identification and outcome assessments for both index cases and their FDRs; blinding status is indicated using black for complete blinding and white for no blinding [20].



1 Identification
Nine centers of western region were invited to participate to the study. Six university hospital centers (Angers, Brest, Nantes, Poitiers, Rennes and Tours) and 3 general hospital centers (Le Mans, La Rochelle and Rennes)

2 Recruitment
The heads of dermatology departments their oral and written consent to participate in the study. The 9 centers agreed to participate in the study.

3 Randomization
Hospital centers were randomly allocated to written and oral information or oral information only in a 1:1 ratio. Allocation schedule was independently performed by a statistician using a computer-generated randomization list. Randomization was stratified on the type of hospitals: university or general.

4 Identification of index cases (initial visit)
All melanoma patients seen in consultation at one of the centers participating in the study during the 9 months following the start of the trial were eligible and were registered to be included at their one-year follow-up visit. No blinding for identification.

5a Tip sheet and oral information
In the intervention group, dermatologists were asked to provide all index cases both usual oral information and a tip sheet intended for their first degree relatives (FDRs) (one tip sheet per FDR). Oral information delivered by dermatologists to patients was usual information including sun-protection and early detection instructions for themselves and their FDRs. No blinding for intervention delivery.

5b Oral information
In the control group, dermatologists were asked to provide all index cases will receive oral information alone. Oral information delivered by dermatologists to patients was usual information including sun-protection and early detection instructions for themselves and their FDRs. Dermatologists from centers randomized to the control group did not have access to the tip sheet developed for the trial. No blinding for intervention delivery.

6 Inclusion of index cases (one-year visit)
At their one-year follow-up visit, all melanoma patients identified as index case during their initial visit were offered to participate in the study. Their non opposition was collected. No blinding for inclusion.

7 Outcome assessment for index cases
We collected data from included melanoma patients including the contact details of their FDRs. No blinding for data collection.

8 Identification of FDRs
FDRs from whom melanoma patients provided contacts were contacted by telephone or email in case of absence of a telephone response. No blinding for identification.

9 Inclusion of FDRs
After expressing their non-opposition, FDRs wer included. In both groups, a non-opposition sheet was mailed to the FDRs (in case they wanted to opt out from the study). No blinding for inclusion.

10 Outcome assessment for FDRs
Primary and secondary outcomes were collected by phone by the clinical research assistant. In the absence of a telephone response, FDRs were contacted by email and in the absence of a response to this email or in the absence of an email address, by post. No blinding for outcome assessment.

Figure 1. Timeline cluster diagram for the trial.

During one of the first initial management visits (initial resection or wide surgery or first follow-up visit), index cases received information on the usefulness of early detection of CM and sun protection behavior for their FDRs: oral and written information in the intervention group (Figure S1) and oral information alone in the control group.

During the follow-up visit one year later, index cases were informed of the study and their non-opposition was recorded. Then, they provided contact details for their FDRs. Questionnaires and scales were completed by the index cases to assess, in particular, the actual distribution of the tip sheet to their FDRs. FDRs were then contacted by telephone. In the absence of a telephone response, FDRs were contacted by email, and in the absence of a response to this email or in the absence of an email address, by mail. After expressing their non-opposition, the data were collected from FDRs on early detection of melanoma and sun-protection behavior. In both the intervention and control groups, an information letter and non-opposition sheet, were mailed to the FDRs.

2.3. Selection Criteria for CM Patients and FDRs

Inclusion criteria for CM index cases were age $\geq$18 years, speaking French, receiving treatment for stage I or II primary cutaneous CM in one of the participating hospitals and consulting at the time of diagnosis (<3 months after the excision). Patients had to have at least one FDR. We excluded patients with mucous or ocular melanoma, who had no FDR, or who did not wish to communicate information concerning melanoma to their FDRs.

Inclusion criteria for FDRs were age $\geq$18 years, speaking French and first-degree related (brothers and sisters, children, father and mother) to an included index case.

2.4. Randomization, Intervention and Control Procedures

Hospital centers were randomly allocated to disseminate written and oral information or oral information only in a 1:1 ratio. The allocation schedule was independently created by a statistician using a computer-generated randomization list. Randomization was stratified on the type of hospital, be it university or general. This was an unblinded study.

In the intervention group, index cases received usual oral information plus a tip sheet for FDRs directly explaining to FDRs they are at increased risk of CM, should perform SSE and TCE and change their behavior regarding sun exposure by seeking shade, wearing a hat and long-sleeved clothes and using sun protection creams. The tip sheet was developed before the beginning of the trial by the medical staff and then submitted to CM patients and some of their relatives in two separate focus-group sessions lasting 2 h each to test their comprehension. The tip sheet was revised according to their feedback. The final version of the tip sheet (used in this trial) is provided in the supplementary file.

In the control group, the index cases received the usual oral information on the fact that their FDRs that they are at increased risk of CM and should perform SSE and TCE and change their behavior regarding sun exposure by seeking shade, wearing a hat and long-sleeved clothes and using sun protection creams.

The Index Cases Received Oral or Oral and Written Advice at Time T, and the Relatives were Questioned in the Month following Inclusion, at T + 12 Months.

2.5. Outcomes

The primary outcome was the FDRs' participation in CM early detection by completing TCE with a dermatologist or a general practitioner performed within one year after the advice (written and oral or oral alone) had been delivered to the index cases.

Secondary outcomes were the FDRs' planned TCE with a dermatologist or a GP with and without a scheduled appointment, FDRs' participation in CM early detection by SSE, frequency of SSE (i.e., >one per month, one or two per year, or never) for FDRs, who declared having performed SSE and sun-protection behavior self-reported by FDRs with a questionnaire. We defined strict sun protection behavior according to the following criteria: use of high-factor sunscreen, wearing of sun-protective clothing, and/or avoidance of sun exposure.

The outcomes for FDRs were assessed during a phone call by a clinical research assistant previously trained to perform data collection. Calls were centralized in order to standardize outcome assessment. To assess the reliability of TCE self-reported by the FDRs without overburdening the trial, a random sample of 50 FDRs (25 in the control group and 25 in the intervention group) declaring to have had TCE was selected, and their physicians (dermatologist or GP) were contacted by telephone to ensure that they actually had a TCE consultation for the early detection of melanoma on the date they reported.

The outcomes for index cases were the transmission of the oral and written information to their FDRs.

2.6. Statistical Analysis

In our sample size calculation, we accounted for clustering at the family level rather than at the hospital level (randomization unit) because clustering was expected to be higher within families than within hospital centers. According to the data from our pilot study, we hypothesized that 60% of the FDRs in the control group will be performed a medical TCE, with an intraclass correlation coefficient at the family level of 0.40. To detect a 15% point increase in TCE, with 90% power and a 5% two-sided significance level, we needed a sample of 900 FDRs. We expected a mean number of four FDRs per CM patient and so planned to include 225 CM patients in our trial.

The baseline characteristics of index cases and FDRs are described by randomization group, with a mean (SD) or median (Q1;Q3) for continuous variables, depending on their distribution, and number (%) for categorical variables. No statistical test was performed on baseline variables.

Analysis of the primary outcome and secondary binary outcomes involved generalized estimating equation methods [21]. We used an exchangeable correlation matrix and robust variance estimators to account for clustering at the family level [22]. We used logistic regression models to estimate odds ratios (OR) with 95% confidence intervals (CIs).

In the main analysis of the primary outcome, the FDRs who performed a TCE before the initial visit of their index case and FDRs who reported a TCE but for whom the date of completion of TCE was unknown were considered not to have completed TCE within the timeframe provided in the protocol. We performed a post-hoc pragmatic sensitivity analysis considering that these FDRs had indeed a TCE (even if this was before the initial visit of the index case or at an unknown date). ICCs for the primary outcome were estimated, per randomization group, using an ANOVA estimator [23]. We estimated clustering at both hospital and family levels.

We also described the median time from the initial consultation of the index case to TCE for FDRs. No statistical test was performed on this outcome, which was not initially in the protocol. The validity of TCE self-declaration was assessed in a random sample of 50 FDRs (25 in the control group and 25 in the intervention group) who reported TCE. We describe the number of false positives (i.e., participants who reported a TCE and for whom the physician did not confirm the TCE) and the positive predictive value.

The analysis was conducted according to the intention-to-treat principle. Randomized index cases were analyzed in the group to which they were allocated. FDRs were also analyzed in the randomization group in which the index case to which they were attached was allocated. $p < 0.05$ was considered statistically significant. All analyses were performed with SAS 9.4. (SAS Institute Inc., Cary, NC, USA).

2.7. Ethics

Ethics review and approval were provided by the research ethics committee of CHRU Tours (no. 2016-s14), and the trial was registered at ClinicalTrials.gov (accessed on 19 June 2022) (NCT02917473).

3. Results

Among the nine centers, five were allocated to the intervention group (2 general hospital centers and three university hospital centers), with 75 included index cases, and four were allocated to the control group (one general hospital centers and three university hospital centers), with 66 included index cases (Figure 2). Two index cases were excluded from the intervention group because they opposed the analysis of their personal data. In the intervention group, a total of 196 FDRs were contacted, and 27 were excluded because they were not reachable or were ineligible. In the control group, five index cases were excluded (four opposed the analysis of their personal data, and one did not wish to communicate information concerning his melanoma to his relatives). In the control group, 140 FDRs were contacted, and 24 were excluded because they were not reachable or were ineligible. Finally, we included 60 index cases with 166 eligible FDRs in the intervention group and 48 index cases with 114 eligible FDRs in the control group.

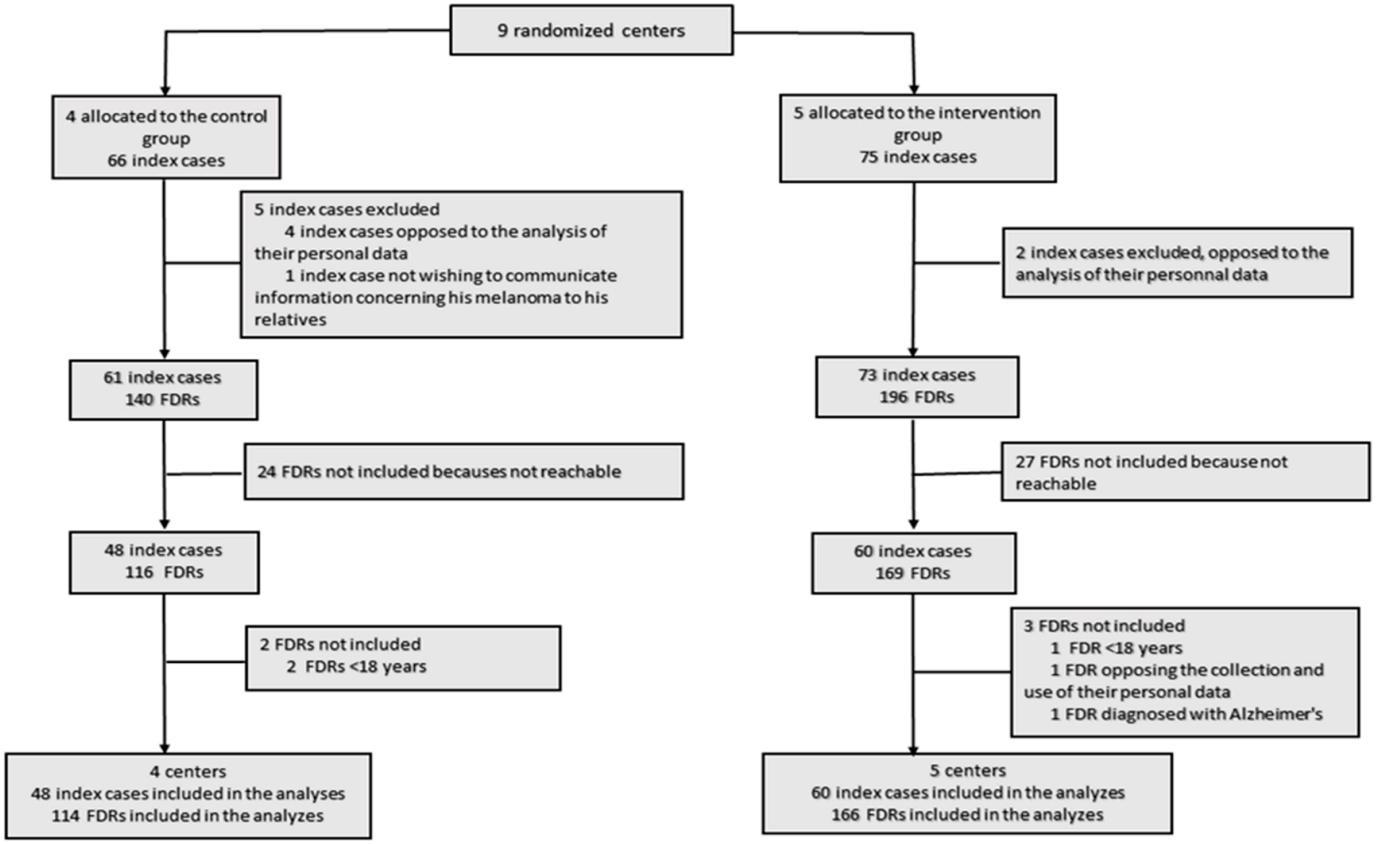


Figure 2. The flow of clusters, index cases and first-degree relatives (FDRs) in the study.

3.1. Study Population

The characteristics of FDRs at inclusion are summarized, by the randomization group, in Table 1. The characteristics of index cases at inclusion were similar in both groups and are summarized in Table S1.

Table 1. Socio-demographic characteristics and melanoma risk factors for first-degree relative (FDRs) at inclusion by randomization group.

	Intervention Group Written and Oral Information n = 166	**Control Group Oral Information Alone n = 114**
FDR characteristics		
Age, n_i = 165, n_c = 114	52 (17)	50 (18)
Men, n_i = 166, n_c = 114	75 (45.2)	45 (39.5)
Level of education, n_i = 166, n_c = 114		
Secondary school	24 (14.5)	15 (13.2)
Certificate of professional competence/Professional study certificate	34 (20.5)	26 (22.8)
High school diploma	43 (25.9)	38 (33.3)
Bachelor's degree	32 (19.3)	16 (14.0)
Master's degree	32 (19.3)	18 (15.8)
Other	1 (0.6)	1 (0.9)
Occupational status, n_i = 166, n_c = 114		
Full-time professional	95 (57.2)	59 (51.8)
Part-time professional	10 (6.0)	7 (6.1)
No activity	1 (0.6)	0 (0.0)
Work time accident/occupational disease	2 (1.2)	0 (0.0)
Student	6 (3.6)	7 (6.1)
Disability	1 (0.6)	0 (0.0)
Unemployment	0 (0.0)	3 (2.6)
Retired	49 (29.5)	36 (31.6)
Other	2 (1.2)	2 (1.8)
Business, n_i = 165, n_c = 113		
Farmer	2 (1.2)	0 (0.0)
Artisan, trader, head of enterprise	10 (6.1)	6 (5.3)
Executive, higher intellectual profession	25 (15.2)	17 (15.0)
Intermediate occupation	11 (6.7)	2 (1.8)
Employed	58 (35.2)	38 (33.6)
Worker	1 (0.6)	3 (2.7)
Other	58 (35.2)	47 (41.6)
Relationship status to index case, n_i = 165, n_c = 114		
Father or mother	20 (12.1)	17 (14.9)
Brother or sister	69 (41.8)	41 (36.0)
Son or daughter	76 (46.1)	56 (49.1)
Family situation, n_i = 165, n_c = 114		
Married	113 (68.5)	76 (66.7)
Single	36 (21.8)	19 (16.7)
Widow (er)	6 (3.6)	5 (4.4)
Separated/divorced	10 (6.1)	14 (12.3)
FDRs with children, n_i =165, n_c = 114	117 (70.9)	80 (70.2)
Number of children, n_i = 117, n_c = 80	2 (2; 3)	2 (2; 3)
History of melanoma, n_i = 161, n_c = 112	6 (3.7)	3 (2.6)
History of another cancer, n_i = 163, n_c = 113	12 (7.4)	10 (8.8)
Skin cancer *, n_i = 12, n_c = 10	5 (41.7)	6 (60.0)
Other type of cancer *, n_i = 12, n_c = 10	7 (58.3)	5 (50.0)

Data are n (%) or mean (SD) or median [Q1; Q3]; * a patient could have several histories of cancer; n_i, intervention group; n_c, control group.

3.2. Information Transmission as Reported by Index Cases

Most index cases declared having informed their FDRs about their melanoma (60/60 in the intervention group and 43/48 in the control group). In the intervention group, 59 of 60 index cases and 34 of 48 index cases in the control group declared having transmitted oral advice to FDRs. Overall, 46 of 60 index cases in the intervention group and 8 of 48 in the control arm declared having transmitted a written tip sheet to FDRs.

3.3. Information Transmission as Reported by FDRs

A total of 55 of 166 FDRs in the intervention group and 4 of 114 in the control group reported that they received a written tip sheet from the index case. In addition, 121 of 166 FDRs in the intervention group and 72 of 114 in the control group declared that oral advice was transmitted by the index case to FDRs.

3.4. Primary Outcome

In the intervention group, 60 of 166 FDRs (36.1%) reported having had a TCE by a dermatologist and/or a GP at a date after the initial consultation of the index case as compared with 45 of 114 FDRs (39.5%) in the control group (OR 0.9 [95% CI 0.5 to 1.5], $p = 0.63$) (Table 2). Sensitivity analysis accounting for all reported TCEs regardless of their date confirmed our primary results (71/166 [42.8%] in the intervention group vs. 56/114 [49.1%] in the control group) (OR 0.8 [95% CI 0.5 to 1.3], $p = 0.39$). The median time from the index case initial visit to the FDRs' TCE was 226 days (Q1; Q3, 129; 319) in the intervention group and 250 days (129; 362) in the control group. The ICCs for reported TCE in the intervention and control groups were 0.009 and 0.06 at the hospital center level and 0.12 and 0.15 at the family level.

Table 2. First-degree relatives' total cutaneous examination (TCE) performed by a dermatologist and/or a general practitioner for the early detection of melanoma by randomized group made an appointment or planning to make an appointment for an early detection examination.

	Intervention Group Written and Oral Information n = 166	Control Group Oral Information n = 114	OR [95% CI]	*p*-Value
Primary outcome: TCE performed strictly within the period of the study *	60 (36.1%)	45 (39.5%)	0.9 [0.5 to 1.5]	0.63
Sensitivity analysis of the primary outcome: all TCEs performed, including TCE performed just before the inclusion period or with a missing date of appointment	71 (42.8%)	56 (49.1%)	0.8 [0.5 to 1.3]	0.39
Secondary outcomes				
TCE planned with a scheduled appointment but not yet completed	8 (4.8%)	1 (0.9%)	5.5 [0.7 to 45.3]	0.11
TCE planned with no scheduled appointment yet	36 (21.7%)	24 (21.1%)	1.1 [0.6 to 1.9]	0.88

OR—odds ratio; 95% CI, 95% confidence interval; * In this analysis, patients who had a screening examination at an unknown date or before the initial visit of their index case were considered as not having had a TCE (i.e., as a "failure" for the primary outcome).

TCE was performed mainly by dermatologists (65/70 [92.9%] in the intervention group vs. 51/56 [91.1%] in the control group) and rarely by GPs (4/70 vs. 3/56) or by both practitioners (1/70 vs. 2/56).

Among the 127 FDRs who had a TCE with a GP or a dermatologist, 29 had a lesion removed after this consultation: 20 of 71 in the intervention group and 9 of 56 in the control group. Three melanoma cases were detected in the intervention group and one in the control group.

Among the 50 FDRs in the randomly selected subsample for validity evaluation (25 in the control group and 25 in the intervention group) declaring to have had a TCE, the self-declaration validity was assessable in 39 cases and was confirmed in all 39 cases (17 in the control group and 22 in the intervention group). Self-declaration validity was not assessable in 11 FDRs because their physicians could not be reached.

3.5. Secondary Outcomes

In the intervention group, 8 of 166 FDRs had a planned TCE with a scheduled appointment compared to 1 of 114 in the control group (OR 5.5 [95% CI 0.7 to 45.3], $p = 0.39$). Moreover, 36/166 FDRs in the intervention group versus 24/114 in the control group planned to perform a medical TCE without any scheduled appointment (OR 1.1 [95% CI 0.6 to 1.9], $p = 0.88$) (Table 2).

SSE was reported by 110 of 166 FDRs (66.3%) in the intervention group versus 69 of 113 (61.1%) in the control group (OR 1.3 [95% CI 0.8 to 2.2], $p = 0.4$). FDRs in the two groups did not differ according to their frequency of performing SSE (i.e., >one per month, one or two per year, or never).

A total of 121 of 166 FDRs in the intervention group and 88 of 114 in the control group reported sun-protection behavior (OR 0.8 [95% CI 0.5 to 1.4], $p = 0.48$). Components of sun-protection behaviors of the participants are summarized in Table 3. There were no participants that differed according to the randomization group.

Table 3. Sun protection behaviors of first-degree relatives.

	Intervention Group Written and Oral Information n = 166	Control Group Oral Information n = 114	OR [95% CI]	p-Value
Use of sun protection	121 (72.9%)	88 (77.2%)	0.8 [0.5 to 1.4]	0.48
Avoidance of sun exposure	39 (23.5%)	22 (19.3%)	1.4 [0.7 to 2.7]	0.40
Wearing protective clothes	70 (57.9%)	38 (43.2%)	1.5 [0.9 to 3.3]	0.13
Use of high index sunscreen	105 (63.3%)	70 (61.4%)	1.1 [0.6 to 1.9]	0.73

OR—odds ratio; 95% CI—95% confidence interval.

4. Discussion

Our study was designed to be pragmatic and, in the case of a positive result, easily generalizable to the daily practice of dermatologists. We hypothesized that a written tip sheet given to CM patients for their FDRs, in addition to the usual oral information, would increase CM early detection in FDRs. We did not see evidence of improved engagement of FDRs in TCE recommended in the written and oral information as compared with oral information alone. Our results were consistent for the secondary outcomes, with no difference in SSE and sun protection behavior between both groups.

4.1. Primary and Secondary Outcomes

For the primary outcome, we found that giving a tip sheet to patients with melanoma for the FDRs did not increase the proportion of medical TCEs in FDRs as compared with the usual oral advice in the control group.

Less than 50% of FDRs in both groups actually had a TCE performed by a dermatologist or GP in the year after the removal of the melanoma in their relative. The low rate of TCE observed in both groups of FDRs is consistent with a study conducted from 2001 to 2003 in the United States [24]. The authors found that 45% of FDRs never had a TCE and that 13.4% had not had a TCE within the past three years. In the same study, 28% of FDRs in the intervention group versus 39% in the control group had not performed SSE in the past year. This proportion is close to 34% in the intervention group in our study. The frequency of SSE among FDRs varies from 30.8% to 71.6%, depending on the study [24,25].

Three previously published randomized trials conducted in the United States found improvements in either TCE or SSE with a tailored intervention directed to FDRs as compared to usual care [26–28]. The interventions assessed in these trials all involved repeated messages dedicated directly to FDRs.

Indeed, our intervention was more pragmatic than those previously evaluated because it consists of a unique message and relies on patients rather than FDRs. However, because our intervention did not directly target the FDRs, to be successful, it required that first the

patient transmit the tip sheet to their FDRs and then the FDRs adhere to the screening and sun protection behaviors. Thus, some elements may explain the failure of our intervention.

First, the rate of FDRs who declared having received the tip sheet was low: 33.1% in the intervention group versus 3.5% in the control group. These results are also in favor of a small between-group contamination or a desirability bias from FDRs in the control group. This low rate of transmission of the tip sheet can be explained by the evolution of what primarily worries the most patients and their families after the announcement of the diagnosis for the patients: at first, family conversations typically focus on the patient, including the diagnosis and the treatment, and after the acute treatment phase, conversations about family risk and prevention predominate [29]. Hence, the distribution of the tip sheet close to the diagnosis stage may have occurred too early. During this second step, melanoma patients may not inform their FDRs because of the perception of low or high risk of occurrence of melanoma in their FDRs according to skin color, ability to sunburn and age [30]. Hence, tailored and repeated interventions should increase the rate of information transmitted to FDRs. During this step, the tip sheet can be lost, forgotten, or difficult to transmit because relatives and index cases do not live in the same city, for example. Thus, the paper support may not be the most suitable support. In addition, FDRs may directly search the Internet for information and find the tip sheet of little use. Indeed, in our study, 72.5% and 64.3% of FDRs in the intervention and control groups, respectively, declared having frequently sought information on the Internet. Moreover, the information on skin cancers, sun protection behaviors, and family risk of cancers has been widely diffused in the media in France, with a "Skin Cancer Day" conducted every year. Accordingly, many more patients and their families are now better informed compared to the early 2000s.

Second, knowing the risks of skin cancer (or other cancers) does not imply engagement in a screening program and reduction or avoidance of risk factors due to an addiction to known carcinogens (e.g., tobacco use or sun exposure) [31].

4.2. Strengths and Limitations

The main strength of our study is its pragmatic design. Because dermatologists have underlined the absence of standardized guidelines and a lack of written material as moderate barriers to melanoma communication, we aimed to design a simple and low-cost intervention to improve sun protection and early detection of melanoma in FDRs. Moreover, the tip sheet was established with the help of two focus groups conducted by a psychologist and a dermatologist with both patients and FDRs. Cluster randomization is also a strength of the trial because it allows all the members of a given family to be in the same randomization group. It could have created a family effect on screening behavior. Another strength of the study is that the main outcome was not only self-declared because we verified that a TCE was indeed performed by calling dermatologists and GPs for a random sample of 50 FDRs who declared TCE.

The main limitation is the lack of power. We planned to include 900 FDRs and finally included 279. Indeed, we overestimated the expected number of FDRs: we estimated the number of FDRs per CM patient to be four, whereas patients reported only 2.6 FDRs, on average, which is the same as reported by Manne et al. [27]. Because of its design, it was not possible to extend the inclusion period of the trial to increase the sample size. Another limitation is that the advice sheet was based on text only, without any visual support such as drawings or pictures [32]. Moreover, after the tip sheet was established with the help of two focus groups, the relevance of the information was not evaluated in individual cognitive interviews with FDRs or further qualitative assessment with Likert scales to assess the value of the content of the tip sheet. Finally, the perceived severity of skin cancer/melanoma assessed among the FDRs as well other psychological factors and the health professional's ability to provide clear explanations [33] may influence the impact of the intervention.

5. Conclusions

Our trial did not provide evidence that a tip sheet added to the usual oral information delivered to melanoma patients for their FDRs improved screening or sun-protection behaviors. Such behavior modification likely requires direct and more intensive personalized and repeated interventions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14163864/s1, Supplementary Figure S1: The French and English version of the tip sheet given to patients intended for their first-degree relatives; Supplementary Table S1: Socio-demographic characteristics and melanoma risk factors of index cases at inclusion by randomization group.

Author Contributions: Study conception and funding acquisition were led by L.M., M.B. and A.C. Intervention delivery and investigation performed by G.Q., A.D., T.L., Y.L.C., E.W.-H., D.L., P.C., H.M. and L.M. Analysis was supervised by A.C. and performed by F.L.V.-A. The first draft of the manuscript was written by D.M., A.C. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the French Cancer National Institute (Institut National du Cancer, INCA_8186).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the research ethics committee of CHRU Tours (no. 2016-s14).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data used in the study are available from the corresponding author upon request.

Acknowledgments: To the patients and their first-degree relatives who took part in the study, to Anne Rebion for data management and to Céleste Verne and Stéphanie Bonte for administrative and technical support.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020, GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Dyba, T.; Randi, G.; Bray, F.; Martos, C.; Giusti, F.; Nicholson, N.; Gavin, A.; Flego, M.; Neamtiu, L.; Dimitrova, N.; et al. The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *Eur. J. Cancer* **2021**, *157*, 308–347. [CrossRef]
3. Swetter, S.M.; Pollitt, R.A.; Johnson, T.M.; Brooks, D.R.; Geller, A.C. Behavioral determinants of successful early melanoma detection. *Cancer* **2012**, *118*, 3725–3734. [CrossRef]
4. Kovalyshyn, I.; Dusza, S.W.; Siamas, K.; Halpern, A.C.; Argenziano, G.; Marghoob, A.A. The impact of physician screening on melanoma setection. *Arch. Dermatol.* **2011**, *147*, 1269–1275. [CrossRef] [PubMed]
5. Dessinioti, C.; Geller, A.C.; Stergiopoulou, A.; Swetter, S.M.; Baltas, E.; Mayer, J.E.; Johnson, T.M.; Talaganis, J.; Trakatelli, M.; Tsoutsos, D.; et al. Association of skin examination behaviors and thinner nodular vs superficial spreading melanoma at diagnosis. *JAMA Dermatol.* **2018**, *154*, 544–553. [CrossRef] [PubMed]
6. Wernli, K.J.; Henrikson, N.B.; Morrison, C.C.; Nguyen, M.; Pocobelli, G.; Blasi, P.R. Screening for skin cancer in adults: Updated evidence report and systematic review for the US preventive services task force. *JAMA* **2016**, *316*, 436–447. [CrossRef] [PubMed]
7. Johansson, M.; Brodersen, J.; Gøtzsche, P.C.; Jørgensen, K.J. Screening for reducing morbidity and mortality in malignant melanoma. *Cochrane Database Syst. Rev.* **2019**, *6*, CD012352. [CrossRef]
8. Buja, A.; Sartor, G.; Scioni, M.; Vecchiato, A.; Bolzan, M.; Rebba, V.; Sileni, V.C.; Palozzo, A.C.; Montesco, M.; Del Fiore, P.; et al. Estimation of direct melanoma-related costs by disease stage and by phase of diagnosis and treatment according to clinical guidelines. *Acta Derm. Venerol.* **2018**, *98*, 218–224. [CrossRef]
9. Olsen, C.M.; Carroll, H.J.; Whiteman, D.C. Familial melanoma: A meta-analysis and estimates of attributable fraction. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 65–73. [CrossRef]

10. Ford, D.; Bliss, J.M.; Swerdlow, A.J.; Armstrong, B.K.; Franceschi, S.; Green, A.; Holly, E.A.; Mack, T.; MacKie, R.M.; Osterlind, A. Risk of cutaneous melanoma associated with a family history of the disease.The Interventional Melanoma Analysis Group (IMAGE). *Int. J. Cancer* **1995**, *62*, 377–381. [CrossRef]
11. Rhodes, A.R.; Weinstock, M.A.; Fitzpatrick, T.B.; Mihm, M.C., Jr.; Sober, A.J. Risk factors for cutaneous melanoma. A practical method of recognizing predisposed individuals. *JAMA* **1987**, *258*, 3146–3154. [CrossRef]
12. ALD n° 30. Guide Médecin sur le Mélanome Cutané. Available online: https://www.has-sante.fr/upload/docs/application/pdf/201203/ald_30_guide_melanome_web.pdf (accessed on 20 April 2022).
13. Losina, E.; Walensky, R.P.; Geller, A.; Beddingfield, F.C., 3rd; Wolf, L.L.; Gilchrest, B.A.; Freedberg, K.A. Visual screening for malignant melanoma, a cost-effectiveness analysis. *Arch. Dermatol.* **2007**, *143*, 21–28. [CrossRef]
14. Coroiu, A.; Moran, C.; Bergeron, C.; Thombs, B.D.; Geller, A.C.; Kingsland, E.; Körner, A. Operationalization of skin self-examination in randomized controlled trials with individuals at increased risk for melanoma: A systematic review. *Patient Educ. Couns.* **2020**, *103*, 1013–1026. [CrossRef]
15. Bowen, D.J.; Hay, J.L.; Mayer, J.; Kuniyuki, A.; Meischke, H.; Harris, J.; Asgari, M.; Shoveller, J.; Press, N.; Burke, W. Predictors of recruited melanoma families into a behavioral intervention project. *Contemp. Clin. Trials* **2012**, *33*, 85–92. [CrossRef]
16. Bowen, D.J.; Albrecht, T.; Hay, J.; Eggly, S.; Harris-Wei, J.; Meischke, H.; Burke, W. Communication among melanoma family members. *J. Health Commun.* **2017**, *22*, 198–204. [CrossRef]
17. Hay, J.; Shuk, E.; Brady, M.S.; Berwick, M.; Ostroff, J.; Halpern, A. Family communication after melanoma diagnosis. *Arch. Dermatol.* **2008**, *144*, 553–554. [CrossRef]
18. Oliveria, S.A.; Heneghan, M.K.; Halpern, A.C.; Hay, J.L.; Geller, A.C. Communication about family members' risk of melanoma: Self-reported practices of dermatologists in the United States. *Arch. Dermatol.* **2012**, *148*, 621–627. [CrossRef]
19. Berton, M.; Perrinaud, A.; Samimi, M.; Delaplace, M.; Vaillant, L.; Machet, L. Utilité d'une fiche de conseils pour la photoprotection et le dépistage des patients et de leur famille après un diagnostic de mélanome. *Ann. Dermatol. Venereol.* **2013**, *140*, S407–S408. [CrossRef]
20. Caille, A.; Kerry, S.; Tavernier, E.; Leyrat, C.; Eldridge, S.; Giraudeau, B. Timeline cluster: A graphical tool to identify risk of bias in cluster randomised trials. *BMJ* **2016**, *354*, i4291. [CrossRef]
21. Zeger, S.L.; Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 124–130. [CrossRef]
22. Eldridge, S.; Kerry, S.M. *A Practical Guide to Cluster Randomised Trials in Health Services Research*; John Wiley & Sons: Chichester, UK, 2011.
23. Ridout, M.S.; Demetrio, C.G.B.; Firth, D. Estimating Intraclass Correlation for Binary Data. *Biometrics* **1999**, *55*, 137–148. [CrossRef]
24. Manne, S.; Fasanella, N.; Connors, J.; Floyd, B.; Wang, H.; Lessin, S. Sun protection and skin surveillance practices among relatives of patients with malignant melanoma: Prevalence and predictors. *Prev. Med.* **2004**, *39*, 36–47. [CrossRef]
25. Sarikaya Solak, S.; Yondem, H.; Cicin, I. Evaluating sun protection behaviors and skin self-examination practices among the family members of melanoma patients in Turkey: A cross-sectional survey study. *Dermatol. Ther.* **2020**, *33*, e14268. [CrossRef]
26. Geller, A.C.; Emmons, K.M.; Brooks, D.R.; Powers, C.; Zhang, Z.; Koh, H.K.; Heeren, T.; Sober, A.J.; Li, F.; Gilchrest, B.A. A randomized trial to improve early detection and prevention practices among siblings of melanoma patients. *Cancer* **2006**, *107*, 806–814. [CrossRef]
27. Manne, S.; Jacobsen, P.B.; Ming, M.E.; Winkel, G.; Dessureault, S.; Lessin, S.R. Tailored versus generic interventions for skin cancer risk reduction for family members of melanoma patients. *Health Psychol.* **2010**, *29*, 583–593. [CrossRef]
28. Bowen, D.J.; Hay, J.; Meischke, H.; Mayer, J.A.; Harris-Wai, J.; Burke, W. Randomized trial of a web-based survivor intervention on melanoma prevention behaviors of first-degree relatives. *Cancer Causes Control* **2019**, *30*, 225–233. [CrossRef]
29. Rodríguez, V.M.; Berwick, M.; Hay, J.L. Communication about melanoma and risk reduction after melanoma diagnosis. *Psycho-Oncology* **2017**, *26*, 2142–2148. [CrossRef]
30. Oliveria, S.A.; Shuk, E.; Hay, J.L.; Heneghan, M.; Goulart, J.M.; Panageas, K.; Geller, A.C.; Halpern, A.C. Melanoma survivors, health behaviors, surveillance, psychosocial factors, and family concerns. *Psycho-Oncology* **2013**, *22*, 106–116. [CrossRef]
31. Loescher, L.J.; Crist, J.D.; Siaki, L.A. Perceived intrafamily melanoma risk communication. *Cancer Nurs.* **2009**, *32*, 203–210. [CrossRef]
32. McWhirter, J.E.; Hoffman-Goetz, L. Visual images for patient skin self-examination and melanoma detection: A systematic review of published studies. *J. Am. Acad. Dermatol.* **2013**, *69*, 47–55. [CrossRef]
33. Loescher, L.J.; Crist, J.D.; Cranmer, L.; Curiel-Lewandrowski, C.; Warneke, J.A. Melanoma high-risk families' perceived health care provider risk communication. *J. Cancer Educ.* **2009**, *24*, 301–307. [CrossRef] [PubMed]

Annexe C

Difficultés de recrutement et d'implémentation de l'intervention dans les ERC en *"stepped-wedge"* : Article

ELSEVIER

Journal of Clinical Epidemiology 148 (2022) 93−103

Journal of Clinical Epidemiology

REVIEW

Recruitment and implementation challenges were common in stepped-wedge cluster randomized trials: Results from a methodological review

Agnes Caille[a,b,c,*], Monica Taljaard[d,e], Floriane Le Vilain—Abraham[b,c], Alexis Le Moigne[a], Andrew J. Copas[f], Florence Tubach[a,1], Agnes Dechartres[a,1]

[a]*Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié-Salpêtrière, Département de Santé Publique, Paris, France*

[b]*Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France*

[c]*INSERM CIC 1415, CHRU de Tours, Tours, France*

[d]*Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada*

[e]*School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada*

[f]*MRC Clinical Trials Unit at University College London, London, United Kingdom*

Accepted 20 April 2022; Published online 26 April 2022

Abstract

Objective: To explore challenges in recruitment and intervention implementation in recent stepped-wedge cluster randomized trials (SW-CRTs).

Study Design and Setting: We searched PubMed to identify primary reports of SW-CRTs (2019−2020). Two reviewers independently screened studies and extracted data from each report. A recruitment challenge was defined as a planned number of clusters or participants not achieved or any reported changes made to the design to address recruitment difficulties. An implementation challenge was defined as early, late, or no implementation of the intervention in at least one cluster.

Results: Of 55 SW-CRTs, 18 (33%) had a recruitment challenge, 23 (42%) had none, and for 14 (26%) it was impossible to judge. At least one implementation challenge was present in 24 (44%), eight (15%) had none, and for 23 (42%) it was impossible to judge. Of the 35 (64%) trials with recruitment or implementation challenges, 18 (72%) had one or more modifications of their design, most often a modification of the trial duration.

Conclusion: Investigators must be aware of the risks of recruitment or implementation challenges when considering the use of an SW-CRT design. Mitigating strategies should be adopted when planning the trial. More transparent reporting of planned and actual design features is required. © 2022 Elsevier Inc. All rights reserved.

Keywords: Stepped-wedge cluster randomized trial; Recruitment; Implementation; Design; Methodological review; Reporting

1. Introduction

Cluster randomized trials (CRTs) are trials that randomize groups, such as hospitals, general practices, or geographical areas, rather than individuals [1]. The most common CRT design is the two-parallel group design in which randomization determines which clusters will receive the intervention and which will receive the control. A recent and increasingly used alternative CRT design is the stepped-wedge CRT (SW-CRT) in which all clusters receive the intervention by the end of the trial; clusters are randomized to sequences and these sequences determine the timing at which a cluster will start implementing the intervention [2]. In each cluster, measurements are repeatedly taken from initial time periods spent in the control condition and subsequent time periods spent in the intervention condition. Advantages over a classic two-parallel group CRT that have been highlighted in the literature include logistical benefits due to staggering of intervention implementation over time, enhanced recruitment because all clusters will receive the intervention

Declaration of Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors declare they have no competing interests.

[1] F.T. and A.D. contributed equally to this article.

* Corresponding author. Tel.: +44-33-2-34-37-96-54; fax: +44-33-2-47-47-46-39.

E-mail address: agnes.caille@med.univ-tours.fr (A. Caille).

https://doi.org/10.1016/j.jclinepi.2022.04.024

0895-4356/© 2022 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Recruitment and implementation challenges are frequent in stepped-wedge cluster randomized trials (SW-CRTs).

What this adds to what was known?

- Many SW-CRT reports do not permit an assessment of whether the planned schedule has been adhered to.
- Reasons for implementation challenges are heterogeneous; some can be avoided, others are unpredictable.

What is the implication and what should change now?

- The risks of recruitment and implementation challenges must be considered before initiating an SW-CRT. Once the decision has been made to adopt an SW-CRT, mitigating strategies should be put in place.
- Transparent reporting of planned and realized design features is essential for correct interpretation of the results.

before the end of the trial, and under certain conditions, better statistical efficiency [3]. Although appealing at the first view, SW-CRTs may be at a higher risk of bias than parallel CRTs [4] and require advanced statistical methods to account for underlying secular trends [5].

Designing an SW-CRT is complex and requires many elements to be specified in advance such as the number of sequences, clusters, and periods [6]. Once started, an SW-CRT is like a race against time; clusters must adhere to the planned schedule, that is, they must comply with the timing of implementation of the intervention and attain the target sample size in each period. However, several challenges in the implementation of the intervention and recruitment have been reported in SW-CRTs [7]. To reach the planned sample size or accommodate unanticipated problems in intervention delivery, changes in the trial design are sometimes decided during the trial such as an extension of trial duration or postponement of intervention implementation in some clusters, which can have implications for clusters yet to receive the intervention [8]. Although these challenges are not specific to the SW design, any changes in the timing of SW-CRTs affect control and intervention observations unevenly and may therefore alter the results and their interpretation. A previous methodological review of 35 SW-CRTs found that only 69% recruited their targeted number of participants and 43% reported difficulties during the study conduct such as a cluster dropout or delayed intervention [9]. Nevertheless, this review did not explore recruitment challenges in detail and to our knowledge, implementation challenges have never been systematically assessed in SW-CRTs.

The aim of this study was to describe recruitment and implementation challenges in recent SW-CRTs and to assess whether and how such challenges were accommodated in the trial design and analysis strategy. We also sought to investigate factors associated with recruitment or implementation challenges.

2. Methods

2.1. Search strategy

We searched MEDLINE via PubMed to identify eligible SW-CRTs. The search algorithm, implemented on September 23, 2020, was based on previously published electronic search strategies [9,10] and used several synonyms to describe the SW design (Supplementary Appendix A).

2.2. Eligibility criteria

We included full reports of SW-CRTs conducted in humans and published in English between January 1, 2019 and September 23, 2020. We restricted the search to this period to focus on trials published after the Consolidated Standards of Reporting Trials (CONSORT) extension for SW-CRTs in November 2018 [6]. We only included primary reports of completed trials; we excluded research letters, protocols, secondary or subgroup analysis papers, and methods papers. To qualify as an SW-CRT, the design had to use cluster randomization and have a minimum of two sequences and three periods. Nonrandomized or quasi-experimental designs and pilot or feasibility studies (as stated by the authors) were excluded. We also excluded designs randomizing fewer than five clusters, even if not described as pilot or feasibility studies, as inferences that can be drawn from such trials would be limited. Finally, SW-CRTs with more than one evaluated intervention were excluded as implementation issues become more complex in this case.

2.3. Selection of articles

Identified references were saved and managed using Zotero 5.0. Duplicates were removed. Two reviewers (A.C. and M.T.) independently screened the titles and abstracts of the identified references to assess eligibility. If necessary, full-text articles were searched and screened. Any reference not meeting the eligibility criteria was excluded and the reason for exclusion was recorded. Any discrepancies in the eligibility of a study were resolved by discussion, with

the help of a third reviewer (A.D.) if needed to reach a consensus.

2.4. Data extraction and management

Two reviewers independently extracted the data in a random computer-generated order (A.C. and one of A.L.M., F.L.V.A., and M.T.). We used a data extraction form tested and revised using four trials. Any discrepancies in data extraction were resolved by discussion between the two reviewers or with the help of a third reviewer to reach a consensus (A.C. and two of A.L.M., F.L.V.A., and M.T.). We attempted to access the protocol and any cited secondary analysis paper to collect complete information on the trials as planned and realized. When no protocol was publicly available, we emailed the corresponding author to request the protocol. We collected and managed extracted data using Airtable (Airtable, San Francisco, CA, USA) [11].

We extracted the following characteristics for each selected trial (definitions of methodological elements used in data extraction are provided in Table 1):

- General characteristics: journal, publication year, location of study recruitment, type of clusters, any reported rationale for the SW-CRT design, any reported prior pilot, or feasibility study;
- Design characteristics: timing of randomization (at a single time point, in batches, or unclear), type of SW-CRTs design (cross-sectional, closed cohort, or open cohort), whether there was a prospective recruitment of participants, source of outcome data collection (exclusively routinely collected data or not), planned and actual number of participants (or observations), number of sequences, number of clusters, number of periods, duration of the trial, complete or incomplete design, and allowance for a transition period;
- Intervention condition: typology of the intervention (targeted at the organization of healthcare or health delivery service, in healthcare professionals, direct participant therapeutic intervention, participant health promotion, or education intervention) and level of the intervention target (cluster, individual, or both levels);
- Control condition (usual care or other);
- Results on the primary outcome (positive or negative).

We classified a trial as having a recruitment challenge if (a) it did not reach its planned number of clusters, (b) it did not reach its planned number of participants with a 10% allowed margin (i.e., less than 90% of the target sample size recruited), and/or (c) it clearly reported that design changes were made in response to recruitment difficulties. We considered only the total number of participants because the number of participants per cluster-period was seldom reported. We defined an implementation challenge as any of (i) early, (ii) delayed, or (iii) no implementation of the intervention in one or more clusters. Implementation challenges were either self-reported by the authors or identified by the reviewers, especially by comparing the diagrams in the protocol and report of the trial. We defined a modification of the trial design as any of the deviation from (i) the planned number of sequences, (ii) the planned number of clusters, (iii) the planned number of periods, (iv) the planned duration (allowing for a 1-month margin), (v) a change from a complete to an incomplete design and vice-versa, or (vi) an addition or withdrawal of a transition period. We considered modifications that were made after the trial initiation in trials for which we identified recruitment or implementation challenges but also in other trials as such modifications can be related to unreported challenges. In case of identified recruitment or implementation challenges, we extracted whether a reason was reported and whether this challenge was accounted for in the analysis strategy (in the primary analysis or any sensitivity analysis).

2.5. Data analysis

Categorical data were summarized using frequency and percentage and quantitative data using mean and standard deviation or median and interquartile range, as appropriate. We described the characteristics of trials with and without a recruitment or implementation challenge without performing any statistical tests of association because of the small sample sizes. We considered the following trial characteristics that we thought a priori might be associated with challenges: a number and type of clusters, availability of a protocol, previous pilot study, trial design, number of participants, prospective recruitment, allowance for a transition period, type and level of experimental intervention, and method for data collection. We also explored whether recruitment or implementation challenges were associated with modifications of the planned design and with positive results for the primary outcome. Statistical analyses were performed with SAS, version 9.4 (SAS Institute Inc., Cary, NC, USA).

3. Results

3.1. Search results

Among the 562 references identified from PubMed, three were duplicates and 440 were excluded as ineligible based on the title and abstract. Of the 119 references assessed on full-text, we excluded 64 that did not meet our eligibility criteria (Fig. 1).

3.2. Characteristics of included studies

The 55 included SW-CRTs were published in 42 journals with an impact factor ranging from 1.85 to 79.32; their general characteristics are reported in Table 2.

A rationale for the SW design was provided in 47 studies (86%). Notably, logistical reasons were reported in 26

Table 1. Definitions of stepped-wedge cluster randomized trial key methodological elements

Variables	Definition
Types of SW-CRT design	Depending on whether measurements taken in the different periods within a cluster come from the same or different participants, the design is classified as closed cohort, open cohort, or cross-sectional type: Closed cohort design: All participants are recruited at the beginning of the trial, they are repeatedly assessed over multiple measurement points and cannot join the trial as it is ongoing. Open cohort design: Almost all participants are recruited at the beginning of the trial, they are repeatedly assessed over multiple measurement points and can join or leave the trial as it is ongoing. Cross-sectional design: Measurements come from different participants at each measurement point.
Complete design	Measurements are taken in each cluster-period.
Incomplete design	Measurements are deliberately omitted in some cluster-periods.
Transition period	Time to embed the intervention, with no performed measurement.
Primary outcome	Main or primary outcome specified by the authors; if not specified, we used the outcome reported for sample size calculation. In case no primary outcome or no sample size calculations were reported or no unique primary outcome could be identified, we considered the first outcome mentioned in the Methods section of the manuscript.
Positive result on the primary outcome	Statistically significant difference in favor of the intervention condition.
Negative result on the primary outcome	Nonsignificant difference or statistically significant difference in favor of the control condition.

(55%) and desire that all clusters receive the intervention in 25 (53%). A prior pilot study was mentioned in 18 (33%). A protocol was available for 44 studies (80%) (including seven obtained after contacting the authors). More than half of the studies randomized hospitals or hospital wards and the vast majority randomized clusters at a single timepoint. The most frequent design for inclusion of participants was cross-sectional (n = 46, 84%) and 24 (44%) studies had a prospective recruitment of participants. Routinely collected data were exclusively used in nearly half (26, 47%). Interventions were delivered only at the cluster level in 38 studies (69%). Results from the main analysis of the primary outcome were negative for 35 (64%) of the studies: either nonsignificant (n = 31) or significant in favor of the control group (n = 4).

3.3. Recruitment and implementation challenges

Thirty five trials (64%) had at least one of the recruitment or implementation challenge. Seven trials (13%) had both recruitment and implementation challenges. Among the 35 trials with a recruitment or implementation challenge, 18 (51%) had a modification in their design.

3.3.1. Frequency and description of recruitment challenges

A recruitment challenge was identified in 18 (33%) trials; under-recruitment of clusters occurred in one trial, under-recruitment of participants in 13 trials, and four trials reported a design adaptation to address recruitment issues (Table 3). In 14 (26%) trials, we had insufficient data to judge whether there were recruitment challenges. Among the 31 trials for which we had the number of participants in the control and intervention conditions reported separately, under-recruitment of participants occurred more frequently in the intervention condition (n = 15) (48%) than in the control condition (n = 12) (39%). Additional details are provided in Appendix B and C. Reasons for recruitment challenges were reported in seven trials; in three trials it was because of fewer than expected eligible participants (Appendix D).

3.3.2. Modifications of the design and adaptation of the analysis strategy in trials with recruitment challenges

Besides the four trials already classified as having a recruitment challenge based on reported modifications of the trial design, six other trials had modifications of their design. The most common modification was a change of the trial duration (n = 8), either extension of the trial duration (n = 6) or shortening of the trial duration (n = 2). We also identified a modification of the number of clusters (n = 3) and periods (n = 4) (Appendix D). In one trial, investigators implemented a rule after trial initiation to allow each cluster to move to their next period once 70% of their target sample size for the ongoing period was attained, leading to variable cluster-period lengths and unpredictable trial duration [12].

An adaptation of the analysis strategy was reported in four trials; three performed sensitivity analyses and one trial adapted the main analysis by excluding the period affected by inclusion issues. In one trial, sensitivity analyses to assess the impact of recruitment challenges qualitatively modified the results (from negative to positive) [13].

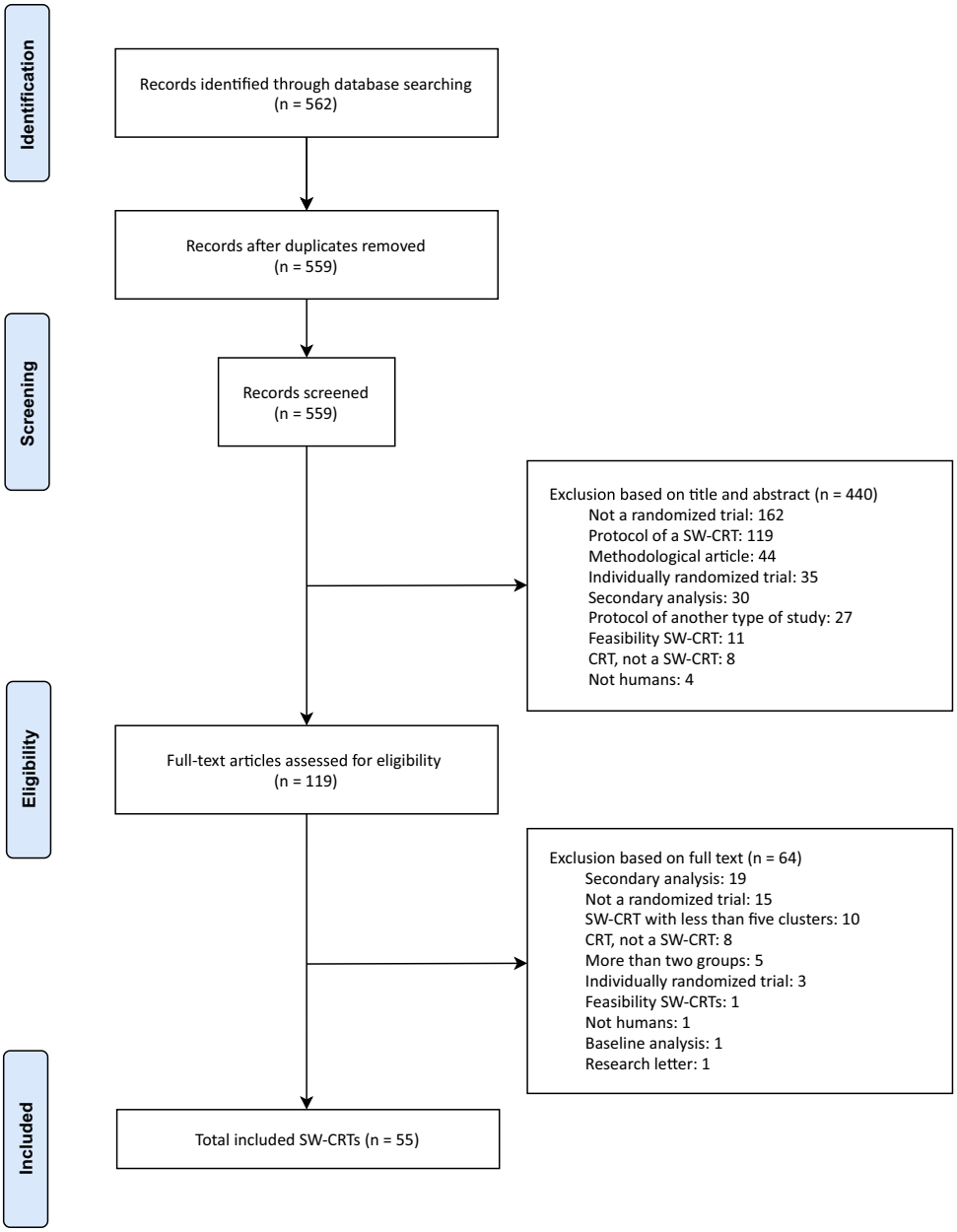


Fig. 1. Flow diagram of the selection process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3.3. Factors associated with recruitment challenges

Trial characteristics as per the presence or absence of recruitment challenges are described in Table 4. All trials with a recruitment challenge had a cross-sectional design and more than half recruited participants prospectively. The median planned sample size was greater in trials with recruitment challenges. A pilot study was performed in 44% of the trials with recruitment challenges as compared to 17% of the trials without recruitment challenges. Trials with a recruitment challenge more often allowed for a transition period (33%) than trials without recruitment challenges (5%).

3.3.4. Frequency and description of implementation challenges

Implementation challenges were identified in 24 (44%) trials; 18 had delayed implementation, eight had early implementation, and two had no implementation of the intervention in some clusters (Table 3). In 23 trials (42%), we had insufficient information to judge whether there were implementation challenges and in eight (15%) it was clearly reported that there were no implementation challenges. Reasons for implementation challenges were reported in 18 trials with three reporting multiple reasons

Table 2. Characteristics of stepped-wedge cluster randomized trials included in the review

General characteristics	*N* = 55
Publication year	
2019	30 (54.5)
2020	25 (45.5)
Location of study recruitment[a]	
North America	15 (27.3)
Europe	12 (21.8)
Asia	10 (18.2)
Oceania	9 (16.4)
Africa	7 (12.7)
Central America/South America/Caribbean	4 (7.3)
Type of clusters	
Hospitals or hospital wards	31 (56.4)
Primary care practices	12 (21.8)
Geographical areas (e.g., villages)	5 (9.1)
Nursing homes	4 (7.3)
Other[b]	3 (5.4)
Rationale for the stepped-wedge design reported	
Yes[c]	47 (85.5)
Logistical/practical reasons	26 (55.3)
Desire that all clusters receive the intervention	25 (53.2)
Statistical reasons (e.g., power considerations)	24 (51.1)
Ethical/equity reasons	13 (27.7)
Opportunistic intervention to be implemented anyway	7 (14.9)
Facilitate recruitment of clusters/participants	4 (8.5)
Other	3 (6.4)
No	8 (14.5)
Available protocol	
Yes	44 (80.0)
No	11 (20.0)
Prior pilot study	
Yes	18 (32.7)
No or unclear	37 (67.3)
Realized design features	
Number of sequences	5 [4; 7]
Number of clusters	11 [7; 18]
Number of periods[d]	6 [5; 10]
Number of participants[e]	7,635 [1,023; 32,194]
Allowed for a transition period	10 (18.2)
Complete design[f]	48 (90.6)
Randomization, design, and outcome data collection	
Timing of randomization	
At a single timepoint	47 (85.5)
In batches	3 (5.5)
Unclear	5 (9.1)
Trial design	
Cross-sectional	46 (83.6)
Continuous recruitment	42 (91.3)
Fixed time point recruitment	4 (8.7)
Open cohort	5 (9.1)
Closed cohort	4 (7.3)

(*Continued*)

Table 2. Continued

General characteristics	*N* = 55
Prospective recruitment of participants	
Yes	24 (43.6)
No	31 (56.4)
Exclusively routinely collected data	
Yes	26 (47.3)
No	29 (52.7)
Intervention and control conditions	
Type of experimental intervention[g]	
Targeted at the organization of healthcare or health delivery service	26 (47.3)
Targeted at healthcare professionals	19 (34.5)
Direct participant therapeutic intervention	7 (12.7)
Participant health promotion or educational intervention	3 (5.5)
Level at which intervention is delivered	
Cluster level	38 (69.1)
Individual level	7 (12.7)
Both cluster and individual levels	10 (18.2)
Type of control group	
Usual care	52 (94.5)
Other[h]	3 (5.5)
Results	
Results for the primary outcome	
Positive	20 (36.4)
Negative	35 (63.6)

Data are expressed as number and percentage, *n* (%) or median [interquartile range]. Percentages may not total 100% due to rounding.
[a] One trial was performed over three locations: Africa, South America, and Asia, so does not sum to 55.
[b] Other cluster types are mental healthcare service providers ($n = 1$), surgeons ($n = 1$), and reception centers for asylum seekers ($n = 1$).
[c] A trial can report multiple reasons, so does not sum to 47.
[d] Only the most prominent was considered.
[e] Two cases were missing.
[f] Four cases were missing.
[g] A complete design implies measurement in every cluster-period of the study. Two cases were missing.
[h] In two trials, control group consisted of minimal application of the experimental intervention and in one trial, it was an attention control intervention.

(Appendix D). The main reasons for delayed implementation were logistical or technical issues (six trials), issues with staff–staff turnover, strike or implementation planned during holidays (five trials), lower than expected recruitments (four trials), and approval issues (two trials). In two trials, intervention was implemented early because it became a standard of care during the trial. In one trial, some clusters refused to implement the intervention and in another, the intervention was not rolled out in three clusters because of technical issues.

3.3.5. Modifications of the design and adaptation of the analysis strategy in trials with implementation challenges

Among the 24 trials with implementation challenges, 14 reported a modification of their design such as extension of the trial duration ($n = 7$), shortening of the trial duration ($n = 4$), addition of periods ($n = 4$), and addition of clusters ($n = 2$) (Appendix D).

Five trials with implementation challenges adapted their analysis strategy; in two the main analysis was performed using an "as implemented" strategy; in three trials, sensitivity analyses were performed to assess the impact of implementation challenges on the intervention effect and found consistent results with the main analysis (one positive and two negative results).

In three trials, the authors had anticipated the risk of implementation challenges at the planning stage of the trial; in two trials, they planned to perform sensitivity analyses in case of deviation from the planned implementation schedule; in one they planned to conduct an analysis based on the randomization schedule regardless of the actual date of implementation. Among these three trials, only one actually reported implementation challenges [14].

3.3.6. Factors associated with implementation challenges

Trials with implementation challenges more often took place in a healthcare setting—hospitals, wards, or

Table 3. Description of recruitment and implementation challenges in included stepped-wedge cluster randomized trials

Characteristics	*N* = 55
Recruitment or implementation challenge	35 (63.6)
Both recruitment and implementation challenges	7 (12.7)
Any recruitment challenge[a]	
Yes	18 (32.7)
None identified	23 (41.8)
Insufficient information to identify any recruitment challenges	14 (25.5)
Type of recruitment challenges encountered	
Underrecruitment of clusters	1
Underrecruitment[b] of participants	13
Design adaptation to address recruitment issues	4
Any implementation challenge[c]	
Yes	24 (43.6)
No, clearly reported that there were no implementation challenges	8 (14.6)
Insufficient information to identify any implementation challenges	23 (41.8)
Type of implementation challenges encountered[d]	
Early implementation	8
Delayed implementation	18
No implementation	2

[a] Recruitment challenge was defined as a planned number of clusters not achieved, a planned number of participants not achieved (less than 90% of the target number of participants recruited), or reported changes made to the design to achieve the planned number of participants.
[b] Recruitment or identification in case there was no prospective recruitment of participants.
[c] Implementation challenge includes early, late, or no implementation of the intervention in clusters that did not drop out of the study.
[d] A trial can report multiple implementation challenges, so does not sum to 24.

practices—(79%) than trials without implementation challenges (38%) (Table 4). They more often had a prospective recruitment of participants (54% vs. 25%) and less often allowed for a transition period (17% vs. 38%).

3.4. Other modifications of the design

Seven trials without recruitment or implementation challenges had at least one modification of the planned design. The most frequent modification was deviation from the planned duration of the study, which was identified in five of those studies and a modification in the number of periods in three. Additional details are provided in Appendix D.

4. Discussion

4.1. Summary of key findings and comparison with other studies

To our knowledge, this is the first methodological review to assess recruitment and implementation challenges in recent SW-CRTs. We found that nearly two-thirds of SW-CRTs had recruitment or implementation challenges. More than half of trials with challenges modified their planned design—most often the trial duration. In some trials, recruitment difficulties led to design modifications and deviation from the implementation schedule.

A previous review found that 33% of SW-CRTs did not reach their prespecified sample size, defined as less than 100% of the planned number of participants [9]. Our definition of under-recruitment was more permissive, allowing for a 10% margin. Using the same definition as the previous study, the prevalence of under-recruitment in our review was 43%, thus showing no improvement in recent SW-CRTs [9]. Moreover, our definition of a recruitment challenge was broader than only under-recruitment of participants and included under-recruitment of clusters and design modifications in response to recruitment difficulties. The same review found that 43% of SW-CRTs had one or more difficulties during the study roll out, data collection, or analysis but it did not specifically focus on implementation challenges [9]. Another article examining six case studies reported that a staggered implementation of SW-CRTs raises new practical challenges of adhering to the planned schedule [7].

Our results also highlight that planned and actual designs are poorly reported, preventing the identification of recruitment or implementation challenges and other changes made after trial initiation. Although the extension of the CONSORT statement for SW-CRTs clearly recommends the reporting of changes to methods after trial commencement, it seems that there is still room for improvement [6]. We identified seven studies with no identified recruitment or implementation challenges but with some modifications of their designs; such modifications could reflect challenges that were not reported. Owing to

Table 4. Characteristics of stepped-wedge cluster randomized trials included in the review as per the presence or absence of recruitment or implementation challenges

	Recruitment challenges present?[a]		Implementation challenges present?[b]	
Characteristics	**Yes (n = 18)**	**No (n = 23)**	**Yes (n = 24)**	**No (n = 8)**
Type of clusters				
Hospitals or hospital wards	12 (66.7)	13 (56.5)	14 (58.3)	3 (37.5)
Primary care practices	4 (22.2)	1 (4.4)	5 (20.8)	0
Nursing homes	0	3 (13.0)	0	3 (37.5)
Geographical areas	0	5 (21.7)	3 (12.5)	1 (12.5)
Other	2 (11.1)	1 (4.4)	2 (8.3)	1 (12.5)
Available protocol				
Yes	14 (77.8)	21 (91.3)	22 (91.7)	7 (87.5)
No	4 (22.2)	2 (8.7)	2 (8.3)	1 (12.5)
Prior pilot study				
Yes	8 (44.4)	4 (17.4)	9 (37.5)	3 (37.5)
No or unclear	10 (55.6)	19 (82.6)	15 (62.5)	5 (62.5)
Number of clusters	11 [7; 19]	12 [6; 16]	11.5 [6.5; 17.0]	8.5 [6.5; 11.0]
Number of participants, as planned	3,200 [915; 14,000][c]	1,800 [640; 32,400]	3,000 [960; 32,400][d]	1,780 [680; 6,763]
Trial design				
Cross-sectional	18 (100.0)	17 (73.9)	21 (87.5)	5 (62.5)
Open cohort	0	3 (13.0)	0	2 (25.0)
Closed cohort	0	3 (13.0)	3 (12.5)	1 (12.5)
Prospective recruitment of participants				
Yes	10 (55.6)	13 (56.5)	13 (54.2)	2 (25.0)
No	8 (44.4)	10 (43.5)	11 (45.8)	6 (75.0)
Allowed for a transition period, as realized				
Yes	6 (33.3)	1[c] (4.5)	4 (16.7)	3 (37.5)
No	12 (66.7)	21 (95.5)	20 (83.3)	5 (62.5)
Type of experimental intervention				
Targeted at the organization of healthcare or health delivery service	9 (50.0)	8 (34.8)	8 (33.3)	6 (75.0)
Targeted at healthcare professionals	5 (27.8)	9 (39.1)	12 (50.0)	2 (25.0)
Direct participant therapeutic intervention	2 (11.1)	5 (21.7)	2 (8.3)	0
Participant health promotion or educational intervention	2 (11.1)	1 (4.4)	2 (8.3)	0
Level at which intervention is targeted				
Cluster level	11 (61.1)	14 (60.9)	16 (66.7)	6 (75.0)
Individual level	3 (16.7)	4 (17.4)	3 (12.5)	0
Both cluster and individual levels	4 (22.2)	5 (21.7)	5 (20.8)	2 (25.0)
Exclusively routinely collected data				
Yes	7 (38.9)	7 (30.4)	12 (50.0)	4 (50.0)
No	11 (61.1)	16 (69.6)	12 (50.0)	4 (50.0)
Modification of the design				
Yes	10 (55.6)	8 (34.8)	14 (58.3)	3 (37.5)
No	8 (44.4)	15 (65.2)	10 (41.7)	5 (62.5)

Data are expressed as number and percentage, n (%) or median [interquartile range]. Percentages may not total 100% due to rounding.

[a] 14 trials in which there was insufficient information to identify any recruitment challenge were excluded. Recruitment covers recruitment and identification in case there was no prospective recruitment of participants.

[b] 23 trials in which there was insufficient information to identify any implementation challenges were excluded.

[c] Information was missing for one trial.

[d] Information was missing for five trials.

Table 5. Recommendations for planning and reporting future SW-CRTs

Stage	Recommendations
Planning	- Write a precise roadmap for implementation of the intervention in clusters, especially when the study team has to implement the intervention in several clusters simultaneously, and ensure that the planned implementation schedule is practicable. - Do not underestimate the likelihood of recruitment and implementation challenges during major holidays and other events. - Obtain all necessary REC and gatekeeper approvals for all clusters before the beginning of the trial. - Before randomization, obtain agreement—preferably in writing—from all participating clusters about the implementation schedule including the required lead time to prepare for implementation. - Plan and prespecify the analysis strategy to explain how any possible deviations from the implementation and recruitment schedule will be handled in the primary and secondary analyses.
Reporting	- Clearly report the planned and actual number of participants (per condition and per cluster-period, when appropriate). - Clearly report the planned and actual schedule for implementation of the intervention. - Rather than merely state that the analysis was conducted on an intention-to-treat basis, clearly report the analysis population, and explain how deviations from the planned schedule were handled. - In case of recruitment or implementation challenges or modification of the planned design, report the reasons to allow an assessment of implications for risks of bias.

Abbreviations: REC, research ethics committee.

poor reporting of reasons in most of the trials, we were unable to judge whether recruitment challenges were preventable. One or multiple reasons were identified in 75% of trials with implementation challenges, some preventable (such as holiday periods) but others completely unpredictable (such as strike or concurrent major reform). Of note, SW-CRTs included in this review were not impacted by the corona pandemic (the most recent ended in August 2019) but such a major event certainly challenged ongoing trials at that time.

Our small sample sizes did not allow us to test whether trial characteristics were associated with recruitment or implementation challenges. Nevertheless, although the use of a feasibility study before an SW-CRT has been proposed as a solution to avoid practical difficulties [15], we observed more recruitment challenges in trials with a prior pilot study than in those without. Our analysis cannot rule out the possibility that pilot studies were more frequently performed in highly complex trials prone to recruitment challenges. The use of a transition period was more frequently reported in trials with identified recruitment challenges but less often in trials with implementation challenges. Contrary to our hypothesis, we observed a similar prevalence of prospective recruitment and exclusive use of routinely collected data among trials with and without recruitment challenges.

When there are implementation challenges, interpretation of the results is complex. The analysis strategy should be prespecified, including how deviations from the planned implementation schedule will be handled. A delayed implementation of the intervention and extension of the trial duration can also happen in a two-parallel group CRT but such deviations are more problematic in an SW-CRT because it affects unevenly intervention and control observations [16]. A first simulation study has shown that early implementation of the intervention can lead to biased intervention effect estimates which may be addressed by using models incorporating fixed or random group-by-time effects [17]. Implementation and recruitment challenges are also likely to impact power but further work is needed to explore how such challenges can affect the results of SW-CRTs.

4.2. Recommendations for future studies

Although several trials did anticipate practical challenges in their protocol, trialists may not fully appreciate the associated logistical complexities, need for time-constrained recruitment, and likelihood of deviations from the implementation schedule. We suggest some recommendations for the planning and reporting of future SW-CRTs (Table 5). These recommendations for planning and reporting can be used by trialists and journal editors to appraise the SW-CRT protocol and results, respectively.

4.3. Limitations

Our study has several limitations. The search strategy might have missed some SW-CRTs. However, our aim was not to be exhaustive but to provide information on a sample of recent SW-CRTs. Owing to poor reporting of a planned and actual design in many trials, it was difficult to identify recruitment or implementation challenges and we had to add a category for trials with insufficient information. Our definition of implementation challenges focused on deviations from the planned timing and did not include elements related to fidelity of the delivered intervention. Indeed, core components of the intervention were hardly ever reported so it was impossible to judge whether clusters were fully exposed to the intervention or not. As we included trials published close to the publication of the CONSORT extension for SW-CRTs [6], most of the included trials were probably designed beforehand and one would expect some improvements in future trials. Finally, sample sizes were small making it difficult to explore factors associated with recruitment or implementation challenges.

5. Conclusion

Recruitment or implementation challenges are frequent in SW-CRTs. The theoretical advantages of an SW-CRT

might be compromised by their organizational and logistical time-constrained requirements. Our practical recommendations may help researchers to enhance the design and reporting of future SW-CRTs.

CRediT authorship contribution statement

Agnès Caille: Conceptualization, Formal analysis, Data curation, Writing − original draft. **Monica Taljaard:** Conceptualization, Data curation, Writing − original draft. **Floriane Le Vilain—Abraham:** Data curation, Writing − review & editing. **Alexis Le Moigne:** Data curation, Writing − review & editing. **Andrew J. Copas:** Conceptualization, Writing − review & editing. **Florence Tubach:** Conceptualization, Supervision, Writing − original draft. **Agnès Dechartres:** Conceptualization, Data curation, Supervision, Writing − original draft.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.04.024.

References

[1] Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
[2] Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 2015;350:h391.
[3] Campbell MJ, Hemming K, Taljaard M. The stepped wedge cluster randomised trial: what it is and when it should be used. Med J Aust 2019;210:253−254.e1.
[4] Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? Int J Epidemiol 2020;49:1043−52.
[5] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 2007;28:182−91.
[6] Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, Thompson JA, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ 2018;363:k1614.
[7] Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials 2015;16:351.
[8] Adrion C, Weiss B, Paul N, Berger E, Busse R, Marschall U, et al. Enhanced Recovery after Intensive Care (ERIC): study protocol for a German stepped wedge cluster randomised controlled trial to evaluate the effectiveness of a critical care telehealth program on process quality and functional outcomes. BMJ Open 2020;10:e036096.
[9] Eichner FA, Groenwold RHH, Grobbee DE, Oude Rengerink K. Systematic review showed that stepped-wedge cluster randomized trials often did not reach their planned sample size. J Clin Epidemiol 2019;107:89−100.
[10] Hemming K, Carroll K, Thompson J, Forbes A, Taljaard M, SW-CRT Review Group. Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review. J Clin Epidemiol 2019;107:77−88.
[11] Airtable. Available at https://airtable.com/. Accessed July 30, 2021.
[12] Schwarze ML, Buffington A, Tucholka JL, Hanlon B, Rathouz PJ, Marka N, et al. Effectiveness of a question prompt list intervention for older patients considering major surgery: a multisite randomized clinical trial. JAMA Surg 2020;155:6−13.
[13] Raphaelis S, Frommlet F, Mayer H, Koller A. Implementation of a nurse-led self-management support intervention for patients with cancer-related pain: a cluster randomized phase-IV study with a stepped wedge design (EvANtiPain). BMC Cancer 2020;20:559.
[14] Lenguerrand E, Winter C, Siassakos D, MacLennan G, Innes K, Lynch P, et al. Effect of hands-on interprofessional simulation training for local emergencies in Scotland: the THISTLE stepped-wedge design randomised controlled trial. BMJ Qual Saf 2020;29:122−34.
[15] Kristunas CA, Hemming K, Eborall H, Eldridge S, Gray LJ. The current use of feasibility studies in the assessment of feasibility for stepped-wedge cluster randomised trials: a systematic review. BMC Med Res Methodol 2019;19:12.
[16] Neal B, Wu Y, Feng X, Zhang R, Zhang Y, Shi J, et al. Effect of salt substitution on cardiovascular events and death. N Engl J Med 2021;385:1067−77.
[17] Rennert L, Heo M, Litwin AH, De Gruttola V. Accounting for confounding by time, early intervention adoption, and time-varying effect modification in the design and analysis of stepped-wedge designs: application to a proposed study design to reduce opioid-related mortality. BMC Med Res Methodol 2021;21:53.

Floriane LE VILAIN--ABRAHAM

Analyse d'un essai randomisé en clusters avec un critère de jugement de type survie

Résumé :

Un essai randomisé en cluster (ERC) est un essai dans lequel on randomise des groupes d'individus, nommés clusters, et non des individus. Ce schéma d'étude induit une corrélation entre les réponses des individus d'un même cluster qui doit être prise en compte lors de l'analyse statistique. Les critères de jugement de type survie ne sont pas rares dans les ERC, pourtant il n'existe pas de recommandations pour les analyser.
En survie, l'effet d'une intervention peut être mesuré par la différence de temps de survie moyen restreint (RMST) entre les bras intervention et contrôle jusqu'au temps t*. Cette mesure, alternative au rapport des risques instantanés, ne repose pas sur l'hypothèse des risques proportionnels et peut être facilement interprétée comme le gain d'espérance de vie obtenu grâce à l'intervention sur une durée t*.
L'objectif principal de ce travail était d'étudier l'utilisation de la différence de RMST pour l'analyse des critères de type survie dans les ERC.
Tout d'abord, nous avons proposé et comparé à l'aide d'une étude de simulation, deux approches pour estimer une différence de RMST pour les ERC (intégration des courbes de Kaplan-Meier et régression sur des pseudo-valeurs). Les méthodes avaient de bonnes performances pour un nombre de clusters suffisamment grand (≥50). Pour les ERC avec un petit nombre de clusters, le recours à un test de permutation avec la régression sur des pseudo-valeurs permettait de contrôler l'erreur de type I, qui dépassait 5% autrement. Nous avons ensuite proposé des méthodes alternatives au test de permutation, visant à réduire le biais de la variance estimée avec la régression sur des pseudo-valeurs. Ces méthodes, en particulier la correction de Fay et Graubard associée à une distribution de Student, présentaient une erreur de type I proche de 5%.
Ce travail ouvre la voie à une meilleure analyse des critères de jugement de type survie et à l'utilisation de la différence de RMST comme mesure de l'effet intervention dans le cadre des ERC.

Mots clés : essai randomisé en clusters, survie, temps de survie moyen restreint, pseudo-valeurs

Summary:

Cluster randomised trials (CRTs) are trials in which groups of individuals, called clusters, are randomised, rather than individuals themselves. In such study design, the responses of the individuals within the same cluster are correlated and statistical analysis must take into account this clustering. Time-to-event outcomes are not uncommon in CRTs, but there is no recommendation on the optimal analysis.
In randomised clinical trials with a time-to-event outcome, the intervention effect could be quantified by a difference in restricted mean survival time (RMST) between the intervention and control groups up to time t^*. This measure is an alternative to the hazard ratio that does not rely on the proportional hazards assumption. It is easily interpretable as the expected survival gain until t^* due to the intervention. The main objective of this work was to study the difference in RMST to analyse CRTs with time-to-event outcomes.
In the first part of this work, we proposed and compared through a simulation study, two approaches to estimate a difference in RSMT in CRTs (the integration of the Kaplan-Meier curves and the pseudo-values regression). The methods showed good performance when there was a sufficient number of clusters (≥50). For CRTs with a limited number of clusters, a permutation test for pseudo-values regression was implemented and controlled the type I error, which was over 5% otherwise. In the second part of the work, we proposed the use of small samples bias-corrections of the sandwich variance estimator, as an alternative to the permutation test. These methods, especially the Fay and Graubard correction associated with a Student distribution, showed type I errors close to 5%.
This work opens the way for better analysing time-to-event outcomes and for quantifying the intervention effect by a difference in RMST in CRTs.

Keywords: cluster randomised trial, time-to-event, restricted mean survival time, pseudo-values