Methodological review showed that time-to-event outcomes are often inadequately handled in cluster randomized trials

Agnès Caille, Elsa Tavernier, Monica Taljaard, Solène Desmée

Methodological review showed that time-to-event outcomes are often inadequately handled in cluster randomized trials

Agnès Caille[1,2*], Elsa Tavernier[1,2], Monica Taljaard[3,4], Solène Desmée[1]

1 Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France

2 INSERM CIC1415, CHRU de Tours, Tours, France

3 Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

4 School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

*Corresponding author: INSERM CIC1415, CHRU de Tours, 2 boulevard Tonnellé, 37044 Tours Cedex 9, France.

E-mail: agnes.caille@med.univ-tours.fr

Word count: 2987/max 3000

Abstract

Objectives: To estimate the prevalence of time-to-event (TTE) outcomes in cluster randomized trials (CRTs) and to examine their statistical management.

Study design and setting: We searched PubMed to identify primary reports of CRTs published in six major general medical journals (2013-2018). Nature of outcomes and, for TTE outcomes, statistical methods for sample size, analysis and measures of intracluster correlation were extracted.

Results: A TTE analysis was used in 17% of the CRTs (32/184) either as a primary or secondary outcome analysis, or in a sensitivity analysis. Among the five CRTs with a TTE primary outcome, two accounted for both intracluster correlation and the TTE nature of the outcome in sample size calculation; one reported a measure of intracluster correlation in the analysis. Among the 32 CRTs with a least one TTE analysis, 44% (14/32) accounted for clustering in all TTE analyses. We identified 12 additional CRTs in which there was at least one outcome not analyzed as TTE for which a TTE analysis might have been preferred.

Conclusion: TTE outcomes are not uncommon in CRTs but appropriate statistical methods are infrequently used. Our results suggest that further methodological development and explicit recommendations for TTE outcomes in CRTs are needed.

Keywords: Cluster randomized trial, time-to-event outcome, methodological review, intracluster correlation coefficient, survival analysis, statistical methods

Running title: Methodological review of time-to-event outcomes in cluster randomized trials

What is new?

Key findings

- Methodological research on cluster randomized trials (CRTs) has so far focused on continuous and binary outcomes. This review of 184 CRTs is the first to estimate the prevalence of time-to-event (TTE) outcomes in CRTs and to examine how TTE outcomes were managed.
- TTE outcomes were not uncommon in CRTs but were often incorrectly analysed, possibly due to lack of available methodology.
- Sample size calculations and statistical analyses in CRTs with TTE outcomes frequently did not adjust for clustering, possibly leading to incorrect sample sizes and inferences. Almost no CRTs with a TTE primary outcome reported a measure of intracluster correlation.
- We identified CRTs with a binary or a continuous variable that might have been more appropriately analyzed as a TTE.

What this adds to what was known?

- Statistical methods used for TTE outcomes in CRTs are often inadequate. This might be related to the scarcity of methodological literature and the lack of practical guidelines for TTE outcomes management in CRTs.

What is the implication and what should change now?

- Further methodological research and guidance are needed to help trialists in adequately planning and analyzing CRTs with TTE outcomes.

Introduction

Cluster randomized trials (CRTs) are trials in which intact social units, such as medical practices, hospitals, or communities, are randomized to intervention or control conditions while outcomes are assessed on individuals within such clusters[1]. This study design is a natural choice to evaluate the impact of interventions delivered at the cluster level or when there is substantial risk of contamination, and its use is rapidly increasing[2]. In CRTs, outcomes assessed on individuals from a given cluster tend to be more similar than outcomes of individuals from different clusters. The degree of clustering is commonly measured by the intracluster correlation coefficient (ICC)[3]. When reporting CRT results, an estimate of clustering should be provided for at least the primary outcome[4,5]. Reporting of an ICC is useful as it may inform future sample size calculations which require advance estimates of clustering to ensure the desired power can be achieved. Clustering must also be accounted for in the statistical analysis because standard statistical tests ignoring the intracluster correlation lead to underestimated standard errors, increased risk of type I error and thus potentially incorrect inferences about the intervention effect.

Most of the developments to quantify and account for clustering in CRTs have considered continuous or binary outcomes, and available methods for time-to-event (TTE) outcomes are limited.

The ICC was first defined for continuous outcomes[6,7] and was then extended to binary outcomes[8]. For a TTE outcome, the ICC has not been clearly defined and there is no recommended measure of clustering for correlated survival data. Kaplan-Meier curves and log rank test statistics are commonly used in survival analyses and the intervention effect is usually expressed with a hazard ratio obtained through a Cox semiparametric proportional hazards model. Two broad approaches are available to obtain a hazard ratio for clustered

survival data: shared frailty models and marginal models. Shared frailty models are cluster-specific models that incorporate a random effect (frailty term) shared by members of the same cluster[9]. Marginal Cox proportional hazards models estimate regression parameters using the usual method for independent observations and clustering is accounted for in estimation of the standard error by the use of a robust sandwich covariance matrix[10]. However, investigators currently have little guidance about the choice of optimal methods for sample size calculation, analysis and ICC estimation for clustered TTE outcomes.

To our knowledge, no methodological reviews have been conducted to determine how TTE outcomes are managed in CRTs. Such a review could inform further work on measurement of clustering and recommendations for analysis of TTE outcomes. The primary aims of this review were: (i) to estimate the prevalence of TTE outcomes in published CRTs; and for CRTs with a TTE outcome, (ii) to assess whether and how clustering was taken into account in sample size calculation and analysis; and (iii) to assess whether and how clustering was quantified.

Methods

Search strategy

In February 2019, one author (AC) searched MEDLINE via PubMed for all CRTs published from January 01, 2013 to December 31, 2018 in six of the highest impact factor general medical journals (Annals of Internal Medicine, Journal of the American Medical Association, New England Journal of Medicine, PLOS Medicine, The BMJ and The Lancet). The search strategy was adapted from a previously published search strategy[11] and is provided in Supplementary Appendix 1.

Study eligibility criteria

We included primary trial reports of CRTs (any type of design including crossover and stepped wedge) conducted in humans. Primary trial reports were defined as the report of the primary analysis of the primary outcome. We excluded non-randomized studies, protocols, feasibility studies, studies reporting only baseline data, studies reporting only secondary analyses, as well as methodological articles.

Study selection

The titles and abstracts of all articles identified by the search were imported into Zotero and screened independently and in duplicate by two reviewers (AC and ET). Disagreements were resolved through discussion. Full-text articles were obtained for all potentially eligible studies and screened independently and in duplicate by two of three reviewers (AC, SD and ET). For studies not meeting eligibility criteria, the reason for exclusion was recorded. Any discrepancies on eligibility were resolved by discussion with the third reviewer whenever necessary to reach a consensus.

Data extraction

Two of three authors (AC, SD and ET) independently extracted data from the identified studies using a data extraction form developed for this review (provided in Supplementary Appendix 2). Studies were randomly allocated to the reviewers with each reviewer extracting data from two-thirds of the studies. The initial data extraction form was pilot tested on a small number of studies and refined accordingly. Data were extracted from the full-text and from any available electronic supplementary files (e.g. protocol or statistical analysis plan), when appropriate. We collected data on the general characteristics of each CRT: publication year, journal, trial objective (superiority or noninferiority), trial design (parallel groups, crossover

or stepped-wedge), number of intervention groups, type of cluster, number of randomized clusters, and number of participants at baseline.

We also collected the nature of the primary outcome as primarily analyzed (binary, continuous, count, time-to-event, ordinal or unclear). When several primary outcomes were reported in the article or when the primary outcome was not identified, we used the outcome reported in the sample size calculation. If sample size calculations had been performed for several outcomes of several natures, the primary outcome nature was considered as unclear. For CRTs with TTE primary outcomes, we recorded whether the sample size calculation accounted for clustering and for the TTE nature of the outcome and whether any measure of clustering was reported in the analysis. For CRTs with a TTE outcome, we extracted the description of the statistical methods used for analysis and whether they accounted for clustering. Of note, it has been shown that the usual Kaplan-Meier estimator is consistent for correlated data[12] but adjustment is needed to estimate the variance of the survival function[13]. Thus, we considered that a CRT with Kaplan-Meier plots without 95% confidence interval (95% CI) and an intervention effect estimate with 95% CI and p-value obtained with an appropriate method, accounted for clustering. Any discrepancies in data extraction were resolved by discussion with the third reviewer whenever necessary to reach a consensus.

Finally, each CRT report without a primary or secondary TTE outcome was assessed to determine whether one or several outcome(s) might have been more appropriately analyzed as a TTE outcome. This evaluation was obtained by consensus between the three reviewers and based on the following decision rules[14,15]: the design of the trial must allow the measurement of the date of the event, and at least one of these two conditions must be met (i) some participants were lost to follow-up (≥10%) and (ii) the time to the event (and not only the occurrence of the event) was of interest. When the period of follow-up where the event

was possible was short, e.g death in the first 28 days of life, we considered that survival analysis was not relevant.

Analysis

We described the included trials using frequency and percentage for categorical variables and mean with standard deviation (SD) or median with interquartile range (IQR), as appropriate, for quantitative variables. Analyses were performed using R v3.3.2 (http://www.R-project.org, the R Foundation for Statistical Computing, Vienna, Austria).

Results

Among the 219 references identified through PubMed searching, 36 were excluded based on title and abstract, leaving 183 to assess for eligibility on full-text. After full-text reading, 4 further records were excluded (3 were secondary analysis of a CRT and 1 was an individually randomized trial) for a final sample of 179 articles corresponding to 184 CRTs (three articles reported on more than one CRT). Figure 1 shows the flow chart of the CRTs selection process for the review. The characteristics of the 184 included CRTs are reported in Table 1. Briefly, the CRTs randomized a median of 46 clusters and included a median of 3888 individuals. The majority had a superiority objective (n=175 [95.1%]) and 159 (86.4%) had a parallel group design.

Prevalence of TTE outcomes in CRTs

Among the 184 included CRTs, the primary outcome was a TTE in five trials (2.7%). The primary outcome was binary in 107 (58.2%), continuous in 46 (25.0%), count in 22 (12.0%), ordinal in one (0.5%) and unclear in three (1.6%). We identified three possibilities for a CRT to have a TTE analysis: CRTs with a TTE primary outcome (n=5), CRTs with at least one

TTE secondary outcome (n=22) and CRTs with a sensitivity analysis using survival methods of their primary outcome, primarily analyzed as binary (n=11). Three trials had both a primary and secondary TTE outcomes and three trials had both a sensitivity analysis of their binary primary outcome and a TTE secondary outcome. Thus, a total of 32 CRTs included at least one TTE analysis either as primary or secondary outcome analysis or in a sensitivity analysis of a binary outcome (Figure 2). The full list of the 32 studies is included with citations in Supplementary Appendix 3. The overall prevalence of TTE analysis in CRTs was 17.4% (95% CI; 12.2% to 23.7%). Among the 181 CRTs where the type of primary outcome was clear, the prevalence of a TTE primary outcome was 2.8% (95% CI; 0.9% to 6.3%).

CRTs presenting at least one TTE analysis were mostly similar to all the included CRTs for other characteristics (Table 1).

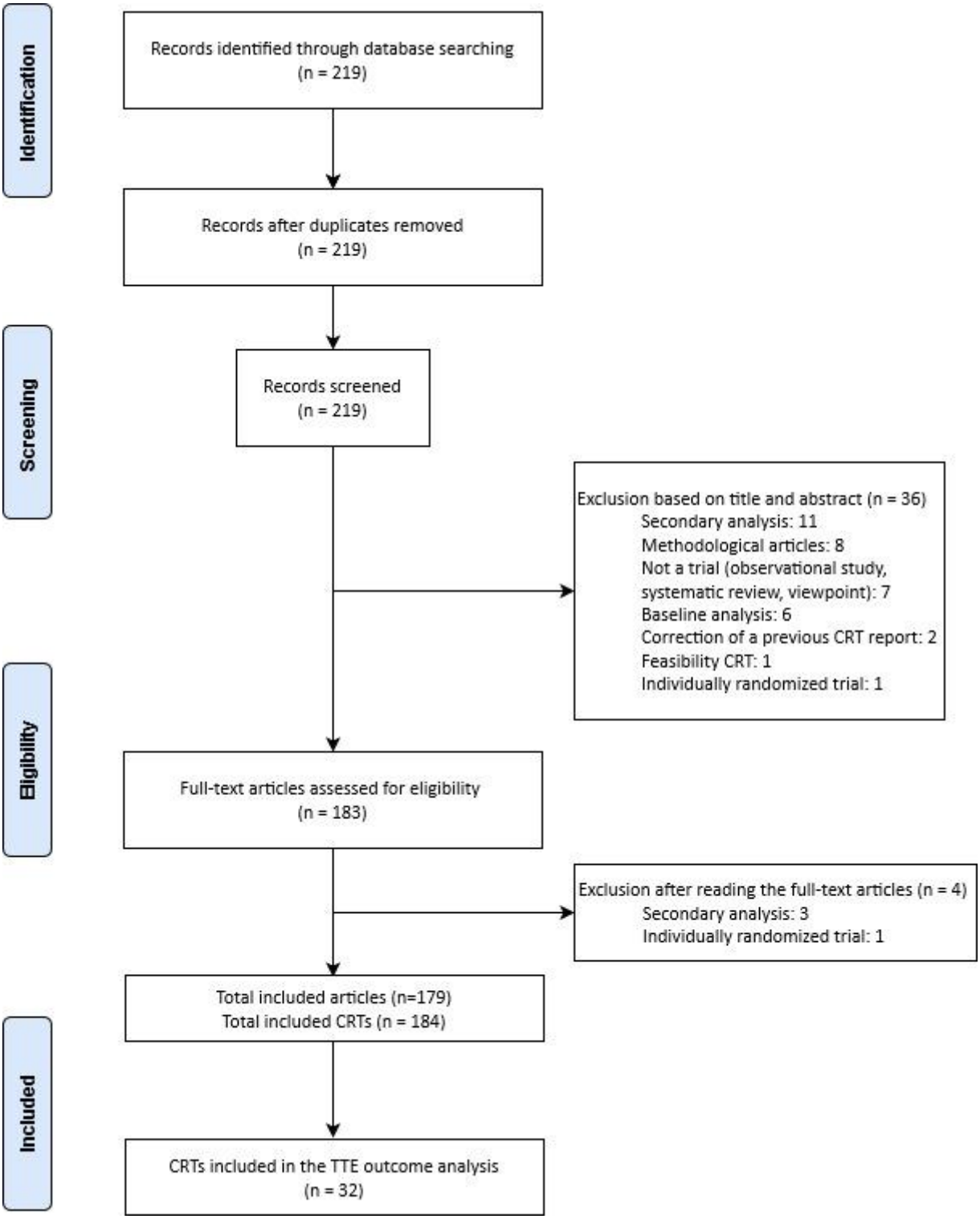

Figure 1. Flow chart summarizing studies identified, selected and included in the systematic review

Table 1. Characteristics of included CRTs (N=184) and CRTs with a TTE outcome analysis (N=32)

Characteristics	All CRTs	CRTs with a TTE

	N=184	N=32
Publication year:		
2013	37 (20.1)	4 (12.5)
2014	24 (13.0)	3 (9.4)
2015	28 (15.2)	9 (28.1)
2016	26 (14.1)	4 (12.5)
2017	33 (17.9)	5 (15.6)
2018	36 (19.6)	7 (21.9)
Journal:		
The Lancet	56 (30.4)	13 (40.6)
Plos Medicine	48 (26.1)	1 (3.1)
JAMA	30 (16.3)	9 (28.1)
BMJ	25 (13.6)	0
NEJM	17 (9.2)	7 (21.9)
Annals of Internal Medicine	8 (4.3)	2 (6.2)
Trial objective:		
Superiority	175 (95.1)	31 (96.9)
Noninferiority or equivalence	9 (4.9)	1 (3.1)
Trial design:		
Parallel groups (including factorial design)	159 (86.4)	24 (75.0)
Stepped-wedge	15 (8.2)	2 (6.2)
Crossover	10 (5.4)	6 (18.8)
Number of groups:		
Two	148 (80.4)	27 (84.4)
More than two	36 (19.6)	5 (15.6)
If more than two groups, number of groups	3.5 [3.0;4.0]	3.0 [3.0;4.0]
Cluster:		
Communities/Residential areas	48 (26.1)	6 (18.8)
Hospitals, hospital units, hospital wards	46 (25.0)	15 (46.9)
Primary care practices	42 (22.8)	6 (18.8)
Schools	14 (7.6)	1 (3.1)
Individual health professionals	5 (2.7)	1 (3.1)
Households, families	5 (2.7)	1 (3.1)
Nursing homes, aged care	4 (2.2)	0
Worksites	1 (0.5)	0
Other	19 (10.0)	2 (6.2)
Number of clusters randomized		
Median [interquartile range]	46 [20;86]	48 [26;115]
Range	4 – 37 724	4 – 1 552
Number of included individuals		
Median [interquartile range]	3 888 [1 404;16 221]	4 438 [2 228;10 836]
Range	78 – 1 291 824	523 – 415 357
Level of primary outcome measure:		
Individual	174 (94.6)	32 (100.0)
Cluster	10 (5.4)	0
Level of primary outcome analysis:		
Individual	147 (79.9)	27 (84.3)
Cluster	31 (16.8)	4 (12.5)[a]
Both	3 (1.6)	0
Unclear	3 (1.6)	1 (3.1)
Primary outcome type (as analyzed):		
Binary	107 (58.2)	19 (59.4)
Continuous	46 (25.0)	1 (3.1)
Count	22 (12.0)	6 (18.8)
TTE	5 (2.7)	5 (15.6)
Ordinal	1 (0.5)	1 (3.1)
Unclear	3 (1.6)	0
Primary outcome analyzed as binary but sensitivity analysis using survival methods	11 (6.0)	-
One secondary outcome of TTE type[b]	22 (12.0)	-

One outcome might have been more appropriately analyzed as a TTE	12 (6.5)	-

CRTs, cluster randomized trials; TTE, time-to-event. Data are expressed as number and percentage, n (%) or median [interquartile range]. Percentages may not total 100% due to rounding.

[a]None of these 4 CRTs had a primary TTE outcome.

[b]For 11 trials the primary outcome was binary, for 6 it was count, for 3 it was TTE, for 1 it was continuous and for 1 it was ordinal. In 3 CRTs, a sensitivity analysis of the binary primary outcome using a survival analysis was also performed.

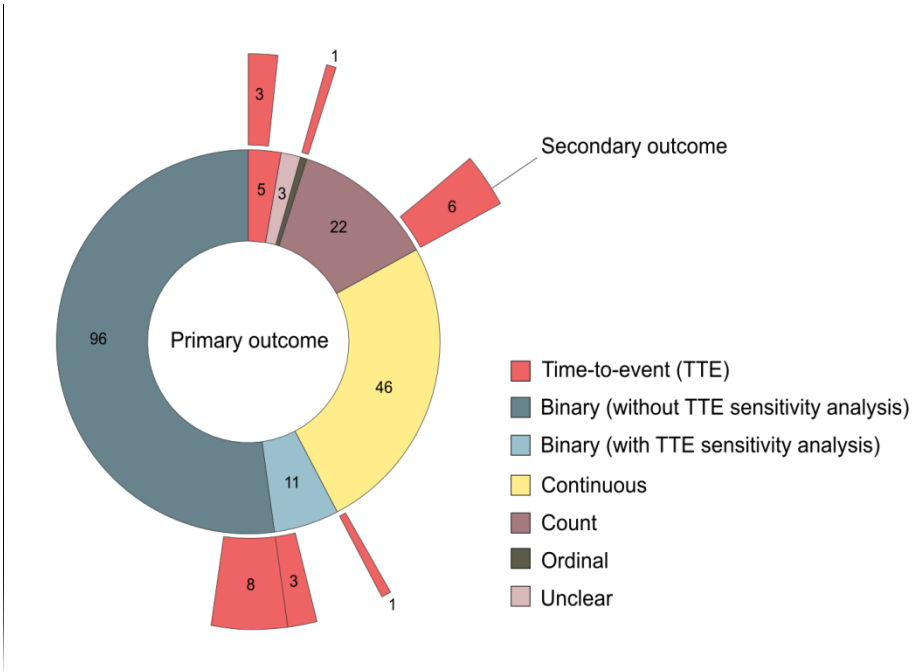

Figure 2. Donut plot of prevalence of time-to-event (TTE) outcomes among 184 cluster randomized trials, either as primary or secondary outcome or in a sensitivity analysis of a primary binary outcome

Sample size calculation for TTE outcomes in CRTs

Among the five CRTs in which the primary outcome was a TTE, clustering was clearly accounted for in sample size calculation in three and it was unclear in two trials (Table 2). The TTE nature of the outcome was clearly accounted for in two trials. It was unclear whether the TTE nature of the outcome was accounted for in two trials and in the remaining trial the outcome was considered as binary in the sample size calculation [A30]. The two trials [A13, A17] accounting for both intracluster correlation and the TTE nature of the primary outcome

used the method from Xie and Waksman[16] which extends the Freedman's formula for independent data to correlated data. In one of the two trials [A17], the authors also referenced the method from Gangnon and Kosorok[17] which extends the Schoenfeld's formula for individually randomized trials to CRTs. Description of both calculation methods is provided in Table 2.

Measure of intracluster correlation for TTE outcomes in CRTs

All trials with a TTE primary outcome were analyzed at the individual level and only one trial [A17] reported a measure of intracluster correlation (Table 2). In this trial, the primary outcome was duration of mechanical ventilation and the authors reported an ICC based on the deviance residuals of a Cox proportional hazards model with robust sandwich variance estimator.

Table 2: Sample size calculation methods and reporting of an intracluster correlation measure in CRTs with a TTE primary outcome

Author and Reference	Sample size method accounted for intracluster correlation?	Sample size method based on TTE?	Method used for sample size calculation and comments	Reporting of an intracluster correlation measure?*	Method used for measure of intracluster correlation
Loeb [A13]	Yes	Yes	Method from Xie and Waksman [16] Derived from Lee, Wei and Amato marginal model Intervention effect estimate is a hazard ratio Clustering accounted for with an ICC on censoring indicator Formula reduces to Freedman's for independent data	No	NA
Moll van Charante [A15]	Yes	Unclear	Methods differ between the protocol and the report.	No	NA
Curley [A17]	Yes	Yes	Methods from Xie and Waksman [16] and Gangnon and Kosorok [17] 1/Xie and Waksman Derived from Lee, Wei and Amato marginal model Intervention effect estimate is	Yes	ICC based on the deviance residuals

			a hazard ratio Clustering accounted for with an ICC on censoring indicator Formula reduces to Freedman's for independent data 2/Gangnon and Kosorok Based on a cluster-level weighted log-rank statistics Intervention effect estimate is a hazard ratio Clustering accounted for with a within-cluster martingale correlation Formula reduces to Schoenfeld's for independent data		
Qadri [A24]	Unclear	Unclear	Method used in unclear as the authors referenced a method described in Donner and Klar book [1] The method described for TTE outcome in Donner's book is the Hayes and Bennett method [18] which allows the sample size calculation for the comparison of incidence rates and includes a coefficient of variation to account for clustering The authors did not report a coefficient of variation in the sample size calculation description	No	NA
Huang [A30]	Unclear	No	Sample size calculation is not the same in the protocol and in the report, in the protocol the ICC is taken into account, in the report, it is unclear whether or not the ICC is accounted for. The outcome was considered as binary in both calculations.	No	NA

CRTs cluster randomized trials ; TTE time to event ; ICC intracluster correlation coefficient; NA not applicable

*All CRTs had primary analysis at individual level

Analysis of TTE outcomes in CRTs

Among the 32 CRTs including at least one TTE analysis, 14 clearly accounted for clustering in all reported TTE analyses (including the 5 CRTs with a TTE primary outcome), 10 did not

account for clustering in all, four accounted for clustering in some but not all analyses, and in three trials it was unclear whether clustering was accounted for (Table 3). One trial [A28] only reported Kaplan-Meier plots without either 95% CIs or p-values: the method was considered incomplete as no between-group comparison was performed.

The most frequently used strategy to adjust for clustering was the addition of a cluster random-effect (13 trials): 12 used a Cox proportional hazards model with shared frailty and one [A13] used a Cox proportional hazards model with nested frailty, accounting for two nested levels of clustering. A marginal strategy, using a Cox proportional hazards model with robust sandwich variance estimator, was used in six trials. One trial [A13] used both approaches: conditional and marginal. In three trials [A14, A19, A31], the authors state that TTE analyses were adjusted for clustering but the adjustment method was not reported. In one of these trials [A31], the model used to obtain hazard ratios was not reported.

Overall, Kaplan-Meier plots were provided in 22 trials and only one [A6] reported 95% CIs of the survival functions, but without adjusting for clustering. Among the 11 trials in which log-rank tests were used, adjustment for clustering was never reported.

Table 3. Analysis methods used for TTE outcomes in the included CRTs

Author and Reference	Type(s) of TTE outcome	Analysis method accounting for intracluster correlation	Methods used for TTE analysis[a]
Shah [A1]	Sensitivity[b]	Not all analyses The Cox proportional hazards model with shared frailty accounted for clustering. The log-rank test did not account for clustering.	Cox proportional hazards model with shared frailty Log-rank test Kaplan-Meier plots
Benger [A2]	SO	Yes	Cox proportional hazards model with shared frailty
Sperry [A3]	Sensitivity[b]	No	Cox proportional hazards model Log-rank test Kaplan-Meier plots
Wang [A4]	SO	No	Cox proportional hazards model Log-rank test Kaplan-Meier plots
White [A5]	SO	No	Gray's semiparametric survival regression model

Martin [A6]	Sensitivity[b]	No	Kaplan-Meier plots with 95% confidence intervals
Semler [A7]	Sensitivity[b]	Yes	Cox proportional hazards model with shared frailty
Guidet [A8]	Sensitivity[b]	Not all analyses The Cox proportional hazards model with robust sandwich variance estimator accounted for clustering. The log-rank test did not account for clustering.	Cox proportional hazards model with robust sandwich variance estimator Log-rank test Kaplan-Meier plots
Vinereanu [A9]	SO	Not all analyses The Cox proportional hazards model with shared frailty accounted for clustering. The log-rank test did not account for clustering.	Cox proportional hazards model with shared frailty Log-rank test Kaplan-Meier plots
Anderson [A10]	SO	Not all analyses The Cox proportional hazards model with robust sandwich variance estimator accounted for clustering. The log-rank test did not account for clustering.	Cox proportional hazards model with robust sandwich variance estimator Log-rank test Kaplan-Meier plots
Henao-Restrepo [A11]	Sensitivity[b]	Yes	Cox proportional hazards model with shared frailty Kaplan-Meier plots
Mortimer [A12]	SO	Yes	Cox proportional hazards model with shared frailty
Loeb [A13]	PO	Yes	Cox proportional hazards model with nested frailty Cox proportional hazards model with robust sandwich variance estimator
Brinkman [A14]	SO and sensitivity[b]	Unclear The authors state that all analyses accounted for the cluster randomized trial design but they don't explain how it is accounted for in the Cox proportional hazard and competing risk models.	Cox proportional hazards model Competing risks model Kaplan-Meier plots
Moll van Charante [A15]	PO and SO	Yes	Cox proportional hazards model with shared frailty Kaplan-Meier plots
Freund [A16]	SO	Yes	Cox proportional hazards model with shared frailty
Curley [A17]	PO and SO	Yes	Cox proportional hazards model with robust sandwich variance estimator Kaplan-Meier plots
Harper [A18]	SO	Yes	Cox proportional hazards model with shared frailty Kaplan-Meier plots Life-table analysis
Kalra [A19]	SO and sensitivity[b]	Unclear The authors state that the Fine and Gray's cumulative incidence curve model was adjusted for clustering but the method of adjustment is not reported. The log-rank test did not account for	Fine and Gray's cumulative incidence curve model adjusted for clustering Log-rank test Kaplan-Meier plots

		clustering.	
Khanna [A20]	SO	Yes	Cox proportional hazards model with robust sandwich variance estimator Kaplan-Meier plots
Kim [A21]	Sensitivity[b]	Yes	Cox proportional hazards model with shared frailty
Pinder [A22]	SO	No	Log-rank test Kaplan-Meier plots
Postma [A23]	SO and sensitivity[b]	Yes	Cox proportional hazards model with shared frailty Kaplan-Meier plots
Qadri [A24]	PO	Yes	Cox proportional hazards model with robust sandwich variance estimator Kaplan-Meier plots
Young [A25]	SO	No	Log-rank test Kaplan-Meier plots
Cox [A26]	SO	No	Log-rank test Kaplan-Meier plots
Oostdijk [A27]	SO	No	Cox proportional hazards model
West [A28]	Sensitivity[b]	Incomplete analysis No between group comparison	Kaplan-Meier plots
Climo [A29]	SO	No	Cox proportional hazards model Kaplan-Meier plots
Huang [A30]	PO and SO	Yes	Cox proportional hazards model with shared frailty
Little [A31]	SO	Unclear The authors state that "*The basic model* [was] *adjusted for baseline prescribing and clustering by physician and practice*." but the method of adjustment is not reported.	Not reported The authors state that "*The basic model* [was] *adjusted for baseline prescribing and clustering by physician and practice*." but they don't explain which models were used to obtain hazard ratios.
Zlotkin [A32]	SO	No	Log-rank test Kaplan-Meier plots

TTE time to event ; CRTs cluster randomized trials ; ICC intracluster correlation coefficient; PO primary outcome; SO at least one secondary outcome

[a]Refers to any methods used for a TTE outcome analysis in each CRT

[b]Sensitivity TTE analysis of a binary primary outcome

Missed opportunities to use a TTE analysis

We identified 12 CRTs (6.5%), currently without TTE outcomes, in which at least one outcome might have been more appropriately analyzed as a TTE. The identified outcome was the primary outcome in 3 CRTs and in 9, it was one or several secondary outcomes. A total of 18 such outcomes from these 12 CRTs were identified: nine were analyzed as a binary variable, four as a continuous variable (a duration), and five were analyzed both considering

the occurrence of the event (binary) and the time before the occurrence of the event (continuous). (Two examples are provided in Boxes 1 and 2, details on the 12 CRTs are in Table 4 and citation list in Supplementary Appendix 4).

Box 1: Illustrative case study in which analysis of an outcome as a time-to-event might have been more appropriate than the one used in the trial – Wouters et al. [B6]
In a CRT randomizing 59 nursing home wards to a multidisciplinary medication review or standard procedures, the primary outcome was discontinuation of use of at least one inappropriate medication after 4 months of follow-up. In the analysis, performed with mixed-effects models, this outcome was considered a binary variable and participants who were lost to follow-up (14.6% in the intervention group and 14.0% in the control group) were considered as discontinuation failure in the primary analysis. The intervention effect estimated with relative risks (RRs) was significant, with RR=1.37 (95% CI 1.02 to 1.75). In a secondary analysis of the primary outcome, the authors removed the patients who were lost to follow-up and the RR became statistically nonsignificant (RR=1.33 [95% CI 0.98 to 1.70]).
Comments: Analysis as a TTE outcome could have been of interest by using the information for censored (lost to follow-up) participants without making the strong assumption that they did not discontinue inappropriate medication during the 4 months follow-up.

Box 2: Illustrative case study in which analysis of an outcome as a time-to-event might have been more appropriate than the one used in the trial – Rat et al. [B7]
In a CRT randomizing 1482 general practices between three groups - two different physician notification strategies or usual care - the primary outcome was patient participation in colorectal cancer screening after 1-year follow-up. The authors used a logistic regression model for analysis considering participation as a binary outcome.
Comments: Time to participation is also of interest because it has a potential important impact on care and prognosis with possibly less invasive care and better prognosis for those with early detection of cancer. The actual analysis did not differentiate between early and late participation in screening and a TTE outcome could have been useful.

Table 4. Cluster randomized trials (CRTs) with outcomes that might have been more appropriately analyzed as TTE outcomes and justification

Author and Reference	Original outcome(s)	Primary (PO) or Secondary Outcome (SO)	Type of the original outcome(s)	Potential TTE outcome(s)	Justifications
Keenan [B1]	- Mortality	PO	Binary	Time to death	- The date of the event was collected - Follow-up from 7 to 25 months - Time to the event is of interest - Percentage lost to follow-up is unclear
Karlsson [B2]	- Stroke, transient ischemic attack (TIA) and systemic thromboembolism	SO	Binary	Time to stroke, transient ischemic attack (TIA) and systemic thromboembolism	- Collection of the date of the event was possible (use of electronic health record) - Follow-up of 12 months - Time to the event is of interest - Percentage lost to follow-up is unclear
Ballard [B3]	- Antipsychotic use - Mortality	SO	Binary Binary	Time to antipsychotic use Time to death	- Date of the event could have been collected - Follow-up of 9 months - Time to the event is of interest - Percentage lost to follow-up was 36.4% and 33.2% in the control group
Elul [B4]	- Linkage to care (+ time from diagnosis to linkage to care) - Antiretroviral therapy (ART) initiation (+time from ART eligibility to ART initiation) - Death	SO	Binary and continuous Binary and continuous Binary	Time from diagnosis to linkage to care Time from ART eligibility to ART initiation Time to death	- Date of the event was collected (electronic medical record) - Follow-up of 12 months - Time to the event is of interest - Percentage lost to follow-up is unclear
McNairy [B5]	- Linkage to care (+Time from diagnosis to linkage) - Antiretroviral therapy eligibility assessment (+Time from diagnosis to ART eligibility assessment) - ART initiation (+Time from Human Immunodeficiency Virus (HIV) testing to ART initiation) - Death	SO	Binary and Continuous Binary and Continuous Binary and Continuous Binary	Time from diagnosis to linkage Time from diagnosis to ART eligibility assessment Time from HIV testing to ART initiation Time to death	- Date of the event was collected - Follow-up of 12 months - Time to the event is of interest - Percentage lost to follow-up was 29% in the intervention group and 49% in the control group
Wouters [B6]	- Successfull discontinuation use of at least 1 inappropriate medication (without relapse symptoms or severe withdrawal effects)	PO	Binary	Time to discontinuation of at least one inappropriate medication	- Date of the event was collected - Follow-up of 4 months - Time to the event is of interest - Percentage lost to follow-up was 14.6% in the intervention group and 14.0% in the control group
Rat [B7]	- Patient participation in colorectal cancer screening	PO	Binary	Time to colorectal cancer screening	- Date of the event was collected - Follow-up of 12 months

					- Time to the event is of interest - Percentage lost to follow-up is unclear
Wardle [B8], Trial A	- Median number of days to return the guaiac faecal occult blood testing (gFBOT) kit	SO	Continuous	Time to return the gBOT kit	- Date of the event was collected - Follow-up of 18 weeks - Time to the event is of interest - Percentage lost to follow-up is unclear
Wardle [B8], Trial B	- Median number of days to return the guaiac faecal occult blood testing (gFBOT) kit	SO	Continuous	Time to return the gBOT kit	- Date of the event was collected - Follow-up of 18 weeks - Time to the event is of interest - Percentage lost to follow-up is unclear
Wardle [B8], Trial C	- Median number of days to return the guaiac faecal occult blood testing (gFBOT) kit	SO	Continuous	Time to return the gBOT kit	- Date of the event was collected - Follow-up of 18 weeks - Time to the event is of interest - Percentage lost to follow-up is unclear
Wardle [B8], Trial D	- Median number of days to return the guaiac faecal occult blood testing (gFBOT) kit	SO	Continuous	Time to return the gBOT kit	- Date of the event was collected - Follow-up of 18 weeks - Time to the event is of interest - Percentage lost to follow-up is unclear
Underwood [B9]	- All-cause mortality	SO	Binary	Time to death	- Date of the event could have been collected - Follow-up of 12 months with an open-cohort design (i.e. variable length of follow-up according to the time of inclusion in the trial) - Time to the event is of interest - Percentage lost to follow-up is unclear

PO: primary outcome; SO: secondary outcome

Discussion

To our knowledge, this is the first methodological review providing an overview of TTE outcomes in CRTs. We reviewed 184 CRTs published in high impact factor journals and described the methods used for TTE analysis. There are three key findings. First, TTE outcomes are not uncommon in CRTs, although rarely used as primary outcome. Second, when TTE outcomes are encountered in CRTs, appropriate statistical methods for sample size, analysis and estimation of clustering are infrequently used. Third, we identified a substantial number of CRTs in which at least one outcome not analyzed as a TTE might have been more appropriately analyzed as a TTE.

Our results are consistent with previous reviews which found that TTE outcomes are seldom used as primary outcomes in CRTs. One review[19] of 100 CRTs did not find any TTE primary outcomes. Another review[20] evaluating the handling of missing data in CRTs excluded 13 CRTs with a primary TTE outcome over 461 full-text articles assessed for eligibility (2.8%). We did not find any review describing the overall frequency of TTE outcomes in CRTs. A review[21] of 469 individually randomized trials reported a TTE primary outcome in 60 among the 219 (27%) published in higher impact journals (same journals than those of our review without PLOS Medicine), suggesting that TTE primary outcomes are more common in individually randomized trials than in CRTs. The reason may be the lack of accessible methodology for TTE outcomes in CRTs.

In our review, sample size calculation accounted for both the TTE nature and clustering in two among the five with a TTE primary outcome. Nevertheless, some methods exist for sample size calculation in CRTs with a TTE outcome, the review from Rutterford et al.[22] retrieved four suitable methods including those two[16,17] that were used in CRTs in our review.

Despite the recommendation of the CONSORT statement for CRTs to report a measure of clustering for each primary outcome, only one among the five CRTs with a TTE primary outcome provided a measure of clustering: an ICC based on the deviance residuals with no citation to any relevant methodological literature to support this method. This result is not surprising as there is no clear definition or guidelines on the ICC or alternative clustering measure for TTE outcomes. To our knowledge, two main strategies have been proposed to estimate the ICC for TTE outcomes: the first is to estimate the ICC from the binary censoring indicators[16] the other is to estimate the ICC from observed event times (excluding observations from censored participants)[23]. A recent simulation study[24] compared these two strategies and concluded that neither approach can be recommended because of bias and inconsistent results.

When clustering was accounted for in the analysis, we found that it was always through an individual-level regression model. As shown in binary outcomes, a conditional approach (with shared frailty) was more often implemented than a marginal approach (with robust variance estimator)[25]. Extensions of the log rank test statistic for clustered data have been recently proposed[26,27] but were not used in any of the 10 CRTs reporting log-rank tests.

In 12 further CRTs, we found that a survival analysis may have been a more appropriate choice than the method used in the report. One explanation for use of these non-optimal analysis strategies may be the lack of available statistical methods for TTE analysis in the context of CRTs. To our knowledge, the only comparison of methods for TTE outcomes in CRTs was published by Stedman et al.[28] and focused on specific scenarios suitable for CRTs randomizing physicians.

Our review has several limitations. First, we focused on CRTs published in six general medical journals with high impact factor. This choice may have selected CRTs in which reporting is better and more appropriate statistical methods are used than in other

journals[21,29]. Our results may thus provide an optimistic overview of the statistical methods used for TTE outcomes in CRTs. Second, our search did not include synonyms of "cluster randomized" such as "community" or "group randomized" and may have missed eligible CRTs published in these journals. Our aim was not to be exhaustive but to obtain a general overview of the handling of TTE in CRTs. Third, assessment of whether a TTE analysis might have been more appropriate for some outcomes was to some extent subjective but relied on specific rules and required the consensus of three reviewers. However, our review also has several strengths. Our findings are based on a large review of 184 published CRTs with duplicate data extraction using a predefined standardized data extraction form. Data extraction was performed by senior statisticians with strong expertise in either CRTs (AC and ET) or survival analysis (SD). As we reviewed CRTs published recently in high impact factor journals, we often had access to the protocol and statistical analysis plan of the trial providing us with important details on statistical methods.

Conclusion

Our review found that management of TTE outcomes could be improved and highlights the need for further methodological research and development of guidelines for optimal handling of TTE outcomes in CRTs.

CRediT authorship contribution statement

Agnès Caille: Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing – review & editing. Elsa Tavernier: Conceptualization, Data curation, Writing - review & editing. Monica Taljaard: Writing - review & editing. Solène Desmée: Conceptualization, Data curation Writing - review & editing.

Funding

This work is independent research supported by the French National Research Agency [ANR-19-CE36-0002 –QUARTET].

Acknowledgments

We thank Véronique Laurent-Buron for her help in retrieving full-text articles of selected references.

Conflict of interest

None declared.

Data Availability Statement

Extraction data are available from the corresponding author.

References

[1] Donner A. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

[2] https://researchmethodsresources.nih.gov/grt.aspx n.d.

[3] Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. International Statistical Review 2009;77:378–94. https://doi.org/10.1111/j.1751-5823.2009.00092.x.

[4] Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012;345:e5661.

[5] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, 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.

[6] Fisher R. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1921;1:3–32.

[7] Donner A. A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. International Statistical Review 1986;54:67–82.

[8] Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics 1999;55:137–48.

[9] Duchateau L, Janssen P. The Frailty Model. Springer Science & Business Media; 2007.

[10] Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. Stat Med 1994;13:2233–47. https://doi.org/10.1002/sim.4780132105.

[11] Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality. Trials 2016;17:575. https://doi.org/10.1186/s13063-016-1685-6.

[12] Ying Z, Wei LJ. The Kaplan-Meier Estimate for Dependent Failure Time Observations. Journal of Multivariate Analysis 1994;50:17–29. https://doi.org/10.1006/jmva.1994.1031.

[13] Williams RL. Product-limit survival functions with correlated survival times. Lifetime Data Anal 1995;1:171–86. https://doi.org/10.1007/BF00985768.

[14] Harrell F. Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

[15] Chevret S. Logistic or Cox model to identify risk factors of nosocomial infection: still a controversial issue. Intensive Care Med 2001;27:1559–60. https://doi.org/10.1007/s001340101066.

[16] Xie T, Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. Stat Med 2003;22:2835–46. https://doi.org/10.1002/sim.1536.

[17] Gangnon RE, Kosorok MR. Sample-size formula for clustered survival data using weighted log-rank statistics. Biometrika 2004;91:263–75. https://doi.org/10.1093/biomet/91.2.263.

[18] Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. Int J Epidemiol 1999;28:319–26.

[19] Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, 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. https://doi.org/10.1186/s13063-016-1571-2.

[20] Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials 2016;17:72. https://doi.org/10.1186/s13063-016-1201-z.

[21] Bala MM, Akl EA, Sun X, Bassler D, Mertz D, Mejza F, et al. Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis. J Clin Epidemiol 2013;66:286–95. https://doi.org/10.1016/j.jclinepi.2012.10.005.

[22] Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol 2015;44:1051–67. https://doi.org/10.1093/ije/dyv113.

[23] Segal MR, Neuhaus JM. Robust inference for multivariate survival data. Statistics in Medicine 1993;12:1019–31. https://doi.org/10.1002/sim.4780121103.

[24] Kalia S, Klar N, Donner A. On the estimation of intracluster correlation for time-to-event outcomes in cluster randomized trials. Stat Med 2016;35:5551–60. https://doi.org/10.1002/sim.7145.

[25] Turner EL, Platt A, Gallis JA, Tetreault K, Easter C, McKenzie JE, et al. Incomplete evidence of impact: Results of a systematic review of reporting of binary effect measures in cluster randomised trials. Submitted n.d.

[26] Jung S-H, Jeong J-H. Rank tests for clustered survival data. Lifetime Data Anal 2003;9:21–33.

[27] Gregg ME, Datta S, Lorenz D. A log rank test for clustered data with informative within-cluster group size. Statistics in Medicine 2018;37:4071–82. https://doi.org/10.1002/sim.7899.

[28] Stedman MR, Lew RA, Losina E, Gagnon DR, Solomon DH, Brookhart MA. A comparison of statistical approaches for physician-randomized trials with survival outcomes. Contemp Clin Trials 2012;33:104–15. https://doi.org/10.1016/j.cct.2011.08.008.

[29] Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ 2017;357:j2490. https://doi.org/10.1136/bmj.j2490.