METHODS AND ANALYSES OF STRATIFIED CLUSTER RANDOMIZED TRIALS

METHODOLOGICAL AND STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF STRATIFIED CLUSTER RANDOMIZED TRIALS

By
ASM (Sayem) BORHAN, B.Sc., M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2020) Hamilton, Ontario
(Health Research Methodology – Biostatistics Specialization)

TITLE: METHODOLOGICAL AND STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF STRATIFIED CLUSTER RANDOMIZED TRIALS

AUTHOR: ASM (Sayem) Borhan, BSc, MSc

SUPERVISOR: Dr. Lehana Thabane

NUMBER OF PAGES: xii, 109

ABSTRACT

Background and Objectives

While the number of adopting stratified cluster randomized trials (CRTs) is increasing, we have limited knowledge about the methodological and statistical issues pertaining to this design.

Our objectives were to (i) survey the literature to assess the methodological and statistical issues and quality of reporting of stratified CRTs; (ii) examine the sensitivity of methods for analyzing data from stratified CRTs; (iii) evaluate the performance of methods for analyzing continuous data from stratified CRTs.

Methods

We conducted a systematic survey and identified the stratified CRTs from the database MEDLINE. Data were abstracted on several methodological and statistical issues including sample size, randomization, and method of analysis. Two empirical studies were conducted to examine the robustness of methods for analyzing continuous and count data from stratified CRTs. Furthermore, a simulation study was performed to evaluate the performance of methods for analyzing continuous data from stratified CRTs under different scenarios including number of clusters, and cluster sizes.

Results and Conclusions

There was significant deficiency in reporting and analysis of data from stratified CRTs. The majority of the studies did not adjust the primary method for both clustering and stratification to assess the intervention effect.

The results from the empirical studies indicated that the methods for analyzing continuous and count data yielded similar conclusions. However, these methods varied in terms of magnitude of the effect sizes and widths of the 95% confidence intervals (CIs). Moreover, these studies demonstrated that, widths of the 95% CIs were narrower, and p-values were lower when adjusted for stratification compared to without adjusted for stratification.

The results from the simulation study showed that, performance of all methods improved as the number of clusters and cluster sizes increases. However, the performance of these methods deteriorated as the value of intra-cluster correlation coefficient (ICC) increases. Generalized estimating equations (GEE) and meta-regression yielded type I error rate of approximately 10% for small number of clusters. Meta-regression was the least powerful and efficient method compared to GEE, mixed-effects, and cluster-level linear regression methods.

The contributions of this thesis will guide the researchers to make informed decision about assessing the intervention effect and reporting of stratified CRTs.

ACKNOWLEDGEMENTS

This thesis would not be possible to complete without the support of so many amazing people, whom I am grateful to have in my side.

First, I would like to express my sincere gratitude to my incomparable supervisor and mentor Dr. Lehana Thabane. I am really thankful for his incredible support, guidance and advice over the course of my PhD study. I feel extremely lucky to have him as my supervisor and mentor who cared so much and who responded to my concerns, questions, and queries so promptly. He has gone above and beyond as a PhD supervisor and also helped me to improve my confidence, collaboration and soft skills. Thank you Dr. Thabane for everything you have done, and I hope you will continue to be my life-long mentor.

I am also immensely grateful to Dr. Jonathan Adachi and Dr. Alexandra Papaioannou for their invaluable support and guidance throughout my study period. Their insightful feedback really instrumental to complete these works. Most importantly, they have introduced me to a different world of research - in the area of aging, frailty, and bone health. Furthermore, their meticulous advice has helped me to complete and prepare some outstanding research works, which brought me several national and international awards. I cannot thank you enough for your tremendous encouragement and consistent support.

I would like to thank Dr. Jinhui Ma for reviewing the manuscripts, and providing thoughtful and constructive feedback, which played a vital role in improving the quality of these manuscripts. In addition, I am thankful to Dr. George Ioannidis, Dr. Courtney

Kennedy, Dr. Rizwana Mallick and Dr. Harsha Kathard for their support with data and feedback on the manuscripts.

I would like to thank my colleagues and staff in the Department of Health Research Methods, Evidence, and Impact and GERAS Centre for their support during my study period.

I am forever indebted to my parents, whose selfless sacrifice, dedication, encouragement, and support help me to come to this point. My father, whom I lost when I was 17, could not enjoy this achievement with us. I miss him everyday and his inspiration is key to move forward. Also, I would like to thank my sister and brother for their constant support and encouragement.

Last but not the least, I would like to thank my wife Syeda and my little Myra for their unconditional love, support and being on my side in this journey.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 2

Chapter 5

DECLARATION OF ACADEMIC ACHIEVEMENT

This thesis is a 'sandwich' thesis, which combined four individual projects prepared for publication in peer-reviewed journals. The following are contributions of Sayem Borhan in all of the papers in the dissertation: developing the research ideas and research questions; conducting all statistical analysis; writing all of the manuscripts; submitting the manuscripts; and responding to reviewers' comments. The work of this thesis was conducted between Fall 2016 and Summer 2020.

Chapter 1

Introduction

1 Introduction

Randomized controlled trials (RCTs) plays a vital role in evidence-based medicine (EBM) as these trials are the 'gold standard' for assessing the efficacy or effectiveness of treatments or interventions. RCTs can be based on individuals – where individual participants are randomized into intervention groups, or clusters – where intact clusters are randomized into intervention groups, which is known as cluster randomized trials (CRTs) [1]. Over the last couple of decades, the number adopting CRTs with stratified design to evaluate the intervention effect has been increasing [2, 3]. However, less attention has been given to methodological and statistical issues pertaining to stratified CRTs.

1.1 Cluster Randomized Trial

In CRTs, intact groups or clusters of individuals are randomly assigned to intervention groups [1]. These clusters can be diverse such as communities [4], schools [5], or geographical areas [6]. For example, in the CHAP trial, intact communities were randomized to assess the efficacy of community-based cardiovascular health awareness program [4].

1.1.1 Reasons for randomizing clusters

The most common reasons for randomizing clusters of individuals, instead of individuals, are:

(i) Type of intervention is suitable for cluster randomization:

In RCTs, we generally assess the efficacy of a treatment or medical interventions applicable to individual patients. However, there are certain type of interventions which are convenient and cost effective to deliver in the communities or other form of groups [7,8]. For example, general practices (GPs) were randomized in the diabetes education and self management for ongoing and newly diagnosed (DESMOND) trial to assess the effectiveness of an educational intervention about type II diabetes [9]. It is appropriate to deliver this intervention in a group, otherwise patients under the same doctor may wonder why some patients receiving different treatment and may demand the same.

(ii) To avoid treatment Contamination:

One of the main reasons for adopting CRTs is to avoid contamination – which occurs when participants in one intervention group receive part or full intervention allocated to another group [8]. In the vitamin D and osteoporosis (ViD*OS*) trial [10], the long-term care (LTC) homes were randomized into intervention knowledge translation (KT) group and control group, to assess the efficacy of KT intervention on improving the prescription of vitamin D, calcium and osteoporosis medications. If individual residents from the same LTC were randomized to different intervention groups, there would be a

chance of contamination as professionals or participants in the same LTC may alter their practice for all residents.

(iii) Convenience:

Sometimes CRTs have greater logistical convenience than the RCTs on individuals [8]. In the Ghana vitamin A supplementation trial (VAST) study [11], more than 20,000 children were enrolled to assess the efficacy of Vitamin A on mortality – a rare outcome. It would be difficult to organize this study by individually randomizing these 20,000 children, especially for the field workers. Because field workers, who delivered the study interventions, Vitamin A or placebo, would require carrying the list of children and check the groups these children were assigned to over the course of study period. Instead of randomizing individual children the investigators divided the study investigators into 185 geographical clusters with more than 100 children per cluster and randomized these clusters into vitamin A or placebo groups, which was much more convenient.

1.1.2 Methodological and statistical issues due to randomizing clusters

There are several methodological and statistical issues that arises due to randomization of intact clusters, which need to be taken into account in the design and analysis CRTs.

1.1.2.1 Within cluster correlation/Between cluster variation

Due to allocation of intact clusters of individuals the outcomes measured on the individuals in the same cluster are likely correlated. Within cluster correlation and between cluster variation represent two separate perspectives of the same phenomenon [8]. The degree of similarity among the outcomes from the same cluster is measured by intra-cluster correlation coefficient (ICC), denoted by the Greek letter ρ [1,7,8], is given by

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Where, σ_b^2 is the between-cluster variance; σ_w^2 is the within-cluster variance

ICC, ρ, generally, fall between 0 and 1[1,8]. A $\rho = 0$ indicates there is no clustering or no between-cluster variance. On the other hand, $\rho = 1$ indicates subjects in the same cluster are perfectly correlated. ICC is analogous to the standard Pearson correlation coefficient between any two observations from the same cluster [1].

Hayes and Moulton [8] defined a new approach to summarize the between cluster variability, known as coefficient of variation (CV), defined as the ratio of the standard deviation between clusters (σ_b) and overall mean of the outcome [7,8].

1.1.2.2 Design effect

In CRTs, data are collected from cluster sample of individuals, instead of simple random sample (SRS), and design effect (DE) is used to measure the inflation in variance due to this sampling [8]. Design effect is defined as the ratio of the variance of the outcome when clustering is taken into account to the variance of the outcome when clustering is not

taken into account and is given by $1 + (\bar{m} - 1)\rho$, where $\bar{m}$ is the average cluster size [1,7,8]. Inflation can be large even with small ICC if the average cluster size is large. For example, for a study with an estimated ICC, $\hat{\rho} = 0.02$, and average cluster size, $\bar{m} = 52$, the estimated DE is $1 + (\bar{m} - 1)\rho = 1 + (52 - 1) * 0.02 = 2.02$, i.e. we need twice as many participants as the RCT on individuals [7]. Design effect is also referred to as the variance inflation factor (VIF) since it measures the increase in the variance resulting from ignoring the clustering to allow for clustering [1,8]. If we conducted a CRT but performed the analysis like an RCT assuming individuals are independent, the standard error of the estimated parameters will be underestimated by $\sqrt{2.02}$ and more likely to lead to spurious statistically significant results [1,7].

1.1.2.3 Unit of analysis

There are two main approaches for analyzing data from CRTs:

(i) Cluster-level analysis

This is a two-stage approach. In the first stage, a summary measure of the outcome of interest for each cluster is obtained. In the second stage, an appropriate statistical method is used to analyze the summary measure from each cluster [8].

(ii) Individual-level analysis

Individual-level analysis is a one-stage process and based on individual-level data. It is possible to measure the effects of covariates through an individual-level analysis.

However, this approach may not perform well for CRTs with a small number of clusters [8].

1.2 Design of cluster randomized trials

In CRTs, clusters are usually allocated to intervention groups using three basic type of designs: (i) completely randomized – involve no stratification or matching on baseline prognostic factors; (ii) matched-pair – involve random assignment of two matched clusters into different intervention groups in each stratum; and (iii) stratified – extension of matched-pair design where more than two clusters in one stratum are randomized into intervention groups.

1.2.1 Stratified cluster randomized trial

In stratified CRTs, the available clusters are first grouped into two or more strata based on some prognostic, regional, socio-economic, epidemiologic, or other factors [1,8]. Then, the clusters within each stratum are randomized into intervention groups. The Mallick et al [2] study is an example of a stratified CRT, where schools were first divided into quintile (1-3: lower and 4-5: higher) and stratified as a high vs low school based on the socio-economic resources [12]. Then, within each stratum schools were randomized into Classroom Communication Resource (CCR) or Usual Care groups to examine the effect of CCR on peer attitude towards children who stutter [12].

A stratified design, aims to reduce the variance of the estimated intervention effect, is falls between completely randomized and matched-pair design. This design can help to

achieve more reduction in the variance of the estimated intervention effect than the pair-matched design [13]. A stratified design has several advantage over a matched-pair design including (i) it is more efficient and powerful due to the loss of fewer degrees of freedom; (ii) since there is more than two clusters in each stratum it is possible to test for variation in intervention effect between strata; (iii) if all strata have more than two clusters, loss of data from one cluster does not lead to any missing strata [8]. A stratified design is suitable for study with both small and large number of clusters since it has superior or similar precision and power compared to matched-pair and completely randomized design, respectively [8].

2 Methodological developments

There has been a great deal of methodological developments in the field of CRTs, over the last five decades, since the publication of the landmark paper by Cornfield in1978 [14] and subsequent publication of statistical notes in BMJ [15-18] and two key books [1, 19]. There are reviews that documented the methodological developments [20, 21]. Recently published books [7, 8, 22] also detailed the development in design and analysis of CRTs including sample size calculation, recruitment of clusters and individuals, ethics considerations, estimation of ICCs, statistical analysis methods, and reporting guideline.

Two individual-level analysis models for CRTs are commonly used in practice: (i) cluster-specific (CS) – estimate the average intervention effect if a participant stays in the same cluster but move from control to treatment arm; (ii) population-average (PA) – estimate the average intervention effect if a participant in the population move from control

to treatment arm [22]. Mixed-effects [23] – a CS approach, and generalized estimating equation (GEE) [24] – a PA approach, are two models commonly used to analyze the continuous, binary and count data from CRTs [1,7,8,22]. In addition, Bayesian methods have been developed for analyzing data from CRTs [25-28].

Several innovative design strategies including stepped wedge and pseudo cluster randomized design have been developed and the methods have been developed to analyze the data from these designs [22]. Moreover, the CONSORT statement has been extended to guide the researchers about reporting of CRTs [29].

2.1 Methodological developments in stratified CRTs

Like standard CRTs, it is necessary to adjust for clustering as well as stratification to assess the intervention effect of a stratified CRT [1,7,8,22]. Ignoring the adjustment for stratification leads to wider confidence intervals and larger p-values [30]. Thus, we may fail to identify the intervention effect when it does exist [30].

Both PA and CS methods discussed in the previous section, adjusting for stratification, can be used to assess the intervention effect from stratified CRTs [1,22]. Researchers have investigated the performance of methods for analyzing data from CRTs [31,32, 33]. On the other hand, sensitivity analysis helps us to assess the robustness of the results obtain from the primary method [34]. However, most of these focused on a completely randomized design and there is very limited knowledge about the sensitivity of methods for analyzing data from stratified CRTs in the literature.

3 Methodological and statistical challenges of stratified CRTs addressed in this thesis

Several methodological and statistical challenges are addressed in this thesis, which would be beneficial for design, analysis, and reposting of stratified CRTs. The objectives of this thesis are to (i) conduct a systematic survey of the literature to assess the statistical and methodological issues and quality of reporting; (ii) investigate the sensitivity of methods for analyzing data; (iii) assess the impact of not adjustment for stratification through empirical comparison; (iv) use simulation to evaluate the performance of methods.

3.1 Systematic survey of the literature to assess the current practice about reporting and analysis of data from stratified CRTs

Systematic survey of the literature and summarizing evidence about the current practice are essential to understanding the design characteristics, reporting and analysis of data from stratified CRTs. Moreover, this summarization helps us to understand whether there a lack of reporting and analysis methods to assess the effect of intervention, or whether there is any need for improvement. Kahan and Morris [30] summarizes the evidence from stratified RCTs on individuals and found that only 26% of the studies adjusted the primary analysis for balancing or stratification factors. Taljaard et al [35] recommended the search terms to identify the CRTs. To our knowledge, there is no such summarization of evidence regarding reporting and analysis of data from stratified CRTs.

In this thesis, we systematically surveyed the literature and identified the stratified CRTs by adding the term ‘strati*’ with the search terms suggested by Taljaard et al [35] from the database MEDLINE since the inception to July 2019. We summarized the

evidence on several aspect of design characteristics including sample size, randomization, method of analysis, and reporting of results.

3.2 Sensitivity of methods for analyzing continuous data from stratified CRTs

It is vital to assess the robustness of the results obtained from the RCTs or CRTs [34]. Sensitivity analysis plays an important role to examine the robustness of the conclusion that obtained from the primary method [34]. For CRTs, sensitivity analysis can be performed in different ways including (i) using methods different from the primary method; (ii) with or without adjusting for clustering; (iii) using different correlation structure [34]. Methods used to analyze the data from completely randomized CRTs can be extended to stratified CRTs [1, 22]. These methods fall into two broad categories: individual-level methods – based on individual-level data and cluster-level methods – based on cluster-level summary measurements. Mixed-effects model [23] and generalized estimating equation (GEE) [24] are individual-level methods. The meta-analytic approach [36] can be used to assess the effect of intervention across all strata of stratified CRTs, like multi-centre trials [37-39]. There is very limited investigation on the sensitivity of method for analyzing continuous data from stratified CRTs.

The outcome from stratified CRTs can be count data. For example, one of the outcomes in the Vitamin D and Osteoporosis Study (ViDOS) [10] was number of falls, which was over-dispersed (mean was smaller than the variance) with excessive zeros. Cluster-specific and population-average extension of Poisson regression can be used to analyze the count data from CRTs [1, 40]. Similarly, Pacheco et al [41] investigated the

performance of methods for analyzing over-dispersed count data from CRTs. However, researchers were mostly focused on completely randomized CRTs.

In this thesis, we investigated the sensitivity of methods for analyzing continuous and count data using the data from Mallick et al [12] and the ViDOS study [10], respectively.

3.3 Assess the impact of not adjustment for stratification

Failure to adjust for stratification leads to wider confidence intervals and a larger p-value of the estimated intervention effect [30]. Kahan and Morris [30] demonstrated the impact of not adjusing for stratification in their study based on RCT on individuals.

In this thesis, we empirically examined the impact of not adjusting for stratification using the data from two stratified CRTs – Mallick et al [12] and ViDOS [10] studies.

3.4 Performance of methods for analyzing data from stratified CRTs

Assessing the performance of methods is essential to help researchers choose the optimal methods to estimate the intervention effect. Researchers have investigated the performance of methods from CRTs, which were mostly limited to completely randomized CRTs [32, 42-45]. Performance of several mixed-effects methods incorporating the individual- and cluster-level association, was examined to analyze the pretest-postest continuous outcome from CRTs [43]. On the other hand, Borhan et al [42] and Austin [32] focused on the performance of methods for analyzing binary data. Chu et al [46]

investigated the performance of several methods including meta-regression to examine the intervention effect from multicentre RCTs. We have very limited evidence regarding the performance of methods for analyzing data from stratified CRTs.

In this thesis, we conducted a simulation study to appraise the performance of several methods for analyzing continuous data from stratified CRTs. The performance of these methods was examined in diverse scenarios including varying the number of clusters, cluster size, effect size, and ICCs.

4 Scope and outline of this thesis

This is a 'sandwich' thesis with four papers. First, we conducted a systematic survey to summarize the evidence about reporting and analysis of data from stratified CRTs. Second, we conducted an empirical study to examine the sensitivity and assess the impact of not adjusting for stratification when the outcome of interest is continuous. Third, we empirically assessed the sensitivity and impact of not adjusted for stratification in the case of count outcome. Fourth, we conducted a simulation study to evaluate the performance of methods for analyzing continuous data from stratified CRTs. Two of these four research works have been published while the other two are under review.

In this thesis, we focused on the following research questions:

1. What is the quality of reporting stratified CRTs?

2. Is the intervention effect assessed through proper adjustments – namely, clustering and stratification?

3. How robust is the methods for analyzing continuous data from stratified CRTs?

4. How robust is the methods for analyzing over-dispersed count data with excessive zeros from stratified CRTs?

5. What is the impact of not adjusting for stratification?

6. How the varying number of clusters, cluster sizes, ICCs, and effect sizes impact the performance of methods for analyzing continuous data from stratified CRTs?

In Chapter 2, we summarized the evidence regarding reporting and analysis of data from stratified CRTs. We focused on several vital methods and design characteristics including sample size, randomization and reporting of results. Also, we summarized the evidence about the primary method for assessing the intervention effect from stratified CRTs.

In Chapter 3, we examined the sensitivity of methods for analyzing continuous data from stratified CRTs. We empirically compared several methods for examining the intervention effect.

Chapter 4 contains the results of the empirical comparison of sensitivity of methods for analyzing count data, especially when the outcome was over-dispersed with excessive zeros, i.e. zero-inflated over-dispersed count data from stratified CRTs.

In Chapter 5, we conducted a simulation study to explore the performance of several methods for analyzing continuous data from stratified CRTs. The performance of these methods was evaluated under different scenarios including varying number of clusters, cluster sizes, ICCs and effect sizes.

Chapter 6 contains the discussion and conclusion.

Certainly, the evidence from this thesis will guide the researchers and decision makers to make informed decision about the reporting of stratified CRTs and methods for assessing the intervention effect from stratified CRTs.

Reference

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold London 2000.
2. Borhan S, Papaioannou A, Ma J, Adachi J, and Thabane L. Analysis and reporting of data from stratified cluster randomized trials. *Trials*, Under review.
3. Bland J. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology* 2004; 4: 21–27.
4. Kaczorowski J et al. Cardiovascular Health Awareness Program (CHAP): a community cluster-randomised trial among elderly Canadians. *Prev Med.* 2008;46(6):537-44.
5. Mallick R, Kathard H, Borhan ASM, Pillay M, Thabane L. A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students. *Trials* 2018; 19: 664.
6. Kroeger A, Avila EV, Morison L. Insecticide impregnated curtains to control domestic transmission of cutaneous leishmaniasis in Venezuela: cluster randomized trial. *British Medical Journal* 2002; 325(7368):810–813.
7. Campbell M, Walters S. *How to design, analyse, and report cluster randomised trials in medicine and health related research*. Wiley 2014. UK
8. Hayes R, Moulton L. *Cluster randomised trials*. CRC/Chapman & Hall 2017. 2nd Edition. UK
9. Davies MJ, Heller S, Skinner TC, et al. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for

people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. *BMJ* 2008; 336(7642): 491-495.

10. Kennedy et al. Successful knowledge translation intervention in long-term care: final results from the vitamin D and osteoporosis study (ViDOS) pilot cluster randomized controlled trial. *Trials* 2015; 16:214.
11. Ghana VSAT study team. Vitamin A supplementation in Northern Ghana: effects on clinic attendances, hospital admissions, and child mortality. *Lancet* 1993; 342: 7-12.
12. Mallick R, Kathard H, Borhan ASM, Pillay M, Thabane L. A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students. *Trials* 2018; 19: 664.
13. Todd J, Carpenter L, Li X, Nakiyingi J, Gray R, Hayes R. The effects of alternative study designs on the power of community randomized trials: evidence from three studies of human immunodeficiency virus prevention in East Africa. Int J Epidemiol. 2003;32(5):755-762.
14. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978; 108(2): 100 -102.
15. Bland J and Kerry S. Statistics notes. Trials randomised in clusters. *BMJ* 1997; 315 (7108): 600.
16. Kerry S, Bland J. Analysis of a trial randomised in clusters. *BMJ* 1998; 316 (7124): 54.

17. Kerry S, Bland, J. Sample size in cluster randomisation. *BMJ* 1998; 316 (7130): 549.
18. Kerry S, Bland J. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998; 316 (7142): 1455.
19. Murray D. *Design and Analysis of Group Randomised Trials*. Oxford University Press 1998, New York.
20. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-432.
21. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine* 2007;26(1):2-19.
22. Eldridge S, Kerry S. *A practical guide to cluster randomised trials in health services research*. Wiley 2012. UK
23. Hedeker D, Gibbons R, Flay B. Random-effects regression models for clustered data with an example from smoking prevention research. *J Consult Clin Psychol* 1994; 62:757–65.
24. Zeger L, Liang K-Y, Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;44: 1049–60.
25. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous response. *Stat Med* 2001; 20: 435-52
26. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med* 2001; 20:453-72.

27. Thompson SG, Warn DE, Turner RM. Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. *Stat Med* 2004; 23:389-410

28. Ma J, Thabane L, Kaczorowski J. et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: The Community Hypertension Assessment Trial (CHAT). *BMC Med Res Methodol* 2009; 9: 37.

29. Campbell MK, Piaggio G, Elbourne DR, Altman DG; for the CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; 345: e5661.

30. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012; 31: 328-40.

31. Murray DM, Hannan PJ, Pals SP, McCowen RG, Baker WL, Blitstein JL. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Stat Med* 2006;25(3):375-388.

32. Austin P. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med* 2007; 26(19):3550-3565.

33. Omar RZ, Thompson SG. Analysis of a cluster randomised trial with binary outcome data using a multilevel model. *Stat Med* 2000; 19:2675.

34. Thabane et al. (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology* 2013; 13(1):92.
35. Taljaard et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Medical Research Methodology* 2010; 10:15.
36. Whitehead A. *Meta-analysis of controlled clinical trials*. Edition 1. Chichester: John Wiley and Sons 2002.
37. Gould A. Multi-centre trial analysis revisited. *Stat Med* 1998;17: 1779-97,
38. Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Stat Med* 2000; 19(8):1115-1139.
39. Fleiss J. Analysis of data from multiclinic trials. *Control Clin Trials* 1986; 7(4):267-275.
40. Young M, Preisser J, Qaqish B, and Wolfson M. Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Statistical Methods for Medical Research* 2007;16: 167-184.
41. Pacheco et al. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. *Statist. Med.* 2009; 28:2989–3011.
42. Borhan, ASM, Klar N, Darlington G. Methods for the Analysis of Pretest-Posttest Binary Outcomes from Cluster Randomization Trials (2012). Electronic Thesis and Dissertation Repository. 825.

43. Klar N, Darlington G. Methods for analyzing change in cluster randomized trials. *Statistics in Medicine* 2004; 23: 2341-57.

44. Ukoumunne O, Carlin J, Gulliford M. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Stat Med* 2007; 26(18):3415-3428.

45. Leyrat C, Morgan K, Leurent B, Kahan B. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology* 2018; 47(1): 321-31.

46. Chu R, Thabane L, Ma J, Holbrook A, Pullenayegum E, Devereaux P. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: A simulation study. *BMC Medical Research Methodology* 2011; 11:21.

Chapter 2

Analysis and reporting of data from stratified cluster randomized trials – a systematic survey

Sayem Borhan[1,2,3], Alexandra Papaioannou[1,4,5], Jinhui Ma[1], Jonathan Adachi[5], Lehana Thabane[1,2,6]

[1] Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

[2] Biostatistics Unit, Research Institute of St Joseph's Healthcare, Hamilton, ON, Canada

[3] Department of Family Medicine, McMaster University, Hamilton, ON, Canada

[4] GERAS Centre, Hamilton Health Sciences, Hamilton, ON, Canada

[5] Department of Medicine, McMaster University, Hamilton, ON, Canada

[6] Departments of Pediatrics and Anesthesia, McMaster University, Hamilton, ON, Canada

Correspondence:

Dr. Lehana Thabane

Biostatistics Unit, Research Institute of St Joseph's Healthcare

3rd Floor, Martha Wing, Room H-325

50 Charlton Avenue East, Hamilton, ON

L8N 4A6

Canada

Abstract

Background

In order to correctly assess the effect of intervention from stratified cluster randomized trials (CRTs) it is necessary to adjust for both clustering and stratification, as failure to adjust can lead to erroneously large p-values and wider confidence intervals. We have conducted a systematic survey the literature to examine the analysis and reporting of stratified CRTs.

Method

We used the search terms to identify stratified CRTs from MEDLINE since the inception to July 2019. In phase 1, we screened the title and abstract for English only study and selected studies, including the protocols, for the next phase. In phase 2, we screened the full text, identified the published main results of the protocol papers of phase 1, and selected studies for data abstraction. Data abstraction form was piloted and developed using REDCap. We abstracted data on multiple study characteristics including whether the primary method adjusted for clustering and stratification, and reporting of sample size, randomization, and effect estimate.

Results

We screened 2686 studies in the phase1 and selected 286 studies for phase 2 - among them 185 studies were selected for data abstraction. Most of the selected studies were two-arm 140/185(76%) and parallel–group 165/185(89%) trials. Twenty-seven (15%) of the 185 studies did not provide any sample size or power calculation, while 105(57%) studies did not mention any method used for randomization. Further, 43(23%) and 150(81%) of 185 studies did not specifically provide information about how the strata were defined and included in the flow chart for all the stratification variables, respectively. More than half 114/185(62%) of the studies did not adjusted for both clustering and stratification and 66/73(90%) of the studies adjusted for stratification as a covariate.

Conclusion

Stratification helps to achieve the balance among intervention groups. But to correctly assess the intervention effect from stratified CRTs, it is important to adjust the primary analysis for both stratification and clustering. Reporting of stratified CRTs require substantial improvement in several areas including definition of strata, inclusion of stratification variable(s) in the flow chart or baseline characteristics table, and the stratum-specific number of clusters and individuals in the intervention groups.

Key words: Stratification, Cluster randomized trial, Systematic survey

Background

The random allocation of intact group of subjects - termed as clusters, into intervention groups are commonly known as cluster randomized trials (CRTs) [1]. The number of adopting CRTs to assess the effect of intervention is increasing [2]. The type of clusters can be diverse such as: geographical areas [3]; health care districts [4]; and schools [5]. There are several type of experimental design strategies that are used to allocate clusters including: completely randomized, stratified and matched-pair. Clusters are randomly allocated to intervention groups within each stratum in stratified design, which is suitable for small number of clusters [6].

The potential degree of similarity among the outcomes from the same cluster, measured through intra-cluster correlation coefficient (ICC), should be taken into account to assess the intervention effect from cluster randomized trials [1]. The failure to account for this correlation may yield a false positive result [1,7]. Scientists have developed and recommended statistical methods that can be used to examine the intervention effect, while taking into account the ICC of clustering [1]. In addition, in the case of a stratified design the statistical methods need to adjust for stratification [1]. It has been shown in the literature that variables used in the randomization process should be adjusted for in the analysis [8-13]. The absence of such adjustment in the analysis can yield large p-values and wider confidence intervals, which could potentially lead to a misleading conclusion that the intervention has no effect [14]. Borhan et al [15] empirically compared the methods for analyzing continuous data from stratified CRTs and reported that confidence intervals were

wider when not adjusted for stratification compared to when adjusted for stratification for the corresponding method.

Thus, to correctly assess the effect of intervention it is important to adjust for stratification variables as it will yield correct p-values and confidence intervals. Kahan and Morris [14] conducted a small-scale review on randomized trials and reported that only 26% of the studies adjusted for the balancing factors in their primary analysis. However, we have limited or no knowledge on how often the assessment of intervention effect from the stratified CRTs adjusted for clustering and stratification occurred.

In this study, we conducted a systematic survey to examine the analysis and reporting of stratified CRTs, which covered several aspects including how often the primary method to examine the effect of intervention adjusted for both clustering and stratification as well as whether the reporting of sample size calculations, randomization, and stratification were adequate.

Method

In this systematic survey we identified the stratified cluster randomized trials and abstracted data on multiple study characteristics including sample size estimation, randomization, analysis and reporting.

Search strategy and study selection

We added the term 'strati*' with the search terms (Table 1) suggested by Taljaard et al [16] to identify the stratified cluster randomized trials from MEDLINE since 1946 to July 2019. First, we performed title and abstract screening and selected the English only studies in the protocol. In the next phase, we screened the full text selected in the first phase and identified the studies for data abstraction. We used the protocol paper to identify the published main study results included in the study. In the case of multiple articles form the same trial we included only the main study results. Selection of studies were performed using EndNote X8. PRISMA flow diagram [17] was used to document the study selection process.

Data abstraction

A data abstraction form was piloted and developed using REDCap. Data abstraction form include data on many study characteristics including country, clinical area, setting of the study, sample size calculation, randomization, analysis of primary outcome, and reporting.

Outcome and analysis

We abstracted data on several methodological and reporting areas related to stratified CRT and descriptive summary: n (%) or mean (SD) or median (Q1, Q3), were

used to analyze the outcomes. There were several outcomes on reporting of sample size or power calculation including whether sample size or power, used level of significance, desired power, and adjustment of sample size for lost to follow-up reported. Similarly, we abstracted data on several issues related to randomization including: randomization unit, number and type of stratification variables and strata and method used for randomization. Several outcomes from the methods of primary outcome analysis were analyzed including: type of primary outcome, unit of analysis, type of primary analysis, whether the primary method adjusted for stratification or clustering or both, how the primary method adjusted for stratification variables, whether missing data were imputed or sensitivity analysis was performed, and statistical significance of intervention effect (only for 2-arm trials). Moreover, we abstracted data on several outcomes related to reporting including whether: study flow chart or baseline characteristics table included stratification variables, number of clusters or individuals for each stratum provided and the estimated ICC reported. See results section for details about the outcomes.

Results

Using the search strategy recommended by Taljaard et al [16] we have identified 2686 papers from MEDLINE since the inception to July 2019 (Table 1). At phase 1 we conducted title and abstract screening and identified 286 papers, including the protocols, for the next phase (Figure 1). In phase 2, we screened the full texts and identified the main

results papers of the protocols from phase 1 for data abstraction. Finally, 185 studies were selected for analysis (Figure 1).

The results of some basic characteristics of the selected studies are provided in Table 2. About 80% of the studies were from 2010 to 2019, while only 7 (4%) studies were from before 2000. Almost half of the studies, 48% were one centred, and most of the studies (31%) were conducted in USA or UK (Table 2). Thirty-six (19%) and 27 (15%) studies were focused on interventions related to child development or primary care/general practices. Almost the same number of studies 36 and 38 were school- or general practice-based, respectively (Table 2). Most of the studies 140 (76%) and 165 (89%) studies were 2-arm and parallel-group trials, respectively (Table 2).

One hundred and fifty-eight (85%) out of 185 studies provided sample size or power calculations while 66% of the studies adjusted for clustering (Figure 2). While more than 80% of the studies reported the level of significance or desired power in sample size or power calculations, only 10% of the studies reported the method used and 28% of the studies adjusted for lost to follow-up in sample size/power calculations (Figure 2). Like the setting of the study almost similar numbers of studies used school or primary care/general practice as the randomization unit (Figure 3). Almost half of the studies had one stratification variable, while only 2% of the studies had 4 or more stratification variables. Most of the stratification variables (35%) were based on geographical location or distance. More than half of the studies (57%) did not provide the method used for randomization, while 23% of the studies specified all the strata (Figure 3).

The results outcomes related to method of analysis of primary analysis are provided in Table 3. The primary outcome of 83% of the studies were continuous or binary. One hundred and forty-three (77%) studies performed individual-level analysis, while for 8% of the studies, it was not clear whether they performed cluster-level or individual-level analyses. More than half (52%) of the studies used an intention-to-treat approach as their primary analysis approach, while 43% of the studies did not mention their primary analysis approach (Table 3). Seventy-one (38%) studies reported primary method/effect estimate adjusted for both clustering and stratification, among the studies adjusted for stratification, 90% of the studies adjusted for stratification by using them as the covariate(s) (Table 3). Twenty-five (47%) of the studies reported their statistically significant intervention effect, among the studies where the effect estimate adjusted for both clustering and stratification. The results of outcomes pertaining to reporting of outcomes are provided in Figure 4. Only 19% and 31% of the studies included stratification variables in the flow chart or baseline characteristics table. Only 10% of the studies reported stratum-specific effect estimate.

Discussion

In this, first-ever, systematic survey we selected 185 stratified cluster randomized trials from MEDLINE since the inception to July 2019 and found that 38% of the studies reported effect estimate adjusted for both clustering and stratification. This results largely supported by the findings of Kahan and Morris [14], as they reported 26% of the studies from their review adjusted for balancing factors.

As we discussed before, in order to correctly assess the effect of intervention, it is important to analyze the primary/secondary outcomes adjusted for stratification variables as well as clustering [14], which is also established from the empirical study of Borhan et al [15]. From this systematic review it is evident that this type of adjustment is still scarce as more than half of the studies did not adjust for both stratification variables and clustering.

Along with performing adjusted analyses we also need to focus on other areas of stratified cluster randomization trials including: sample size calculation and randomization. Like randomized controlled trials on individuals it is necessary to report all the information used to calculate the sample size including detectable difference, level of significance, and desirable power. Further, it is also necessary to report the randomization method used to allocate clusters to intervention groups for each stratum, which was not reported by most of the studies in this survey.

It is noteworthy from this systematics survey that there are significant deficiencies in reporting the results from the stratified CRT. Reporting on the following areas, at minimum, would better represent and help the audience to better understand the stratified nature of this type of study: (1) only a few studies provided the reasoning for stratification. Reporting the reasoning for stratified design and choosing the stratification variable(s) would be helpful; (2) more than 20% of the studies did not provide the definition of all the strata. For a stratified design it is essential to report how all the strata are defined; (3) almost all the studies provided the study flow chart, while only 19% and 31% of the studies included stratification variables/strata in the flow chart or in baseline characteristic table, respectively. Inclusion of stratification variables in the flow chart or baseline characteristics

table would provide the clear depiction of the design; (4) only 20% and 11% of the studies reported the stratum-specific number of clusters and individuals in the intervention groups, respectively. Thus, more attention is needed to report these numbers; (5) reporting the stratum-specific, if possible, would help the readers to know the intervention effect in each stratum.

The major strength of this study was that we used the search terms recommended by Taljaard et al [16] to select the stratified cluster randomized trials from one of the largest database MEDLINE. Also, we included the published main trial results of the protocols selected in title and abstract screening. Further, this survey was based on the time period from 1946 to 2019. The major limitation of this study that, only one reviewer conducted this survey. Despite multiple checking or best effort, it is possible that, the reviewer may have failed to include some of the eligible studies.

A well-designed large-scale systematic review would depict a more complete picture about the analysis and reporting status of stratified cluster randomized trials. Furthermore, a guideline for analysis and reporting of stratified cluster randomized trials would be helpful to guide the researchers.

Conclusion

In this, first-ever, systematic survey we identified and selected stratified cluster randomized trials since inception 1946 to July 2019 for analysis. More than half (57%) and 15% of the studies did not report the method used for randomization and sample size or

power calculation, respectively. Similarly, primary method or reported intervention effect for more than half (62%) of the studies were not adjusted for both clustering and stratification. Also, there was substantial lack in reporting as only 23% of the studies did not provide details on how all the strata were defined, while 81% and 69% of the studies, respectively, did not include stratification variables in the study flow chart and did not provide stratum-specific summary statistics. To assess the intervention effect using stratified cluster randomized trial the analysis method should be adjusted for both stratification and clustering. Further, this type of study requires substantial improvement in reporting such as details about sample size/power calculation and randomization, definition of all strata, inclusion of stratification variable(s)/strata in study flow chart or baseline characteristics table, and stratum-specific number of clusters and individuals in the intervention groups. A reporting guideline focusing on stratified cluster randomized trial would help guide the researchers about analysis and reporting of data from this type of study.

Declarations

Ethics approval and consent to participants

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declared no competing interests for this study.

Funding

There was no funding for this study.

Authors' contribution

SB and LT were conceptualized this study. SB developed and designed this study. SB was involved with screening, study selection, data abstraction, preparing the results and wrote the first draft of this manuscript. SB prepared the first draft and all authors contributed equally to further improve this manuscript.

Reference

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold London 2000.
2. Bland J. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology* 2004; 4: 21–27.
3. Kroeger A, Avila EV, Morison L. Insecticide impregnated curtains to control domestic transmission of cutaneous leishmaniasis in Venezuela: cluster randomized trial. *British Medical Journal* 2002; 325(7368):810–813.
4. Jordhoy M, Fayers P, Saltnes T, Ahlner-Elmqvist M, Jannert M, Kaasa S. A palliative-care intervention and death at home: a cluster randomized trial. Lancet 2000; 356(9233):888–893.
5. Mallick R, Kathard H, Borhan ASM, Pillay M, Thabane L. A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students. *Trials* 2018; 19: 664.
6. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med* 1997; 16:1753-64.
7. Murray D, Varnell S, Blitstein J. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health* 2004; 94:423–32.
8. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Stat Med* 1999;18:1905-42.

9. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012;31:328-40.
10. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol* 1999;52:19-26.
11. Parzen M, Lipsitz SR, Dear KBG. Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial?. *Biom J* 1998;40:385-402.
12. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Control Clin Trials* 2000;21:330-42.
13. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. *Control Clin Trials* 2002;23:662-74.
14. Kahan B, Morris T. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* 2012;345:e5840.
15. Borhan S, Mallick R, Pillay M, Kathard H, Thabane L. Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study. *Contemporary Clinical Trial Communications* 2019; 15:100405.
16. Taljaard et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Medical Research Methodology* 2010, 10:15.

17. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group . Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement. PLoS Med. 2009;doi:10.1371/journal.pmed1000097.

Table 1: Search terms used to identify studies from MEDLINE since the inception to July 2019

1	randomized controlled trial.pt. (485792)
2	animals/ (6439569)
3	humans/ (17863499)
4	2 not (2 and 3) (4567683)
5	1 not 4 (474298)
6	(clusters$ adj2 randomi$).tw. (203)
7	((communit$ adj2 intervention$) or (communit$ adj2 randomi$)).tw. (7588)
8	group$ randomi$.tw. (3177)
9	6 or 7 or 8 (10908)
10	intervention?.tw. (861625)
11	cluster analysis/ (59383)
12	health promotion/ (70103)
13	program evaluation/ (59930)
14	health education/ (59114)
15	10 or 11 or 12 or 13 or 14 (1051673)
16	9 or 15 (1053569)
17	16 or 5 (1434924)
18	16 and 5 (92943)
19	strat*.mp. (1177180)
20	17 and 19 (160437)
21	18 and 19 (11717)
22	strati*.mp. (168454)
23	17 and 22 (24676)
24	18 and 22 (2686)

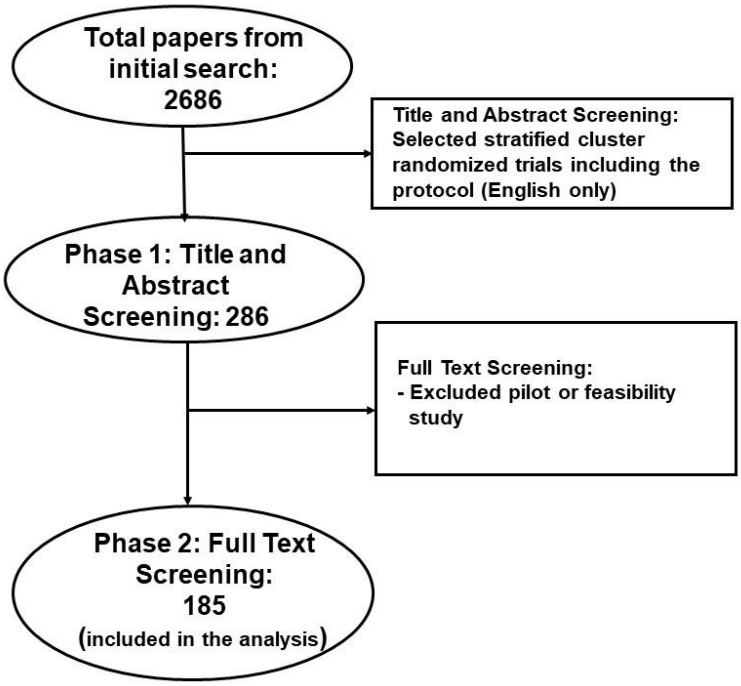

Figure 1: Flow chart of study selection

Table 2: Results of study characteristics

Characteristics	**Number of studies included; n=185**
Publication year; n (%)	
Before 2000	7 (4)
Between 2001 and 2010	30 (16)
Between 2011 and 2019	148 (80)
Centre of study; n (%)	
One	89 (48)
Two	44 (24)
Three or more	52 (28)
Country of the study; n (%)	
UK	32 (17)
USA	26 (14)
Canada	8 (4)
India	7 (4)
Australia	15 (8)
Denmark	7 (4)
Germany	5 (3)
Netherlands	6 (3)
South Africa	8 (4)
Others	71 (39)
Clinical area; n (%)	
Child development	36 (19)
Primary care	27 (15)
Maternal and child health	16 (9)
HIV	12 (6)
Cancer	9 (5)
Malaria	9 (5)
Cardiovascular	5 (3)
Cognitive and mental health	10 (5)
Others	61 (33)
Setting of the study; n (%)	
School	36 (19)
General practice/Primary care	38 (21)
Community	36 (19)
Hospital	18 (10)
Village	10 (5)
Family	6 (3)
Others	41 (22)
Design of the study; n (%)	
Parallel	165 (89)
Cross-over	2 (1)

Stepped wedge	3 (2)
Factorial	7 (4)
Matched pair	6 (3)
Split-plot	1 (1)
Zelen design	1 (1)
Arm of the study; n (%)	
2	140 (76)
3	28 (15)
4	14 (8)
5	2 (1)
6 or more	1 (1)

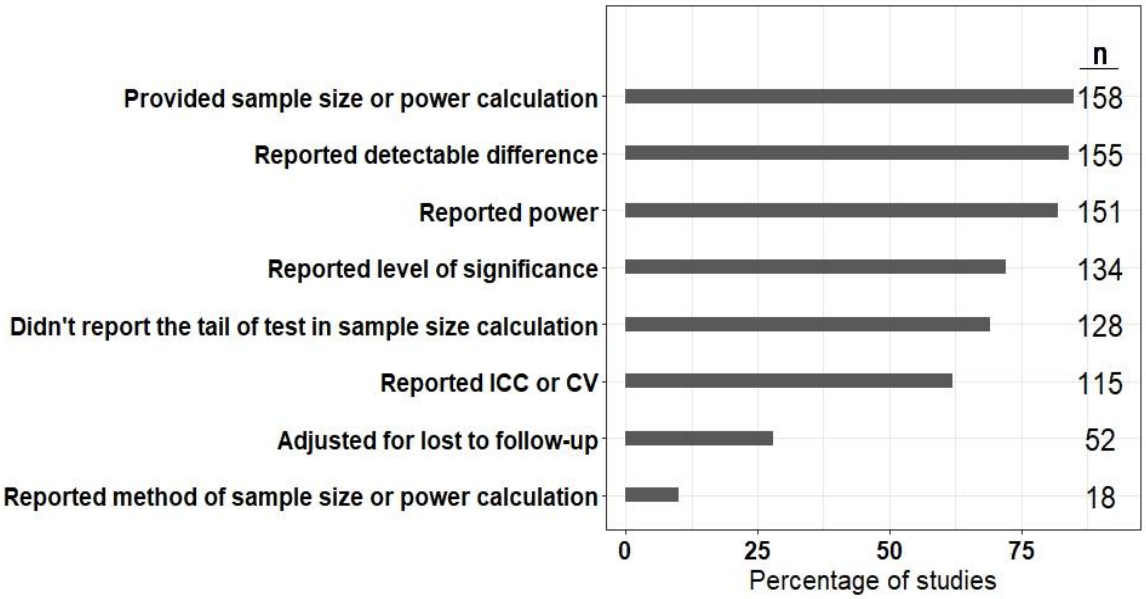

Figure 2: Results of outcomes related to sample size or power calculation among all the studies (n=185)

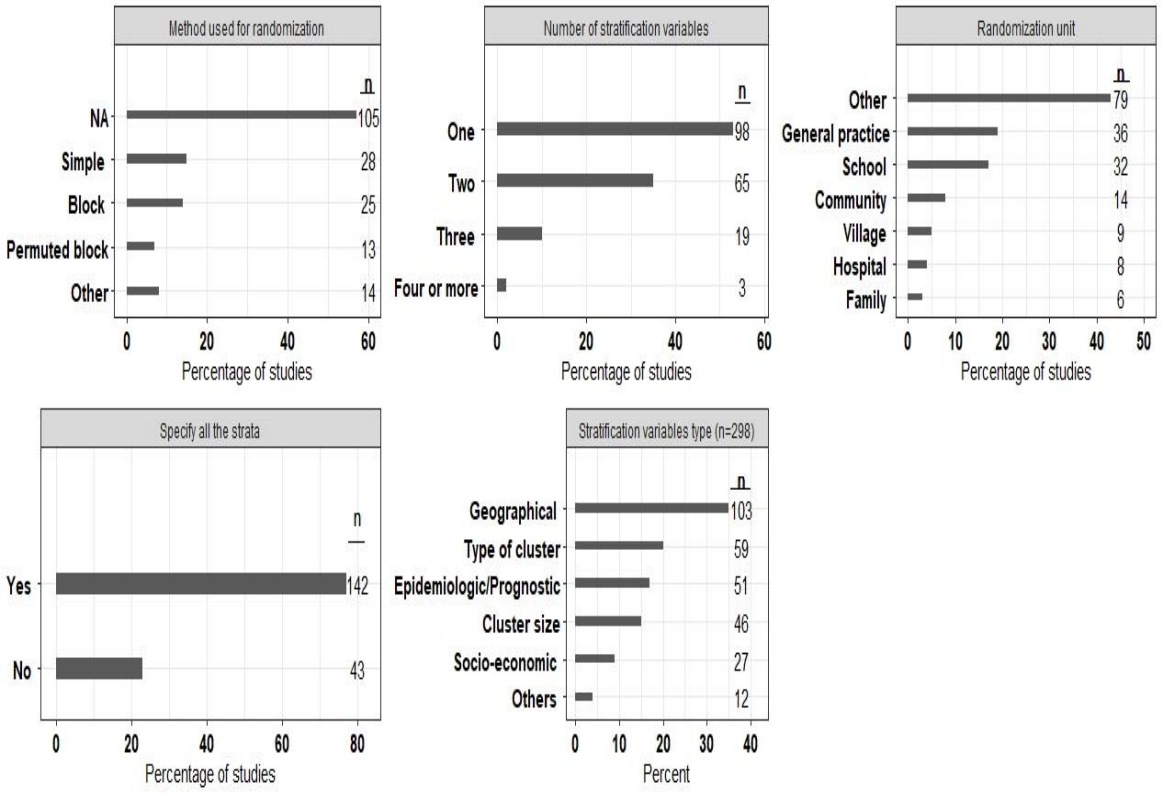

Figure 3: Results of outcomes related to randomization among all the studies (n=185) except type of stratification variables

Table 4: Results of outcomes related to analysis method among all the studies (n=185) except for significance of intervention effect

Outcome	**n=185**
Type of primary outcome; n (%)	
Continuous	64 (35)
Binary	88 (48)
Count	28 (15)
Time to event	3 (2)
Other	2 (1)
Unit of analysis; n (%)	
Cluster-level	28 (15)
Individual-level	143 (77)
Not clear	14 (8)
Primary approach of analysis; n (%)	
Intention-to-treat	96 (52)
Per-protocol	9 (5)
Not available	80 (43)
Primary method/Reported effect estimate adjusted for clustering or stratification; n (%)	
Clustering and stratification	71 (38)
Clustering only	92 (50)
Stratification only	2 (1)
None	20 (11)
Type of adjustment for stratification; n (%) [n=73]	
As a covariate	66 (90)
Stratum specific estimate and then combine	5 (7)
Stratum specific estimate	2 (3)
Imputed missing data; n (%)	
No	150 (81)
Yes	35 (19)
Performed sensitivity analysis; n (%)	
No	127 (69)
Yes	58 (31)
Intervention effect significant; n (%) [2-arm trials only; n=140]	
No	76 (54)
Yes	64 (46)
Significance of intervention effect among those adjusted for both clustering and stratification; n (%) [n=58]	
No	32 (55)
Yes	26 (45)
Significance of intervention effect among those not adjusted for both clustering and stratification; n (%) [n=82]	
No	44 (54)
Yes	38 (46)

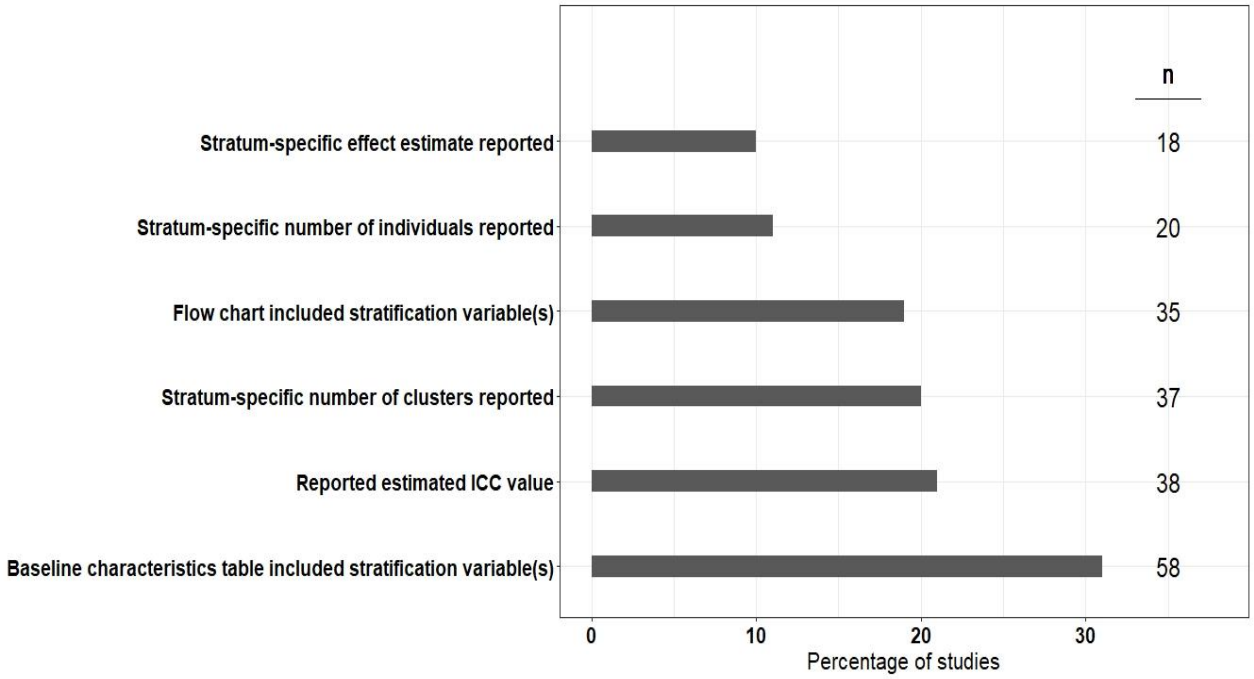

Figure 4: Results of reporting outcomes among all the included studies (n=185)

Appendix: PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	There was no protocol and registration for this systematic survey.
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	4, 10
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-	4

		analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, we didn't assess the risk of bias of individual study.
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no summary of the outcomes.
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no meta-analysis or assessment of heterogeneity.

Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no risk of bias across studies.
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no sensitivity or pre-specified subgroup analyses.
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	4,5
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	5, 12
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, we didn't assess the risk of bias of individual study.

Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was results of individual studies.
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no meta-analysis.
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no risk of bias across studies.
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	This was a systematic survey on reporting analysis of data from stratified cluster randomized trials. Since this was not a clinical systematic review, there was no sensitivity or pre-specified subgroup analyses.

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	5,6
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	6
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	6,7
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	7

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Chapter 3

Contemporary Clinical Trials Communications 15 (2019) 100405

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study

Sayem Borhan[a,b,c,**], Rizwana Mallick[d], Mershen Pillay[e], Harsha Kathard[d], Lehana Thabane[a,b,f,*]

[a] Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
[b] Biostatistics Unit, Research Institute of St Joseph's Healthcare, Hamilton, ON, Canada
[c] Department of Family Medicine, McMaster University, Hamilton, ON, Canada
[d] University of Cape Town, Rondebosch, Cape Town, South Africa
[e] University of KwaZulu Natal, Durban, South Africa
[f] Department of Pediatrics and Anesthesia, McMaster University, Hamilton, ON, Canada

ARTICLE INFO

Keywords:
Stratification
Cluster randomized trial
Sensitivity analysis
Continuous

ABSTRACT

The assessment of the sensitivity of statistical methods has received little attention in cluster randomized trials (CRTs), especially for stratified CRT when the outcome of interest is continuous. We empirically examined the sensitivity of five methods for analyzing the continuous outcome from a stratified CRT - aimed to investigate the efficacy of the Classroom Communication Resource (CCR) compared to usual care to improve the peer attitude towards children who stutter among grade 7 students. Schools – the clusters, were divided into quintile based on their socio-political resources, and then stratified by quintile. The schools were then randomized to CCR and usual care groups in each stratum. The primary outcome was Stuttering Resource Outcomes Measure. Five methods, including the primary method, were used in this study to examine the effect of CCR. The individual-level methods were: (i) linear regression; (ii) mixed-effects method; (iii) GEE with exchangeable correlation structure (primary method of analysis). And the cluster-level methods were: (iv) cluster-level linear regression; and (v) meta-regression. These methods were also compared with or without adjustment for stratification. Ten schools were stratified by quintile, and then randomized to CCR (223 students) and usual care (231 students) groups. The direction of the estimated differences was same for all the methods except meta-regression. The widths of the 95% confidence intervals were narrower when adjusted for stratification. The overall conclusion from all the methods was similar but slightly differed in terms of effect estimate and widths of confidence intervals.

Trialregistration: Clinicaltrials.gov, NCT03111524. Registered on 9 March 2017.

1. Background

Randomization of intact groups, namely clusters, into intervention groups are known as cluster randomized trials (CRT) [1]. Over the years, the number of adopting CRTs is increasing [2]. Diverse types of clusters can be allocated in CRTs including: geographical areas [3]; health care districts [4]; and schools [5]. Like trials on individuals', most CRTs use one of the following three experimental design strategy such as: (a) completely randomized; (b) matched-pair; or (c) stratified. A completely randomized design is satisfactory with substantial number of clusters while stratified design is suitable for small number of clusters [6]. In stratified designs, clusters are randomly allocated to the intervention and control groups within each stratum. For example, Mallick et al. [5] conducted a school-based CRT to investigate the effect of the Classroom Communication Resource (CCR), vs Usual Care, to improve the peer attitude towards children who stutter (CWS). In this trial, schools were first divided into quintile (1–3: lower and 4–5: higher) and stratified as a high vs low school based on the socio-economic resources [5].

Due to the randomization of intact clusters, the outcome from the same cluster may be similar. The intra-cluster correlation coefficient (ICC) is used to measure the degree of similarity [1]. The variance of the estimated intervention effect is inflated due to this correlation and may produce spurious statistically significant results [1,7]. This

* Corresponding author. Biostatistics Unit, Research Institute of St Joseph's Healthcare, 3rd Floor, Martha Wing, Room H-325, 50 Charlton Avenue East, Hamilton, ON, L8N 4A6, Canada.
** Corresponding author. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada.
E-mail addresses: borhana@mcmaster.ca (S. Borhan), thabanl@mcmaster.ca (L. Thabane).

https://doi.org/10.1016/j.conctc.2019.100405
Received 6 March 2019; Received in revised form 20 June 2019; Accepted 3 July 2019
Available online 05 July 2019

inflation can be quantified by the design effect, given by $1 + (\bar{m} - 1)ICC$, where $\bar{m}$ is the average cluster size [1]. Thus, the statistical methods should take into account the potential correlation among the outcomes from the same cluster. Further, the methodologies need to be adjusted for stratification due to stratified design. Researchers have recommended several approaches to analyze the continuous data from completely randomized CRT, which can be extended to stratified designs [1]. The methodologies are broadly classified into two categories: individual- and cluster-level methods. Individual-level methods use the individual-level data such as mixed-model [8] or generalized estimating equation (GEE) [9]. Similarly, we can employ the meta-analytic approach (cluster-level method), which commonly used to combine the results from different studies [10]. This approach helps to aggregate the treatment effects over multiple stratum, like multicentre trials [11–13].

In addition, it is vital to assess the robustness of the results obtained from the randomized controlled trials [14]. The sensitivity analysis helps us to assess the robustness of the results [14]. For CRTs, we can perform several sensitivity analyses. First, we can conduct sensitivity analyses with or without considering the clustering. Secondly, results are compared using different correlation structures [14]. For stratified designs, we can also assess robustness by comparing the methods with or without adjusted for stratification. The GEE with exchangeable correlation structure was used as the primary method of analysis in the Mallick et al. [5] study.

In this study, we empirically examined the sensitivity of methods for analyzing continuous outcome from the stratified CRT using the data from the Mallick et al. [5] study, which in turn demonstrated the robustness of the results obtained using the primary GEE method.

2. Methods

2.1. Overview of the mallick et al. study

The details about the Mallick et al. study can be found elsewhere [5,15]. In brief, this was a cluster randomized trial aimed at examining the effect of Classroom Communication Resource (CCR) on peer attitude towards Children Who Stutter (CWS) in South African schools in the Western Cape. Schools were the unit of randomization and the participants of this trial were the grade 7 students. The selected schools were first stratified to high or low quintile groups and then randomized to CCR or usual care groups. The grade 7 teachers in the intervention group received training on CCR and administered the intervention (including a social story, role-play and facilitated discussion) while participants in the control group received usual curriculum. The participants were assessed 6-month post intervention. The primary outcome was Stuttering Resource Outcomes Measure (SROM) completed at baseline and 6-month post intervention. The study flow chart is presented in Fig. 1.

2.2. Statistical methods

Both individual-level and cluster-level methods were used to analyze the data from the Mallick et al. [5] study. The cluster-and individual-level methods can be adjusted for cluster-level covariates, while individual-level methods can be adjusted for individual-level covariates. The adjustment for stratification covariate, quintile, was applicable for cluster- and individual-level methods, since this was a cluster-level covariate. The results from the analyses were reported in terms of difference (Intervention - Control) along with 95% confidence interval (CI) and associated p-value. All statistical tests were two-sided at the significance level of 0.05. The p-value less than 0.001 were reported as < 0.001 The reporting of the results follows the CONSORT (Consolidated Standards for Reporting Trials) guidelines for reporting cluster-randomized trials [16].

Data were analyzed using both intention-to-treat (ITT) and per-protocol principles. Missing data were imputed using multiple imputation technique assuming missing data follows a missing at random (MAR) pattern. Overall, five datasets were generated, and pooled estimates were reported. All analyses were performed using statistical software R [17].

2.2.1. Individual-level methods

2.2.1.1. Linear regression

The linear regression can be expressed as

$$Y_{ijkl} = \beta_0 + \beta_1 X_{ijkl} + e_{ijkl}$$

Where Y_{ijkl} is the outcome of the l-th subject in the k-th cluster, j-th intervention group and i-th stratum. X_{ijkl} represents the intervention assignment ($X_{ijkl} = 1$ for the treatment group; $X_{ijkl} = 0$ for the control), and e_{ijkl} is the random error assumed to follow a normal distribution with mean 0 and variance σ_e^2. The intercept (β_0) represents the mean outcome for the control group in all clusters, while the slope (β_1) represents the effect of the treatment on the mean outcome.

The linear regression model assumes that data from the participants are independent. This model was implemented using R package lm().

2.2.1.2. Mixed-effects regression model

The mixed-effects regression model is given by

$$Y_{ijkl} = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{ijkl} + C_{ijk} + e_{ijkl}$$

In this model, β_1 and β_2 represents the treatment and stratum effect, respectively, which are fixed. Random cluster effect is represented by C_{ijk}, which follows a normal distribution with mean 0 and variance σ_b^2. The intra-cluster correlation that measures the correlation among the outcomes within cluster is given by $\frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}$, assumed equal for all clusters. We fitted this model using lme4() package in R with restricted maximum likelihood (REML) method [18,19].

2.2.1.3. Generalized estimating equation (GEE)

The generalized estimating equation (GEE) [9] has the advantage of taking into account the correlation of the outcomes through specification of working correlation structure. The estimated treatment effect from the GEE model reflects the both within- and between – cluster relationship [20]. The sandwich covariance estimator yields a robust estimate of treatment effect in the case when the correlation structure is misspecified [21]. Also, small number of clusters leads to an underestimate of variance [22].

For the primary GEE analysis, the exchangeable correlation structure, which based on the assumption that the individuals within the same cluster are equally correlated, was used. Also, this analysis was performed using sandwich method for standard error estimation. This analysis was performed using geepack () package in R.

2.2.2. Cluster-level methods

2.2.2.1. Cluster-level linear regression

This method consists of first estimating a summary measure by cluster such as mean, and then fitting a linear regression based on these summary measures [1].

2.2.2.2. Meta-regression

This is a meta-analytic approach where cluster-level summary is used [10]. This can be extended to perform a stratified analysis on the mean difference in outcome between intervention and control arms within stratum. The overall treatment effect is estimated by a weighted average of individual mean differences across all strata. The principle of inverse-variance weighting is often used [10]. We implemented this method using the metacont() package in R.

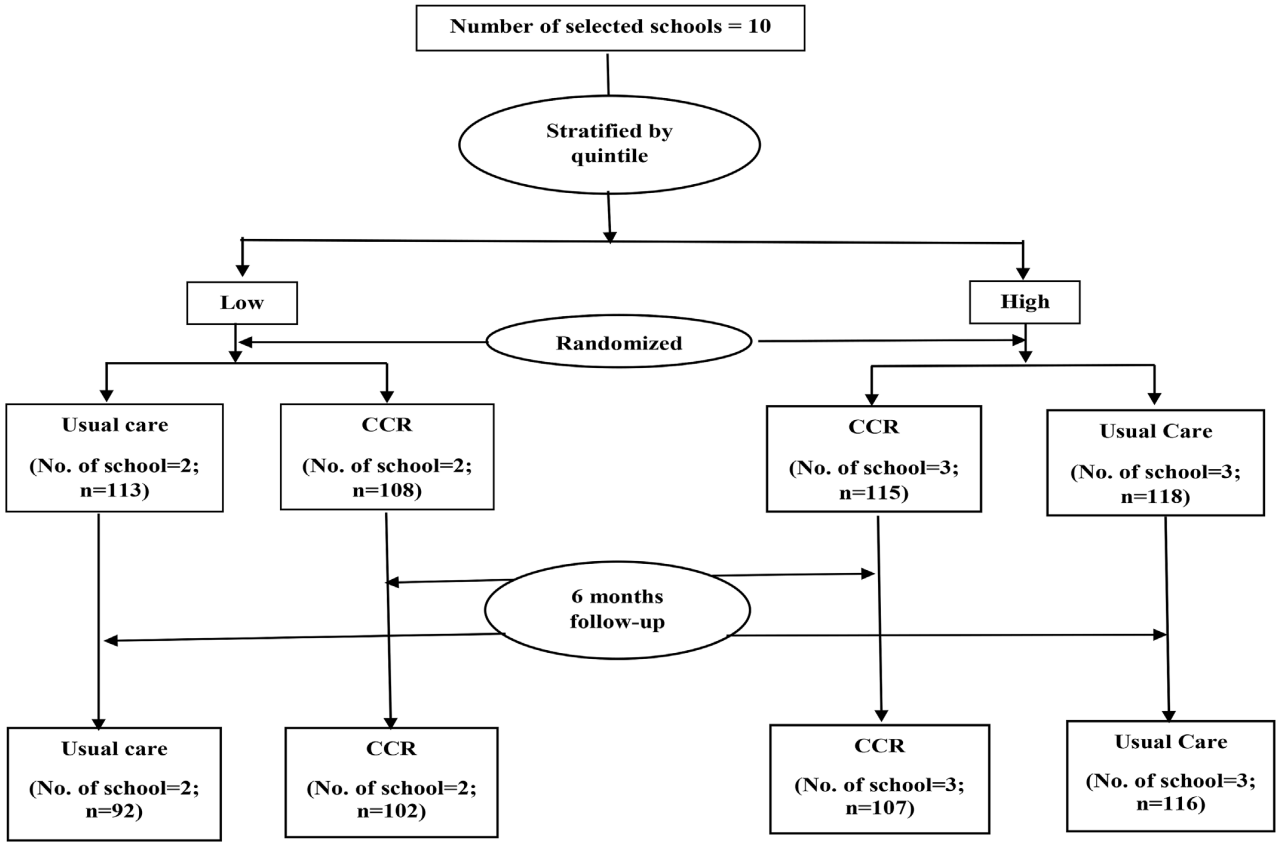

Fig. 1. Study flow chart of the Mallick et al. study.

3. Results

In total, the selected 10 schools were stratified into two groups: higher quintile (6 schools) and lower quintile (4 schools). The schools were then randomized into the intervention CCR group and control usual care group. The average cluster size was 45 (range: 30–54) and 46 (range: 18–68) in the CCR and usual care groups, respectively. Overall, 454 students (223 in the CCR group and 231 in the usual care group) participated in this study. The average age was 13 years for both groups.

We used the methods discussed above (see statistical methods section) to evaluate the effect of intervention. The results of the estimated intervention effect, using ITT, are provided in Fig. 2 with and without adjustment for stratification. Results from all the methods, for the outcome SROM, indicated that the intervention CCR had no statistically significant effect as all the p-values were greater than the nominal level of 0.05 (Fig. 2). The estimated mean differences (MDs) were negative for all the methods except meta regression approach when adjusted for stratification (MD = 0.01[-0.48, 0.50]) (Fig. 2). The p-values for all the methods were similar or lower when adjusted for stratification compared to the same method when not adjusted for stratification, while cluster-level linear regression yielded the lowest p-value (Fig. 2). The magnitude of the widths of the confidence intervals were narrower for cluster-level linear regression (1.06 (when adjusted for stratification); 1.36 (when not adjusted for stratification)) and meta regression (0.98 (when adjusted for stratification); 1.07 (when not adjusted for stratification)) compared to other methods. The widths of the confidence intervals were wider when the methods were not adjusted for stratification compared to the same method adjusted for stratification.

The estimated results of the intervention effect using per-protocol principle are provided in Fig. 3 with and without adjusted for stratification. Similar to ITT analyses, results from per-protocol analyses yielded that the intervention CCR had no statistically significant effect on the outcome SROM as all the p-values were greater than the nominal level of 0.05 for both with and without adjustment for stratification (Fig. 3). The p-values were lower for all the methods when adjusted for stratification (Fig. 3). Also, like ITT, the estimated mean difference was positive (MD = 0.08 [-0.99, 1.15]) for the meta regression method in case of per-protocol analysis. The magnitude of the effect size was higher in the per-protocol analyses compared to ITT analyses except GEE with exchangeable correlation structure (when not adjusted for stratification) (Fig. 3).

For both ITT and per-protocol approaches, the standard errors (SEs) were lower for methods when adjusted for stratification compared to the same method when not adjusted for stratification (results are not presented here).

4. Discussion

In this study, we had empirically investigated the sensitivity of several methods for analyzing continuous outcome from the stratified cluster randomized trial using data from the Mallick et al. [5] study. We used five methods in a frequentist framework to assess the effect of the intervention CCR on SROM compared to usual care. These methods can be differentiated by whether they account the clustering effect or adjust for stratification or both. The overall conclusion, based on intention-to-treat and per-protocol analyses, from all the methods was similar to the primary method (GEE with exchangeable correlation structure) i.e. there was no significant difference between the intervention groups – Classroom Communication Resources (CCR), and the control group –

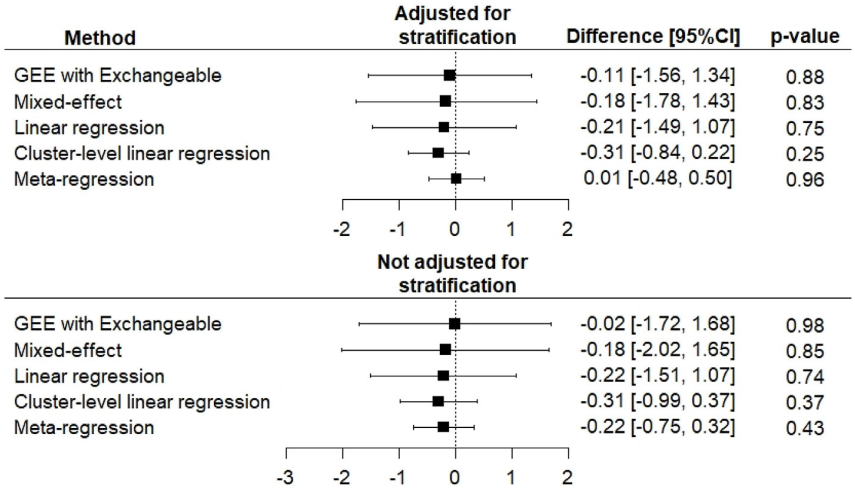

Fig. 2. Results of ITT analyses from different methods with and without adjustment for stratification.

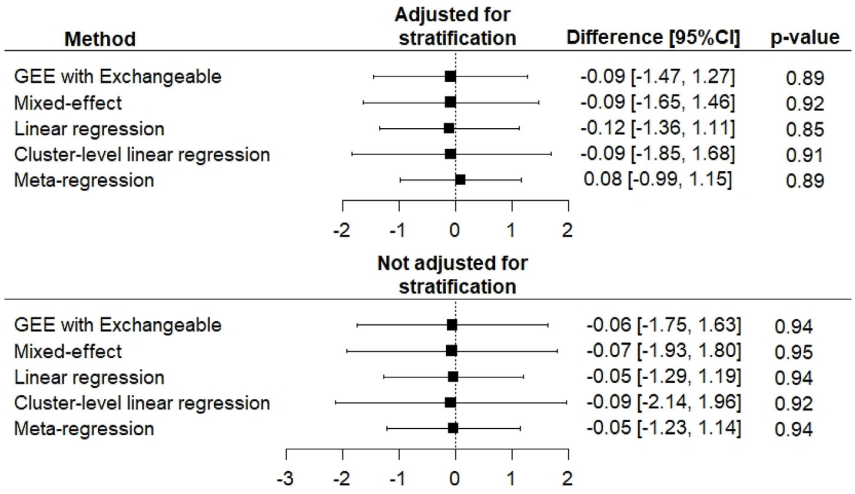

Fig. 3. Results of per-protocol analyses from different methods with and without adjustment for stratification.

usual care, in improving the peer attitude towards CWS.

The conclusion from the linear regression method was matched with other methods, but this method is not appropriate for analyzing data from CRT, as this method does not account the potential correlation among the outcomes from the same cluster. The meta-regression method yielded the narrowest 95% confidence intervals compared to other methods. There is very little variation among the summary measure, mean, of control and intervention groups in low and high stratum i.e. very low heterogeneity, which might lead to yield narrowest confidence interval for meta-regression since the width of the confidence interval decreases as the heterogeneity decreases [23]. Further, the direction of the estimated difference was opposite (positive) for this method compared to other methods when adjusted for stratification. The cluster-level linear regression yielded the widest confidence intervals for per-protocol approach, which are similar to the findings of Walter et al. [22]. However, for ITT approach, the cluster-level methods yielded narrower 95% confidence interval compared to individual-level methods. These results support the findings of Ukoumunne et al. [24] as the authors reported that the cluster-level method performed well, in case of binary data, when ICC is small.

The magnitudes of the estimated differences were similar among the methods with or without adjusted for stratification. However, the widths of the 95% confidence intervals were narrower for adjustment of stratification compared to without adjustment for stratification. These findings matched with the findings of Ma et al. [25] and Kahan et al. [26], where the authors compared several methods for analyzing binary data from stratified CRT and continuous data from stratified randomized controlled trial on individual, respectively. The p-values for all the methods were lower or similar when adjusted for stratification compared to the same method when not adjusted for stratification, which is in line with the findings of Kahan et al. [26].

The failure to adjust for clustering or centre in a multicentre trial results in inflated standard error and wider confidence interval [22,27]. Walters et al. [22] recommended to use cluster-level methods for number of cluster less than 15 per group as individual-level methods may not be reliable in this situation [28,29]. The estimates from the GEE and mixed-effect methods are connected through ICC [30] and in our case the estimates were similar due to the smaller ICC of 0.01.

We compared the results of five methods in several scenarios including: ITT and per-protocol analyses; with and without stratification; and account for potential correlation among the outcomes from the same cluster, which were pertaining to analyze continuous data from stratified CRTs. Moreover, we compared methods based on both individual-level and cluster-level summary data. Sensitivity analyses

might help researchers to make informed decisions, since there is very limited guidance on which method is the best [14]. Furthermore, these analyses help to assess the sensitivity to conclusions to different scenarios such as, with or without clustering. However, we need to be cautious that, like binary data, the interpretation of the treatment effect using the marginal model and the mixed-effect model are may be different [31]. We only considered the multiple imputation technique to impute the missing data. Further investigation using other missing data imputation techniques are warranted.

Based on a simulation study on binary data it has been showed that, the statistical power of GEE is the highest compared to *t*-test, Wilcoxon rank sum test, permutation test, adjusted chi-square test and logistic random-effects model for the analysis of CRTs [32]. However, the estimated variance of from GEE is biased when the number of clusters is small for both binary and continuous data [33–35]. Researchers have reported the need for large number of clusters, 30–40 for mixed models and 40–50 for GEEs, in CRTs [1,36]. Also, some corrections have suggested - for mixed models corrections on degrees-of-freedom and for GEEs corrections to standard error estimations, for analyzing CRTs with small number of clusters [37–41]. Further studies are warranted to investigate how these corrections perform in the case of stratified cluster randomized trials.

5. Conclusion

We have empirically examined the sensitivity of five statistical methods for analyzing continuous outcome from stratified CRTs. The overall conclusions from all methods were similar i.e. no significant effect of the CCR intervention on improving the attitude of peers towards children who stutter. The adjustment for stratification yielded narrower standard errors and confidence intervals, thus it is important to adjust for stratification. Similarly, cluster-level methods yielded narrower confidence intervals compared to individual-level methods. However, further studies are warranted to assess the performance of these methods in wide ranging scenarios.

Funding

There is no funding for this study.

Conflicts of interest

All authors confirm that there are no known conflicts of interest associated with this study and there has been no financial support for this work that could have influenced its outcome.

Acknowledgement

We wish to acknowledge the following for their contributions to the Mallick et al. study: (1) the University of Cape Town; (2) the SA National Research Fund (NRF); and (3) the Carnegie African Diaspora Fellowship Programme (CADFP).

References

[1] A. Donner, N. Klar, Design and Analysis of Cluster Randomization Trials in Health Research, Arnold London, 2000.
[2] J. Bland, Cluster randomised trials in the medical literature: two bibliometric surveys, BMC Med. Res. Methodol. 4 (2004) 21–27.
[3] A. Kroeger, E.V. Avila, L. Morison, Insecticide impregnated curtains to control domestic transmission of cutaneous leishmaniasis in Venezuela: cluster randomized trial, Br. Med. J. 325 (7368) (2002) 810–813.
[4] M. Jordhoy, P. Fayers, T. Saltnes, M. Ahlner-Elmqvist, M. Jannert, S. Kaasa, A palliative-care intervention and death at home: a cluster randomized trial, Lancet 356 (9233) (2000) 888–893.
[5] R. Mallick, H. Kathard, A.S.M. Borhan, M. Pillay, L. Thabane, A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students, Trials 19 (2018) 664.
[6] N. Klar, A. Donner, The merits of matching in community intervention trials: a cautionary tale, Stat. Med. 16 (1997) 1753–1764.
[7] D. Murray, S. Varnell, J. Blitstein, Design and analysis of group-randomized trials: a review of recent methodological developments, Am. J. Public Health 94 (2004) 423–432.
[8] D. Hedeker, R. Gibbons, B. Flay, Random-effects regression models for clustered data with an example from smoking prevention research, J. Consult. Clin. Psychol. 62 (1994) 757–765.
[9] L. Zeger, K.-Y. Liang, P. Albert, Models for longitudinal data: a generalized estimating equation approach, Biometrics 44 (1988) 1049–1060.
[10] A. Whitehead, Meta-analysis of Controlled Clinical Trials, first ed., John Wiley and Sons, Chichester, 2002.
[11] A. Gould, Multi-centre trial analysis revisited, Stat. Med. 17 (15–16) (1998) 1779–1797 discussion 1799-800.
[12] A. Agresti, J. Hartzel, Strategies for comparing treatments on a binary response with multi-centre data, Stat. Med. 19 (8) (2000) 1115–1139.
[13] J. Fleiss, Analysis of data from multiclinic trials, Contr. Clin. Trials 7 (4) (1986) 267–275.
[14] Thabane, et al., A tutorial on sensitivity analyses in clinical trials: the what, why, when and how, BMC Med. Res. Methodol. 13 (1) (2013) 92 2013.
[15] R. Mallick, H. Kathard, L. Thabane, M. Pillay, The Classroom Communication Resource (CCR) intervention to change peer's attitudes towards children who stutter (CWS): study protocol for a randomised controlled trial, Trials 19 (2018) 43.
[16] M. Campbell, D. Elbourne, D. Altman, CONSORT group, CONSORT statement: extension to cluster randomised trials, BMJ 328 (7441) (2004) 702–708.
[17] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018 URL https://www.R-project.org/.
[18] H. Brown, R. Kempton, The application of REML in clinical trials, Stat. Med. 13 (16) (1994) 1601–1617.
[19] R. McLean, W. Sanders, Approximating the degrees of freedom for SE's in mixed linear models, Proceedings of the Statistical Computing Section of the American Statistical Association. New Orleans, Louisiana, 1988.
[20] J. Twisk, Applied Longitudinal Data Analysis for Epidemiology: a Practical Guide, Cambridge University Press, 2003.
[21] K.-Y. Liang, L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[22] S. Walters, C. Morrell, P. Slade, Analysing data from a cluster randomized trial (cCRT) in primary care: a case study, J. Appl. Stat. 38 (10) (2011) 2253–2269.
[23] Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011], in: J.P.T. Higgins, S. Green (Eds.), The Cochrane Collaboration, 2011 Available from:www.handbook.cochrane.org.
[24] O. Ukoumunne, J. Carlin, M. Gulliford, A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials, Stat. Med. 26 (18) (2007) 3415–3428.
[25] Ma, et al., Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT), BMC Med. Res. Methodol. 9 (2009) 37.
[26] B. Kahan, T. Morris, Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis, BMJ 345 (2012) e5840.
[27] R. Chu, L. Thabane, J. Ma, A. Holbrook, E. Pullenayegum, P. Devereaux, Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study, BMC Med. Res. Methodol. 11 (2011) 21.
[28] R. Hayes, L. Moulton, Cluster Randomised Trials, Chapman and Hall/CRC, Boca Raton, FL, 2009.
[29] A. Petrie, C. Sabin, Medical Statistics at a Glance, second ed., Blackwell, Oxford, 2005.
[30] M. Campbell, A. Donner, N. Klar, Developments in cluster randomized trials and statistics in medicine, Stat. Med. 26 (1) (2007) 2–19.
[31] P. FitzGerald, M. Knuiman, Use of conditional and marginal odds-ratios for analysing familial aggregation of binary data, Genet. Epidemiol. 18 (3) (2000) 193–202.
[32] P. Austin, A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes, Stat. Med. 26 (19) (2007) 3550–3565.
[33] R. Prentice, Correlated binary regression with covariates specific to each binary observation, Biometrics 44 (4) (1988) 1033–1048.
[34] L. Mancl, T. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (1) (2001) 126–134.
[35] Leyrat, et al., Cluster randomized trials with a small number of clusters: which analyses should be used? Int. J. Epidemiol. 47 (1) (2018) 321–331.
[36] N. Ivers, M. Taljaard, S. Dixon, et al., Impact of CONSORT extension for cluster randomized trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8, BMJ 343 (2011) d5886.
[37] M. Kenward, J. Roger, Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53 (1997) 983–997.
[38] M. Fay, B. Graubard, Small-sample adjustments for Wald-type tests using sandwich estimators, Biometrics 57 (2001) 1198–1206.
[39] L. Mancl, T. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (2001) 126–134.
[40] P. Li, D. Redden, Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analysing binary outcome in small sample cluster-randomized trials, BMC Med. Res. Methodol. 15 (2015) 38.
[41] P. Li, D. Redden, Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes, Stat. Med. 34 (2015) 281–296.

Chapter 4

Contemporary Clinical Trials Communications 17 (2020) 100539

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: http://www.elsevier.com/locate/conctc

Research paper

An empirical comparison of methods for analyzing over-dispersed zero-inflated count data from stratified cluster randomized trials

Sayem Borhan [a, b, c, d, *], Courtney Kennedy [d], George Ioannidis [d], Alexandra Papaioannou [d, e], Jonathan Adachi [e], Lehana Thabane [a, b, f]

[a] Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
[b] Biostatistics Unit, Research Institute of St Joseph's Healthcare, Hamilton, ON, Canada
[c] Department of Family Medicine, McMaster University, Hamilton, ON, Canada
[d] GERAS Centre, Hamilton Health Sciences, Hamilton, ON, Canada
[e] Department of Medicine, McMaster University, Hamilton, ON, Canada
[f] Departments of Pediatrics and Anesthesia, McMaster University, Hamilton, ON, Canada

ARTICLE INFO

Keywords:
Cluster randomized trial
Stratification
Count
Zero inflated
Overdispersed
Sensitivity

ABSTRACT

Background: The assessment of methods for analyzing over-dispersed zero inflated count outcome has received very little or no attention in stratified cluster randomized trials. In this study, we performed sensitivity analyses to empirically compare eight methods for analyzing zero inflated over-dispersed count outcome from the Vitamin D and Osteoporosis Study (ViD*OS*) – originally designed to assess the feasibility of a knowledge translation intervention in long-term care home setting.

Method: Forty long-term care (LTC) homes were stratified and then randomized into knowledge translation (KT) intervention (19 homes) and control (21 homes) groups. The homes/clusters were stratified by home size (< 250/> = 250) and profit status (profit/non-profit). The outcome of this study was number of falls measured at 6-month post-intervention. The following methods were used to assess the effect of KT intervention on number of falls: i) standard Poisson and negative binomial regression; ii) mixed-effects method with Poisson and negative binomial distribution; iii) generalized estimating equation (GEE) with Poisson and negative binomial; iv) zero inflated Poisson and negative binomial — with the latter used as a primary approach. All these methods were compared with or without adjusting for stratification.

Results: A total of 5,478 older people from 40 LTC homes were included in this study. The mean (=1) of the number of falls was smaller than the variance (=6). Also 72% and 46% of the number of falls were zero in the control and intervention groups, respectively. The direction of the estimated incidence rate ratios (IRRs) was similar for all methods. The zero inflated negative binomial yielded the lowest IRRs and narrowest 95% confidence intervals when adjusted for stratification compared to GEE and mixed-effect methods. Further, the widths of the 95% confidence intervals were narrower when the methods adjusted for stratification compared to the same method not adjusted for stratification.

Conclusion: The overall conclusion from the GEE, mixed-effect and zero inflated methods were similar. However, these methods differ in terms of effect estimate and widths of the confidence interval.

Trial registration: ClinicalTrials.gov: NCT01398527. Registered: 19 July 2011.

1. Background

Randomized trials involving allocation of intact groups or clusters of subjects, instead of independent individuals, are commonly referred to as cluster randomized trials [1]. The rate of adopting cluster randomization trials is increasing [2]. Allocation units are diverse in such studies, and can include families or households, classrooms or schools [3], long-term care homes [4] or even entire communities [5].

Depending on the allocation of clusters, most cluster randomization trials can be classified as using one of three basic types of de-

* Corresponding author. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada.
E-mail address: borhana@mcmaster.ca (S. Borhan).

https://doi.org/10.1016/j.conctc.2020.100539
Received 21 October 2019; Received in revised form 11 January 2020; Accepted 26 January 2020
Available online 1 February 2020

signs: (a) completely randomized, (b) matched-pair, or (c) stratified. Completely randomized designs omit pre-stratification and matching on baseline prognostic factors. This design is most suited for trials enrolling fairly large numbers of clusters [6]. Random assignment of one of the two clusters in a stratum to each intervention group is termed a matched-pair design [6]. The stratified design extends the matched-pair design where more than two clusters are randomly allocated to intervention groups within strata. For example, Vitamin D and Osteoporosis Study (ViD*OS*) [4,7] conducted a pilot stratified cluster randomized trial – where long-term care (LTC) home were stratified by size and profit status, to assess the effect of a multifaceted knowledge translation (KT) intervention on prescribing vitamin D, calcium and osteoporosis medication in long-term care home.

Random allocation of clusters may result in similarity among the outcomes from the same cluster, which is measured using an intra-cluster correlation coefficient (ICC) [1]. This correlation among the responses from the same cluster invalidates the application of statistical techniques which assume independence of observations. Thus, standard statistical methodology needs to be adjusted for this clustering effect, which can be quantified by the design effect, or variance inflation factor, given by $1 + (\bar{m} - 1)\,ICC$, where $\bar{m}$ is the average cluster size [1].

Donner and Klar [1] discussed about several approaches to analyze count data from cluster randomized trials including cluster-specific and population-average extension of Poisson regression. They also discussed we can easily extend these approaches for stratified cluster randomized trials. Similarly, Young et al. [8] compared the performance of cluster-specific and population-average extension of Poisson regression using data from a non-randomized study while Pacheco et al. [9] investigated the performance of methods for analyzing over-dispersed – variance is greater than the mean, count outcome from completely randomized CRT. Further, to account the count outcome with excess zeros we need to use the zero-inflated models. To the best of our knowledge, no study examined the methods for analyzing over dispersed and zero-inflated count data from stratified cluster randomized trials.

On the other hand, Thabane et al. [10] rightfully emphasized the importance of performing a sensitivity analysis, which help us to assess the robustness of the results. For cluster randomized trials we can perform sensitivity analyses with or without taking clustering into account. We can also compare the methods with or without considering the stratification. Borhan et al. [11] examined the sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials and found the overall conclusion from all the methods were similar.

In this study, we performed sensitivity analyses to empirically compare eight methods for analyzing zero inflated over-dispersed count outcome from the ViD*OS* study [4].

2. Methods

2.1. Motivating example: ViDOS study

We used the data from an LTC-based pilot stratified cluster randomized trial – details can be found elsewhere [4,7], for this study. A total of 5,478 older people from 40 LTC homes (19 Intervention and 21 Control) were randomized into two groups KT intervention and control groups. The LTC homes were stratified by size (<250 vs ≥ 250 beds) and profit status (profit vs non-profit). Seven LTC homes withdrew before the study began. The outcome, number of falls were measured at 6- and 12-month post-randomization. For this study, we used the number of falls measured at 12-month. The variance of the number of falls is greater than the mean number of falls (variance = 6 > mean = 1). Similarly, for each cluster the mean number of falls is smaller than the variance of the number of falls. Thus, the number of falls was over-dispersed. Further, the number of falls was zero inflated as 72% and 46% of the number of falls were zero in the control and intervention groups, respectively.

2.2. Statistical analysis methods

Both cluster-specific (mixed-effect method) and population-average (generalized estimating equation) methods were used to analyze the number of falls from the ViD*OS* study. The mixed-effect zero-inflated negative binomial model was considered as the primary method since it can take into account both overdispersion and zero-inflation as well as clustering. The adjustment for stratification covariates – home size and profit status, were applicable for cluster- and individual-level methods, since these were cluster-level covariates. The results from the analyses were reported in terms of the incidence rate ratios (IRRs) along with 95% confidence intervals (CIs) and associated p-values. All statistical tests were two-sided at the significance level of 0.05. The p-value less than 0.001 were reported as <0.001 The reporting of the results follows the CONSORT (Consolidated Standards for Reporting Trials) guidelines for reporting cluster-randomized trials [12].

Data were analyzed using Intention-to-treat (ITT) principles and missing data analysis approach – where missing data were imputed using multiple imputation technique assuming missing data follows a missing at random (MAR) pattern. Overall, five datasets were generated, and pooled estimates were reported.

2.3. Standard Poisson/Negative binomial (NB) model

The standard Poisson and negative binomial model for count data is given by

$$\log(E(Y_{ijkl}) = \mu_{ijkl}) = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{1ijkl} + \beta_3 S_{2ijkl} + e_{ijkl}$$

Where, Y_{ijkl} is the outcome, number of falls, of the $i-th$ subject of the $j-th$ cluster in the $k-th (k = 0,1)$ and $l-th (l = 0,1)$ stratum. X_{ijkl} is the intervention (0: Control; 1: KT Intervention). S_{1ijkl} (0: <250; 1> = 250) is the home size and S_{2ijkl} (0: Non-profit; 1: Profit) is the profit status of the cluster.

Here, β_1 represents the treatment effect while β_2 and β_3 represents the two strata effect corresponding to home size (0: <250; 1: ≥250) and profit status (0: Non-profit; 1: Profit), respectively.

We considered two distributional assumptions for number of falls:

(a) Number of falls follows a Poisson distribution i.e. $Y_{ijkl} \sim Poi(\mu_{ijkl})$, with variance function $V(Y_{ijkl}) = \varphi v(\mu_{ijkl}) = \mu_{ijkl}$, where φ is assumed to be 1 i.e. mean and variance are equal.
(b) Number of falls follows a Negative Binomial (NB) distribution i.e. . $Y_{ijkl} \sim NB(s, \mu_{ijkl})$, with variance function $V\left(Y_{ijkl}\right) = \varphi v\left(\mu_{ijkl}\right) = \varphi\left(\mu_{ijkl} + s\mu_{ijkl}^2\right)$, where φ is assumed to be 1 and s is the overdispersion parameter indicating that the NB distribution models overdispersion implicitly by its parameter s. The NB distribution is preferred when there is overdispersion in the data i.e. mean < variance.

The standard Poisson and negative binomial model were fitted using glm() and glm.nb() in R [13].

2.4. Mixed-effect model (Poisson/Negative binomial)

The mixed-effect model for count data is given by

$$\log(E(Y_{ijkl}) = \mu_{ijkl}) = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{1ijkl} + \beta_3 S_{2ijkl} + C_{ijk} + e_{ijkl}$$

In this model, like the previous model, β_1 represents the treatment effect while β_2 and β_3 represents the two stratum effect corresponding

to home size (0: < 250; 1: ≥250) and profit status (0: Non-profit; 1: Profit), respectively, which are fixed. Random cluster effect is represented by C_{ijk}, which follows a normal distribution with mean 0 and variance σ_b^2. The intra-cluster correlation that measures the correlation among the outcomes within cluster is given by $\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$, assumed equal for all clusters. β_1 is the log of the Rate Ratio (RR) of the intervention X_{ijkl} (0 = Control, 1 = KT Intervention). We used glmer() and glmer.nb() in R to fit mixed-effect with Poisson and negative binomial, respectively.

2.5. Generalized estimating equation (GEE) (Poisson/Negative binomial)

The GEE model for count data is given by

$$\log(E(Y_{ijkl}) = \mu_{ijkl}) = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{1ijkl} + \beta_3 S_{2ijkl}$$

Like before, β_1 represents the treatment effect while β_2 and β_3 represents the two stratum effect corresponding to home size (0: < 250; 1: ≥250) and profit status (0: Non-profit; 1: Profit), respectively. Similar to mixed-effect method we considered two distributional assumption for count data: Poisson and negative binomial. For GEE method we considered exchangeable working correlation structure. GEE with Poisson was fitted using geeglm() in R while GEE with negative binomial was fitted using PROC GENMOD in SAS [14]. GEE with negative binomial was the primary method of analysis.

2.6. Zero inflated models (Poisson/Negative binomial)

For zero inflated models the distribution of Y_{ijkl} is

$$Y_{ijkl} = \begin{cases} 0\,; & \text{with probability } \varphi_{ijkl} \\ \text{Poisson or NB}\left(\mu_{ijkl}\right); & \text{with probability } \left(1-\varphi_{ijkl}\right) \end{cases}$$

The mixed-effect zero inflated Poisson or negative binomial model is given by:

$$logit(\varphi_{ijkl}) = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{1ijkl} + \beta_3 S_{2ijkl} + C_{ijk} + e_{ijkl}$$

$$\log(E(Y_{ijkl}) = \mu_{ijkl}) = \beta_0 + \beta_1 X_{ijkl} + \beta_2 S_{1ijkl} + \beta_3 S_{2ijkl} + C_{ijk} + e_{ijkl}$$

The zero inflated Poisson and negative binomial models were fitted using the R package GLMMadaptive.

3. Results

Overall 40 clusters were randomized into KT intervention (19 clusters) and control (21 clusters) groups. The clusters were stratified by two variables cluster size and profit status. The average cluster size in the KT group was 115 (minimum = 43, maximum = 294) while the average cluster size in the control group was 157 (minimum = 49, maximum = 375). At the end of the follow-up there were 2,209 participants in the intervention group and 3,382 participants in the control group. The average age of the participants in both groups were 84 years while approximately 70% were female.

We used the methods discussed above to assess the effect of KT intervention on number of falls with mixed-effect zero-inflated with negative binomial distribution as the primary method of analysis. The results of the ITT analyses with or without adjusted for stratification are given in Fig. 1. The direction of the effect estimate incidence rate ratios were similar for all the methods. The standard Poisson and negative binomial regression methods yielded statistically significant results as p-values lower than the nominal level of 0.05 while the other methods yielded non-significant results (Fig. 1). The estimated IRRs varies from 1.11 to 1.37 when adjusted for stratification and 1.03 to 1.49 when not adjusted for stratification. The effect estimates IRRs were slightly higher for mixed-effect methods compared to other methods. The magnitude of the widths of the 95% confidence intervals were higher for mixed-effect Poisson and negative binomial methods compared to other methods when adjusted or not adjusted for stratification (Fig. 1). The Akaike's Information Criteria (AIC) were slightly lower when the methods adjusted for stratification compared to without such adjustment. Further, the AIC values were lower for negative binomial models (8391.00 and 8333.24 for mixed-effect and zero-inflated negative binomial models respectively) compared to GEE models (10858.00 and 9093.10 for mixed-effect and zero-inflated Poisson models respectively).

The results of the missing data analysis were given in Fig. 2. Unlike ITT approach, standard Poisson and negative binomial did not yield statistically significant results (Fig. 2). Similar to ITT approach, direction of effect estimate for all the methods were similar. The estimated IRRs varies from 1.35 to 2.12, when adjusted for stratification and 1.41 to 1.96 when not adjusted for stratification. The magnitudes of the widths of the 95% confidence intervals were higher for all methods compared to ITT approach. Similar to ITT 95% confidence intervals were wider for mixed-methods, when not accounted for zero inflation, compared to other methods (Fig. 2).

For all methods, the estimated IRRs were very similar with or without adjusting for stratification for both ITT and missing data analysis approaches (Figs. 1-2). Further, it is noticeable, that the estimated IRRs were slightly higher, for all methods, in missing data analysis approach compared to ITT approach (Figs. 1-2). Also, for ITT approach, the 95% confidence intervals were slightly narrower when

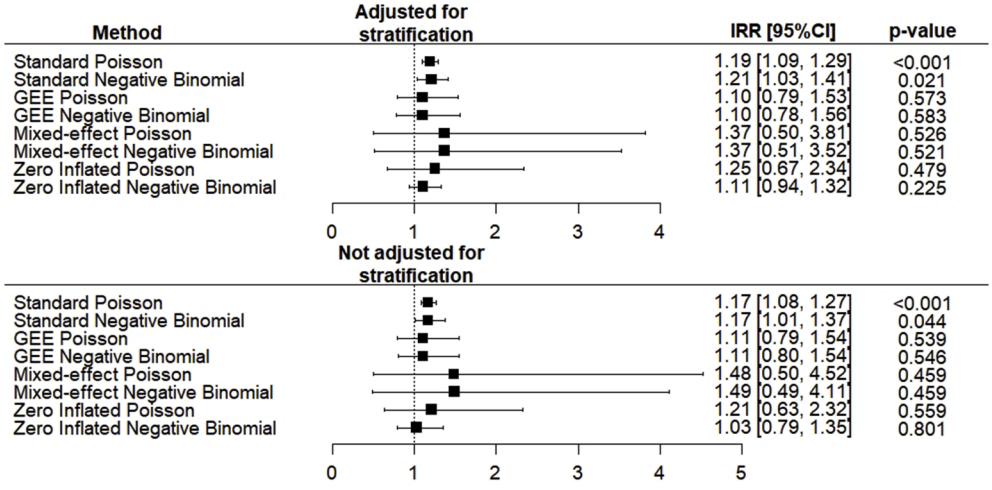

Fig. 1. Results of ITT analysis using different methods with/without adjusted for stratification.

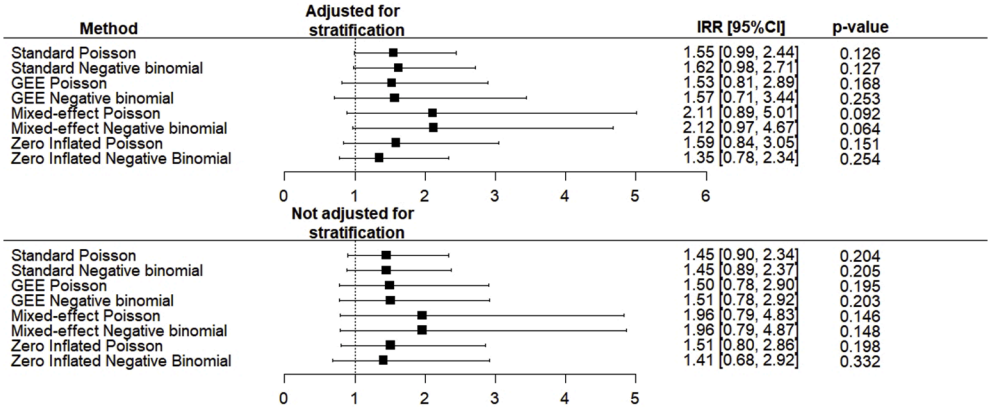

Fig. 2. Results of missing data analysis using different methods with/without adjusted for stratification.

adjusted for stratification (Fig. 1). The difference among the methods in terms of p-values were smaller for missing data analysis approach compared to ITT approach (Figs. 1 and 2).

4. Discussion

In this study, we empirically investigate the methods for analyzing overdispersed zero inflated count outcome from stratified cluster randomized trial using data from the Vi*DOS* study – which was designed to investigate the effect of a KT intervention. We compared eight methods to assess the effect of KT intervention on number of falls. The direction of effect of estimate incidence rate ratios (IRRs) were similar for all methods for both adjusted and not adjusted for stratification. The conclusions from both ITT and missing data analyses indicated that, KT intervention had no effect on number of falls.

For ITT analyses, both standard Poisson and negative binomial methods yielded statistically significant results that the RRs of number of falls were slightly higher in the intervention group compared to control group. However, these two methods were not appropriate for analyzing count data from CRT as these methods do not take into account the degree of similarity among the outcomes from the same cluster.

In this study, we considered mixed-effect with zero-inflated negative binomial as the primary method of analysis to assess the effect of KT intervention on over dispersed number of falls. We performed sensitivity analyses to examine the robustness of the findings of the primary method. The overall conclusion from all the methods were similar. These findings match with the findings of the Borhan et al. [11] when they investigated the sensitivity of several methods for analyzing continuous outcome from the stratified CRT.

Overall, for all methods, the estimated IRRs and the corresponding widths of the 95% confidence intervals were slightly lower for ITT analyses compared to missing data analyses. GEE and mixed-effect with Poisson and negative binomial distributions, respectively, yielded approximately similar IRRs. The estimated IRRs and widths of the 95% confidence intervals were lower for zero inflated models compared to mixed-effect methods with Poisson and negative binomial distribution. The widths of the 95% confidence intervals were lower for GEE methods compared to mixed-effect methods for both ITT and missing data analyses. This is consistent with the findings of Pacheco et al. [9]. The authors reported that, GEE yielded the highest power and narrow CIs when the authors investigated the performance of methods for analyzing overdispersed count data from CRT. However, GEE underestimate the covariance among observations yielding downward biased standard errors when the number of clusters is small [15]. Also, we need to be cautious that, GEE method yields elevated type I error rates in small sample situations (<40 clusters) [9].

We also compared the methods with or without adjusting for stratification. Zero inflated negative binomial yielded the lowest IRRs and narrowest 95% confidence intervals when adjusted for stratification among the valid methods. For ITT approach, the estimated IRRs and the widths of the 95% confidence intervals were almost similar or lower for both GEE methods. Similarly, for mixed-effect methods the estimated RRs and the magnitude of the widths of the 95% confidence intervals were slightly lower when we adjusted for stratification. These findings matched with the findings of Borhan et al. [11], Ma et al. [16] and Kahan et al. [17], where the authors compared several methods for analyzing continuous and binary data from stratified CRT and continuous data from stratified randomized controlled trial on individual, respectively. Similarly, for missing data approach, GEE yielded the similar results with or without adjusted for stratification. For all methods, the p-values were lower when adjusted for stratification compared to same method when not adjusted for stratification and matched with the findings of Kahan et al. [17].

The major strength of this study that, we empirically examined eight methods, including both cluster-specific and population-average methods, for analyzing count outcome from a stratified CRT - Vi*DOS* study, under different scenarios including accounting for clustering and adjusting for stratification. We also compared the methods through ITT approach and imputing the missing data. In addition, we used appropriate method such as negative binomial to account for overdispersion and zero inflated models to account for excess zeros. Thus, this study will guide researchers about the sensitivity of these methods since there is no study, to the best of our knowledge, investigate the performance of these methods for analyzing count data from stratified CRT.

The major limitation of this study, that Vi*DOS* study was a pilot trial designed to investigate the feasibility of the KT intervention. However, Vi*DOS* was stratified by two cluster-level covariates cluster size and profit status, which is very rare in real life. It is possible that, we might have missed some falls data as it is difficult to measure the number of falls and varies between LTCs.

Data from 7 clusters were missing in the intervention group as 6 clusters declined to actively participate after randomization and 1 cluster withdrew after baseline measurement. Further study on missing data imputation techniques when the whole cluster is missing would be an important addition. Furthermore, a well-designed simulation study is warranted to examine the performance of these methods under different scenarios. It requires large number of clusters (>30) to get valid estimate using GEE and mixed-effect methods [18–21]. Researchers have suggested some corrections to address the requirement of large number of clusters [22–26] which can be extended to stratified CRT, especially when the outcome is count.

5. Conclusion

In this study, we empirically compared the eight methods for analyzing count outcome using the data from ViD*OS* study - a pilot stratified cluster randomized trial. The overall conclusion from all the methods were similar that the KT intervention had no effect on number of falls. The zero inflated negative binomial model yielded the lowest IRR and narrowest 95% confidence interval, when adjusted for stratification, compared to GEE and mixed-effect methods. A well-designed simulation study is warranted to assess the performance of these methods.

Acknowledgements

ViDOS study was supported by an operating grant from the Canadian Institutes of Health Research (Funding Reference Number: MOP-114982).

References

[1] A. Donner, N. Klar, Design and Analysis of Cluster Randomization Trials in Health Research, Arnold London, 2000.
[2] J. Bland, Cluster randomised trials in the medical literature: two bibliometric surveys, BMC Med. Res. Methodol. 4 (2004) 21–27.
[3] R. Mallick, H. Kathard, A.S.M. Borhan, M. Pillay, L. Thabane, A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students, Trials 19 (2018) 664.
[4] Kennedy, et al., Successful knowledge translation intervention in long-term care: final results from the vitamin D and osteoporosis study (ViDOS) pilot cluster randomized controlled trial, Trials 16 (2015) 214.
[5] J. Kaczorowski, et al., Cardiovascular Health Awareness Program (CHAP): a community cluster-randomised trial among elderly Canadians, Prev. Med. 46 (6) (2008 Jun) 537–544.
[6] N. Klar, A. Donner, The merits of matching in community intervention trials: a cautionary tale, Stat. Med. 16 (1997) 1753–1764.
[7] C.C. Kennedy, G. Ioannidis, L.M. Giangregorio, J.D. Adachi, L. Thabane, S.N. Morin, et al., An interdisciplinary knowledge translation intervention in long-term care: study protocol for the vitamin D and osteoporosis study (ViDOS) pilot cluster randomized controlled trial, Implement. Sci. 7 (2012) 48.
[8] M. Young, J. Preisser, B. Qaqish, M. Wolfson, Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials, Stat. Methods Med. Res. 16 (2007) 167–184.
[9] Pacheco, et al., Performance of analytical methods for overdispersed counts in cluster randomized trials: sample size, degree of clustering and imbalance, Stat. Med. 28 (2009) 2989–3011.
[10] Thabane, et al., A tutorial on sensitivity analyses in clinical trials: the what, why, when and how, BMC Med. Res. Methodol. 13 (1) (2013) 92 2013.
[11] S. Borhan, R. Mallick, M. Pillay, H. Kathard, L. Thabane, Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study, Contemp. Clin. Trials Commun. 15 (2019) 100405.
[12] M. Campbell, D. Elbourne, D. Altman, CONSORT group, CONSORT statement: extension to cluster randomised trials, BMJ 328 (7441) (2004) 702–708.
[13] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019 URL. https://www.R-project.org/.
[14] SAS/STAT_ *User's Guide*, SAS Institute, Inc., 2019.
[15] L.A. Mancl, T.A. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (1) (2001) 126–134.
[16] Ma, et al., Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT), BMC Med. Res. Methodol. 9 (2009) 37.
[17] B. Kahan, T. Morris, Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis, BMJ 345 (2012) e5840.
[18] R. Prentice, Correlated binary regression with covariates specific to each binary observation, Biometrics 44 (4) (1988) 1033–1048.
[19] L. Mancl, T. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (1) (2001) 126–134.
[20] Leyrat, et al., Cluster randomized trials with a small number of clusters: which analyses should be used?, Int. J. Epidemiol. 47 (1) (2018) 321–331.
[21] N. Ivers, M. Taljaard, S. Dixon, et al., Impact of CONSORT extension for cluster randomized trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8, BMJ 343 (2011) d5886.
[22] M. Kenward, J. Roger, Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53 (1997) 983–997.
[23] M. Fay, B. Graubard, Small-sample adjustments for Wald-type tests using sandwich estimators, Biometrics 57 (2001) 1198–1206.
[24] L. Mancl, T. DeRouen, A covariance estimator for GEE with improved small-sample properties, Biometrics 57 (2001) 126–134.
[25] P. Li, D. Redden, Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analysing binary outcome in small sample cluster-randomized trials, BMC Med. Res. Methodol. 15 (2015) 38.
[26] P. Li, D. Redden, Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes, Stat. Med. 34 (2015) 281–296.

Chapter 5

Performance of methods for analyzing continuous data from stratified cluster randomized trials – a simulation study

Sayem Borhan[1,2,3], Jinhui Ma[1], Alexandra Papaioannou[1,4,5], Jonathan Adachi[5], Lehana Thabane[1,2,6]

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

[2] Biostatistics Unit, Research Institute of St Joseph's Healthcare, Hamilton, ON, Canada

[3] Department of Family Medicine, McMaster University, Hamilton, ON, Canada

[4] GERAS Centre, Hamilton Health Sciences, Hamilton, ON, Canada

[5] Department of Medicine, McMaster University, Hamilton, ON, Canada

[6] Departments of Pediatrics and Anesthesia, McMaster University, Hamilton, ON, Canada

Correspondence:

Sayem Borhan

Department of Health Research Methods, Evidence, and Impact

McMaster University

1280 Main Street West, Hamilton, ON

L8S 4K1

Canada

Email: borhana@mcmaster.ca

Abstract

Background

Adoption of cluster randomized trials (CRTs) with stratified design is increasing. While we know that the number of clusters have substantial impact on the performance of statistical analysis methods, but we have limited knowledge about the performance of methods for analyzing data from stratified CRTs. In this simulation study, we evaluated the performance of several commonly used methods for analyzing continuous data from stratified CRTs with a single stratification variable.

Methods

We compared 4 methods: mixed-effect, generalized estimating equation (GEE), cluster-level (CL) linear regression and meta-regression to analyze the continuous data from stratified CRTs using a simulation study with varying number of clusters, cluster sizes, and intra-cluster correlation coefficients (ICCs). We considered a stratified CRT with one stratification variable with two strata. Total number of clusters were equally divided into two strata. The performance of the methods was evaluated in terms of type I error rate, statistical empirical power, root mean squared error (RMSE), and width of the 95% confidence interval (CI).

Results

GEE and meta-regression methods yielded more than or approximately 10% type I error rates for study with small number of clusters. GEE had higher or similar power

compared to other methods. All methods had similar accuracy, measured through root mean square error, except meta-regression. Similarly, all methods but meta-regression had similar widths of the 95% CI. Performance of all methods worsened as the ICC increased to 0.06 from 0.03.

Conclusions

The performance of all methods improved as the number of clusters increased along with cluster sizes. Meta-regression was the least powerful and least efficient compared to other methods.

Key words: Cluster randomized trials, Stratified, Simulation, Continuous

Background

Cluster randomization trials (CRTs) involve randomization of intact clusters, rather than individual participants, into intervention groups [1]. The type of clusters can be distinct including: geographical region [2], health care area [3], and schools [4]. Over the years, the number of adopting CRTs is increasing [5] as well as the number of CRTs with stratified design, which is suitable when the number of clusters is small [6]. In stratified designs, clusters are randomly allocated to the intervention and control groups within each stratum. For example, Mallick et al [4] conducted a school-based stratified CRT, where schools were first divided into quintile (1-3: lower and 4-5: higher) based on socio-economic resources and then stratified into low versus high quintile. Schools within each stratum were then randomly allocated to intervention and control groups [4].

Randomization of intact clusters may lead to the situation where outcomes from the same cluster may be similar. This similarity inflated the variance of the estimated intervention effect and failure to account this similarity may yield false significant intervention effect [1, 7]. This similarity or clustering is measured by intra-cluster correlation coefficient (ICC) and statistical methods should adjust for this clustering [1]. For stratified design, these methods need to further adjust for stratification [1]. Both individual-level (based on individual-level data) and cluster-level (based on cluster level summary) methods can be used to examine the effect of intervention from the stratified CRTs. The individual-level methods include: mixed-effect model [8] or generalized estimating equation (GEE) [9], while cluster-level methods include: cluster-level linear regression or meta-analytic approach [10] – which can be used to assess the treatment

effects over strata [11-13]. Chu et al [14] compared the performance of several methods including meta-regression to analyze continuous data from multicentre randomized controlled trials.

While it is important to adjust for stratification only 26% of the randomized controlled trials adjusted the primary analysis for the balancing factors [15]. In addition, from a recent systematic survey conducted by our team – where we added the term 'strati*' with the search terms suggested by Taljaard et al [16] to identify the stratified CRTs from the database MEDLINE, since the inception to July 2019, we found that, only 38% of the 185 selected studies adjusted the primary method for both clustering and stratification to assess the intervention effect from stratified CRTs [17]. Failure to adjust for stratification leads to wider confidence intervals and larger p-values [15,17-19]. Borhan et al [18, 19] empirically compared several methods for analyzing continuous and count data from stratified CRTs using data from the Mallick et al [4] and Vi*DOS* study [20], respectively. Moreover, researchers have investigated the performance of methods from completely randomized CRTs [21-26]. Klar and Darlington [22] investigated the performance of several mixed-effect methods to analyze pretest-posttest continuous data from completely randomized CRTs incorporating the individual-and cluster-level associations. On the other hand, Borhan et al [21] and Austin P [23] investigated the performance of methods for analyzing binary data. Based on our systematic survey [17], it is evident that, we have limited evidence about the performance of the methods to assess the intervention effect from stratified CRTs, especially when the outcome of interest was continuous.

In this study, we conducted a simulation study to examine the performance of methods for assessing the intervention effect from stratified CRTs. We evaluated several methods, in terms of type I error rate, empirical power, root mean square error rate, and width of 95% confidence intervals, for analyzing continuous data from stratified CRTs.

Methods

This was a simulation study, where we evaluated the performance of several methods for assessing the intervention effect, when the outcome of interest was continuous from stratified CRTs.

Statistical methods

Both individual-level and cluster-level methods were used to assess the intervention effect. These methods were adjusted for stratification.

Individual-level methods

Mixed-effects regression model (mixed-effect)

The mixed-effects regression model is given by

$$Y_{ijks} = \beta_0 + \beta_1 X_{ijks} + \beta_2 S_{ijks} + C_{ijk} + e_{ijks}$$

Where Y_{ijks} is the outcome of the i-th subject in the j-th cluster, k-th intervention group and s-th stratum. X_{ijks} represents the intervention assignment (X_{ijks}=1 for the treatment group; X_{ijks}=0 for the control), S_{ijks} represents the dichotomous stratification variable with value 0 and 1, and e_{ijks} is the random error assumed to follow a normal distribution with mean 0 and variance σ_e^2.

In this model, β_1 and β_2 represents the treatment and stratum effect, respectively, which are fixed. The random cluster effect is given by C_{ijk}, which follows a normal distribution with mean 0 and variance σ_b^2. The ICC represents the correlation between two randomly chosen subjects in the same cluster. A single common ICC given by $\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$ was assumed for all clusters. The R package lme4() was used to fit this model with restricted maximum likelihood (REML) method [26, 27].

Generalized estimating equation (GEE)

The generalized estimating equation (GEE) model is given by

$$E(y_{ijks}) = \beta_0 + \beta_1 X_{ijks} + \beta_2 S_{ijks}$$

Like mixed-effect model, β_1 and β_2 represents the treatment and stratum effect, respectively. The working correlation structure in the GEE model take into account the correlation among the outcomes from the same cluster and the sandwich covariance estimator yields a robust estimate of the treatment effect even if the correlation structure is

mis-specified [28]. In the GEE analysis, we assumed the correlation structure followed an exchangeable pattern. R package geepack() was used to fit the GEE model.

Cluster-level methods

Cluster-level linear regression (CL Linear Regression)

Cluster-level method is based on cluster-level summary measure, such as mean [1]. We first calculated the mean for each cluster, then a linear regression was fitted, adjusted for stratification, using these mean values.

Meta-regression

The meta-regression approach is based on cluster-level summary measure [10]. We extended this method for stratified design and used the mean difference, in outcomes, between the intervention and control arms within each stratum. Random-effect model was used to estimate the treatment effect and was conducted using the R package metacont().

Simulation study

We conducted a simulation study, to assess the performance of the statistical methods to analyze the continuous outcome from the stratified cluster randomized trials, using the approach adopted by Arnold et al [29] and Moerbeek & Schie [30]. We

considered a stratified design with one stratification variable with two strata. The outcome, Y, was simulated, separately for each stratum, using the following mixed-effects linear regression model: $Y_{ijk} = \beta_0 + \beta_1 X_{ijk} + C_{ij} + e_{ijk}$; where, Y_{ijk} is the outcome of the i-th subject in the j-th cluster, in the k-th intervention group; $X_{ijk}(= 0,1)$ represent the dummy variable for treatment allocation ($i = 1, \dots, n_j; j = 1, \dots, J$); C_{ij} is the cluster-level random effect while e_{ijk} is the individual-level random error term. Both C_{ij} and e_{ijk} follow normal distributions with mean 0 and standard deviations σ_c and σ_e, respectively. Random effects and error term related to the ICC as ICC($= \frac{\sigma_c^2}{\sigma_c^2+\sigma_e^2}$) is the ratio of between cluster variance to the total variance [1]. Without loss of generality, the total variability was fixed at $\sigma^2 = \sigma_c^2 + \sigma_e^2 = 1$ and $\beta_0 = 0$. The other parameters for this simulation study such as, the number of clusters, cluster sizes, intervention effect size, and ICC, were selected, given in Table 1, based on the studies that had continuous outcome as the primary outcome from our recently conducted systematic survey [17]. For each of these designs, 1000 simulations were run for each combination of $\beta_1 = 0, 0.11$; number of clusters =6, 24, 34, 68; number of individuals per cluster=5,10,15,20,25,30,35,40,45,50 and ICC=0.03, 0.06. This simulation study was conducted using R [31].

Comparison of methods

We applied the methods, discussed in the statistical methods section, to assess the effect of intervention for each simulated data set. The following quantities were used to evaluate the performance of these methods: (1) empirical type I error rate measured as the

proportion of time the test statistic reject the null hypothesis of treatment effect $H_0: \beta_1 = 0$, when the true treatment effect was $\beta_1 = 0$; (2) empirical power was measured as the proportion of the time the test statistic reject the null hypothesis of treatment effect $H_0: \beta_1 = 0$, when the true treatment effect was $\beta_1 = 0.11$; (3) root mean squared error (RMSE) measured as the $\sqrt{E[(\widehat{\beta_1} - \beta_1)^2]}$, where $\widehat{\beta_1}$ and β_1 are the estimated and true value of treatment effect; (4) average width of the 95% confidence intervals was measured as the average of the difference between the upper limit and lower limit across all 1000 replications.

Results

Type I error rate

The results of type I error rate for all the methods and study types are given in Figure 1 for both ICC=0.03 and 0.06. The type I error rates were more than 10% and around 10% for the GEE and meta-regression methods, when the number of clusters was 6. This rate was around 7% for the GEE when the number of clusters was 24. On the other hand, mixed-effect and CL linear regression yielded approximately 5%, nominal level, type I error rates. All the methods followed the similar pattern as ICC=0.03 when the ICC was 0.06 (Figure 1).

Overall, GEE and meta-regression yielded liberal type I error rates for study with small number of clusters. Both of these methods yielded approximately 5% type I error

rates as the number of clusters increased. Mixed-effect and CL linear regression yielded 5% type I error rates almost in all cases.

Empirical power

The results of empirical power for all methods are provided in Figure 2 for both ICC=0.03 and 0.06. GEE and meta-regression had more power compared to mixed-effect and CL linear regression methods when the number of clusters was 6. CL linear regression and mixed-effect had power around 10%. For the number of clusters 24, mixed-effect, CL linear regression and meta-regression had almost similar power, while GEE had slightly more power compared to these methods. Meta-regression had the lowest power compared to other methods for the number of clusters 34 and 68. GEE, mixed-effect and CL linear regression had almost similar power when the number of clusters was 68.

The power for all methods followed almost the similar pattern as ICC=0.03 for the ICC=0.06. However, the power for all methods were slightly lower when the ICC was 0.06 compared to ICC=0.03. No method yielded 80% power when the ICC was 0,06 (Figure 2).

Overall, the power for all methods increased as the number of clusters and cluster size increased. GEE had the highest power compared to other methods for small number of clusters (6, 24). As the number of clusters increased CL linear regression and mixed-effect had similar power as GEE, while meta-regression yielded the lowest power.

Root Mean Squared Error (RMSE)

Overall, the average RMSEs were decreased as the number of clusters and cluster sizes increased. All methods had almost similar RMSEs, except meta-regression, for each combination of cluster size and number of clusters (Figure 3). Meta-regression had slightly higher average RMSEs for study with small number of clusters.

The average RMSEs follows the similar pattern of ICC=0.03 for the ICC=0.06 for all methods. However, RMSEs were slightly higher for each combination of cluster size and number of clusters for all methods (Figure 3).

Width of 95% confidence intervals

The results of the average widths of the 95% confidence intervals for ICC=0.03 and 0.06 for all methods are given in Figure 4. For the number of clusters 6, GEE had the narrowest widths compared to other methods, while CL linear regression had the widest widths. Meta-regression had the widest widths compared to other methods for the number of clusters 24, 34 and 68. CL linear regression and mixed-effect methods had almost similar widths for the number of clusters 24, while GEE had slightly narrower widths compared to these methods.

Overall, average widths of the 95% confidence intervals decreased as the number of clusters and cluster sizes increased. CL linear regression had the widest widths for small number of clusters. However, as the number of clusters increased meta-regression had the

widest widths compared to other methods. The average widths for all methods follow the same pattern as of ICC=0.03 for ICC=0.06. However, the widths were slightly higher for each combination of cluster size and number of clusters for all study types (Figure 4).

Discussion

In this simulation study, we investigated the performance of several methods to assess the intervention effect from stratified CRTs with a single stratification variable. We have compared 4 different methods: GEE, mixed-effect, CL linear regression and meta-regression methods. It is evident that, the number of clusters and cluster sizes, and ICC had impacted the performance of these methods evaluated through type I error rate, power, root mean square error, and width of 95% confidence interval.

GEE and meta-regression methods yielded liberal type I error rates for the small number of clusters. On the other hand, CL linear regression and mixed-effect methods yielded satisfactory, approximately 5%, type I error rates. Borhan et al [21] investigated the performance of methods for analyzing pretest-posttest binary data from completely randomized CRTs and found that the GEE method yielded liberal type I error rates for small number of clusters. Similarly, Klar and Darlington [22] reported that mixed-effect methods yielded satisfactory 5% type I error rate when analyzed continuous data from completely randomized CRTs. These findings were in line with our findings.

GEE method yielded higher empirical power compared to other methods for small number of clusters, which matched with the findings of Austin 2007 [23] as the author

found GEE method yielded more power compared to other methods when the author investigated the methods for analyzing binary data from CRTs. Researchers noticed that, sandwich covariance method underestimate the standard error in the case of small number of clusters, which inflated the type I error rate and empirical power [24, 32]. Further, the researchers have suggested, it required at least 40 clusters to get reliable estimate using GEE [7]. Overall, meta-regression yielded the lowest power compared to other methods, for number of clusters more than 6, which matched with the findings of Chu et al [14]. Power of all methods decreased as the ICC increased. Chu et al [14] demonstrated that, empirical power for all methods decreased as the ICC increased.

GEE and meta-regression had almost similar performance in terms of type I error rates and power for small number of clusters. Further study to evaluate the performance of GEE with small sample adjustment [25] is warranted. In addition, it requires further methodological investigation, especially small sample adjustment, to improve the performance of meta-regression in the context of stratified cluster randomization trials.

The average RMSEs for all methods were similar except meta-regression and RMSEs were getting lower as the number of clusters and cluster sizes increases for all the methods. Meta-regression yielded slightly higher RMSEs compared to other methods for the small number clusters, which was in line with the findings of Chu et al [14]. Meta-regression method yielded the widest 95% confidence intervals compared to other methods for number of clusters more than 6.

There were several limitations of this study: first, we considered only 1 stratification variable with 2 strata; secondly, we considered only 1:1 allocation of clusters

into interventions groups and fixed cluster size in each stratum; thirdly, this study was limited to two-arm parallel group trial; finally, we adjusted for stratification by using the stratification variables as covariate(s). From our recent systematic survey, we found that most of the stratified cluster randomized trials were two-arm parallel-group trials and adjusted method by using stratification variables as covariate(s) [17].

Researchers have emphasized that, it is important to adjust for stratification, in addition to clustering, to correctly assess the intervention effect from stratified CRTs [15,18,19]. This study shed light on performance of methods – adjusted for stratification and clustering, for analyzing continuous data from stratified CRTs. Future studies are warranted to examine the performance of these methods with varying number of clusters and cluster sizes across strata.

Conclusions

In this simulation study, we investigated the performance of four methods, adjusted for stratification and clustering, for analyzing continuous data from stratified cluster randomized trials with one stratification variable. The performance of all methods improved as the number of clusters and cluster sizes increased, while the performance of all methods worsened as the ICC increased. GEE had more than 10% and meta-regression had approximately 10%, type I error rates for small number of clusters. Meta-regression was the least powerful and least efficient compared to other methods.

List of abbreviation

CRT: Cluster Randomized Trial

GEE: Generalized Estimating Equation

CL: Cluster-level

ICC: Intra-cluster Correlation Coefficient

REML: Restricted Maximum Likelihood

RMSE: Root Mean Square Error

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

This was simulation study and no real data were used.

Funding

There is no funding for this study.

Competing interests

All authors confirm that there are no known conflicts of interest associated with this study and there has been no financial support for this work that could have influenced its outcome.

Authors' Contributions

SB, JM, AP, JA and LT conceived the research question. SB conceptualized, developed and designed the study. LT contributed to the conceptualization and design of the study. SB performed the analyses, prepare the results and draft the manuscript. SB, JM, AP, JA and LT contributed equally to further improve this manuscript.

Authors' Information

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada; [2] Biostatistics Unit, Research Institute of St Joseph's Healthcare, Hamilton, ON, Canada; [3] Department of Family Medicine, McMaster University, Hamilton, ON, Canada; [4] GERAS Centre, Hamilton Health Sciences, Hamilton, ON, Canada; [5] Department of Medicine, McMaster University, Hamilton, ON, Canada ; [6] Departments of Pediatrics and Anesthesia, McMaster University, Hamilton, ON, Canada

Acknowledgement

Not applicable

Reference

1. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold London 2000.
2. Kroeger A, Avila EV, Morison L. Insecticide impregnated curtains to control domestic transmission of cutaneous leishmaniasis in Venezuela: cluster randomized trial. *British Medical Journal* 2002; 325(7368):810–813.
3. Jordhoy M, Fayers P, Saltnes T, Ahlner-Elmqvist M, Jannert M, Kaasa S. A palliative-care intervention and death at home: a cluster randomized trial. *Lancet* 2000; 356(9233):888–893.
4. Mallick R, Kathard H, Borhan ASM, Pillay M, Thabane L. A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students. *Trials* 2018; 19: 664.
5. Bland J. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology* 2004; 4: 21–27.
6. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med* 1997; 16:1753-64.
7. Murray D, Varnell S, Blitstein J. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health* 2004; 94:423–32.
8. Hedeker D, Gibbons R, Flay B. Random-effects regression models for clustered data with an example from smoking prevention research. *J Consult Clin Psychol* 1994; 62:757–65.
9. Zeger L, Liang K-Y, Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;44: 1049–60.
10. Whitehead A. *Meta-analysis of controlled clinical trials*. Edition 1. Chichester: John Wiley and Sons; 2002.
11. Gould A. Multi-centre trial analysis revisited. *Stat Med* 1998, 17(15-16):1779-97, discussion 1799-800.

12. Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Stat Med* 2000, 19(8):1115-1139.
13. Fleiss J. Analysis of data from multiclinic trials. *Control Clin Trials* 1986, 7(4):267-275.
14. Chu R, Thabane L, Ma J, Holbrook A, Pullenayegum E, Devereaux P. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: A simulation study. *BMC Medical Research Methodology* 2011; 11:21.
15. Kahan B, Morris T. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. BMJ 2012;345:e5840.
16. Taljaard et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. BMC Medical Research Methodology 2010, 10:15.
17. Borhan S, Papaioannou A, Ma J, Adachi J, and Thabane L. Analysis and reporting of data from stratified cluster randomized trials. Trials, Under review.
18. Borhan S, Mallick R, Pillay M, Kathard H, Thabane L. Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study. Contemporary Clinical Trial Communications 2019; 15:100405.
19. Borhan S, Kennedy C, Ioannidis G, Papaioannou A, Adachi J, Thabane L. An empirical comparison of methods for analyzing over-dispersed zero-inflated count data from stratified cluster randomized trials. Contemporary Clinical Trial Communications 2020; 17:100539.
20. Kennedy et al. Successful knowledge translation intervention in long-term care: final results from the vitamin D and osteoporosis study (ViDOS) pilot cluster randomized controlled trial. Trials 2015; 16:214.
21. Borhan, ASM, Klar N, Darlington G. Methods for the Analysis of Pretest-Posttest Binary Outcomes from Cluster Randomization Trials (2012). Electronic Thesis and Dissertation Repository. 825.

22. Klar N, Darlington G. Methods for analyzing change in cluster randomized trials. Statistics in Medicine 2004;23:2341-57.
23. Austin P. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. Stat Med 2007, 26(19):3550-3565.
24. Ukoumunne O, Carlin J, Gulliford M. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med 2007, 26(18):3415-3428.
25. Leyrat C, Morgan K, Leurent B, Kahan B. Cluster randomized trials with a small number of clusters: which analyses should be used? International Journal of Epidemiology 2018;47(1):321-31.
26. Brown H, Kempton R. The application of REML in clinical trials. Stat Med 1994, 13(16):1601-1617.
27. McLean R, Sanders W. Approximating the degrees of freedom for SE's in mixed linear models. Proceedings of the Statistical Computing Section of the American Statistical Association. New Orleans, Louisiana; 1988.
28. Zeger L, Liang K-Y, Albert P. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44: 1049–60.
29. Arnold B, Hogan D, Colford Jr J, Hubbard A. Simulation methods to estimate design power: an overview for applied research. BMC Health Research Methodology 2011; 11:94.
30. Moerbeek M and Schie S. How large are the consequences of covariate imbalance in cluster randomized trials: a simulation study with a continuous outcome and a binary covariate at the cluster level. BMC Health Research Methodology 2016; 16:79.
31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2018. URL https://www.R-project.org/.

32. Mancl L, DeRouen T. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001, 57(1):126-134.

Table 1: Selected parameters for simulation study

Variable	**Summary from studies that had continuous outcome as their primary outcome[a]**	**Selected for simulation study**
Number of clusters in the intervention groups	Mean= 50; Median= 34; Q1=24; Q3=68; Min=6; Max=228	Total number of clusters: 6, 24, 34, 68
Number of individuals per cluster	-	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
Number of stratification variables	Mean= 2; Median= 1; Q1=1; Q3=2; Min=1; Max=3	1
Effect size	Mean= 0.36; Median= 0.11; Q1=-0.12; Q3=1.02; Min=-9.12; Max=7.17	0, 0.11
ICC	Mean= 0.09; Median= 0.03; Q1=0.00; Q3=0.06; Min=0.00; Max=0.58	0.03, 0.06

[a]Summary of these variables were calculated form our recently conducted systematic survey [17]

Q1 = 25th percentile; Q3 = 75th percentile; ICC = intra-cluster correlation coefficient

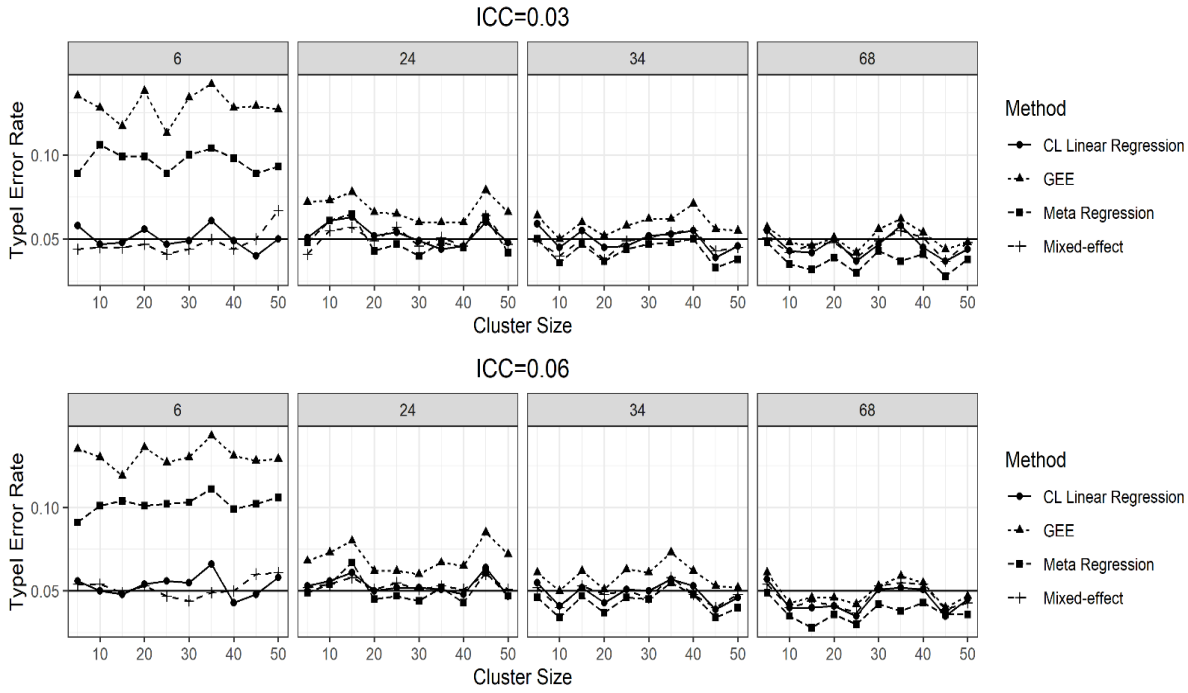

Figure 1: Results of type I error rate for testing null treatment effect, when the true treatment effect was 0, over 1000 simulations for ICC=0.03 & 0.06 and number of clusters 6, 24, 34 and 68

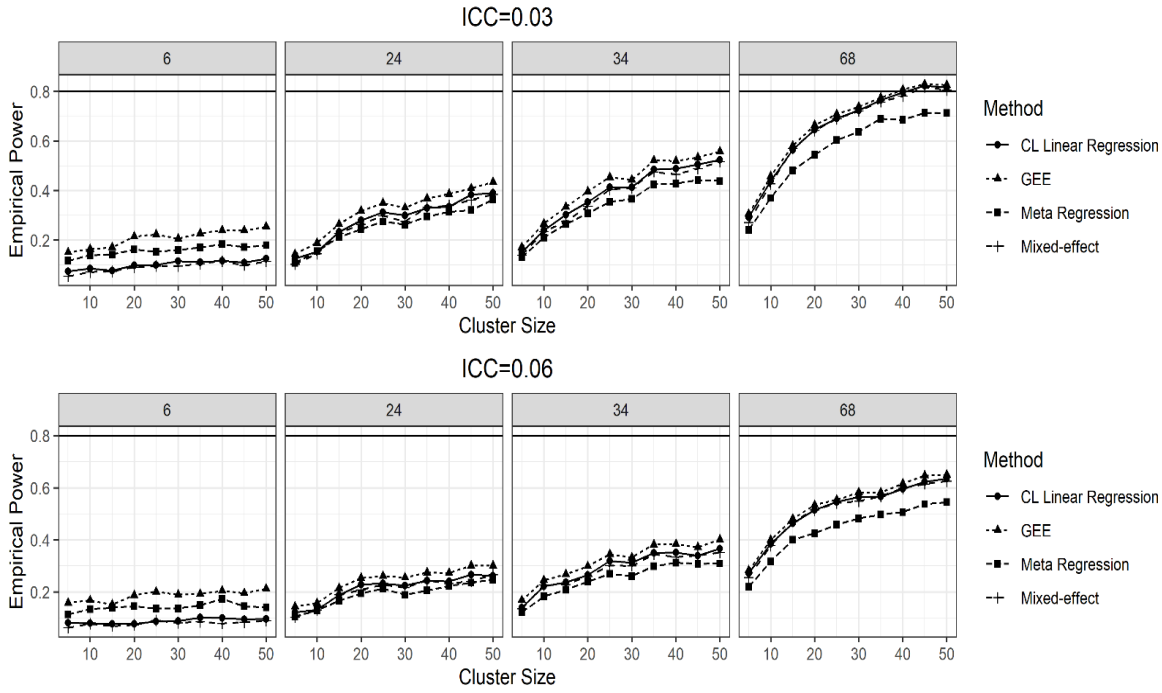

Figure 2: Results of empirical power for testing null treatment effect, while the true treatment effect was 0.11, over 1000 simulations for ICC=0.03&0.06 and number of clusters 6, 24, 34, and 68

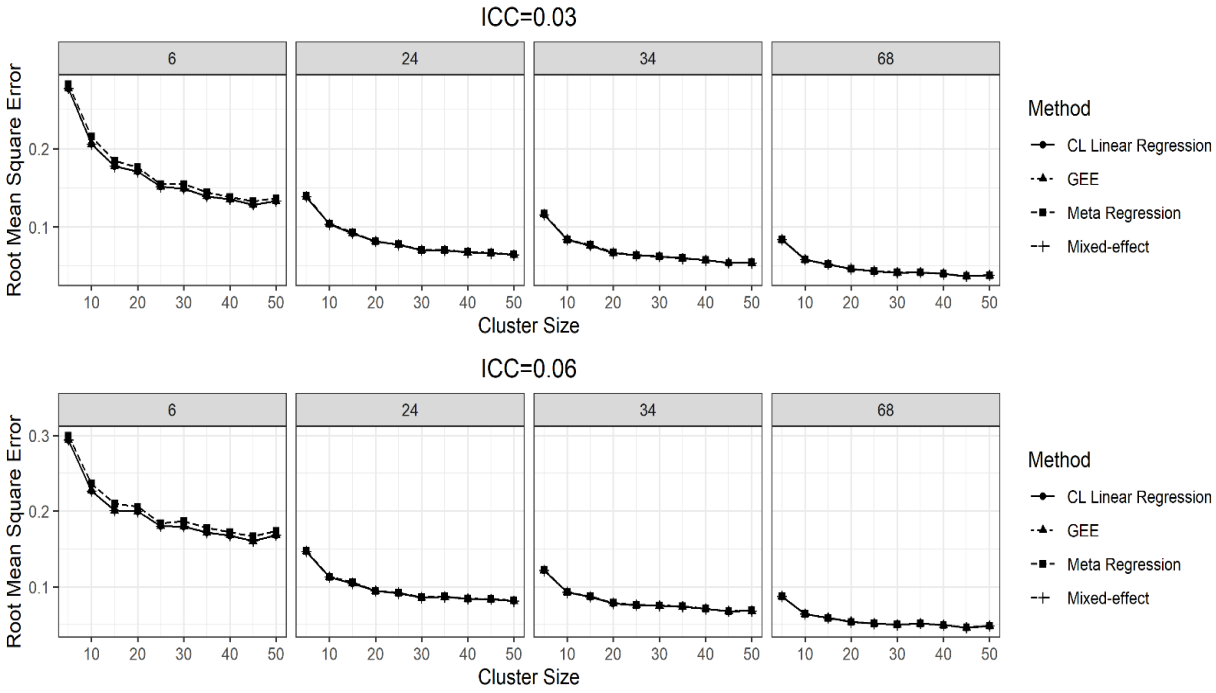

Figure 3: RMSE for testing null treatment effect, while the true treatment effect was 0.11, over 1000 simulations for ICC=0.03 & 0.06, and number of clusters 6, 24, 34, and 68

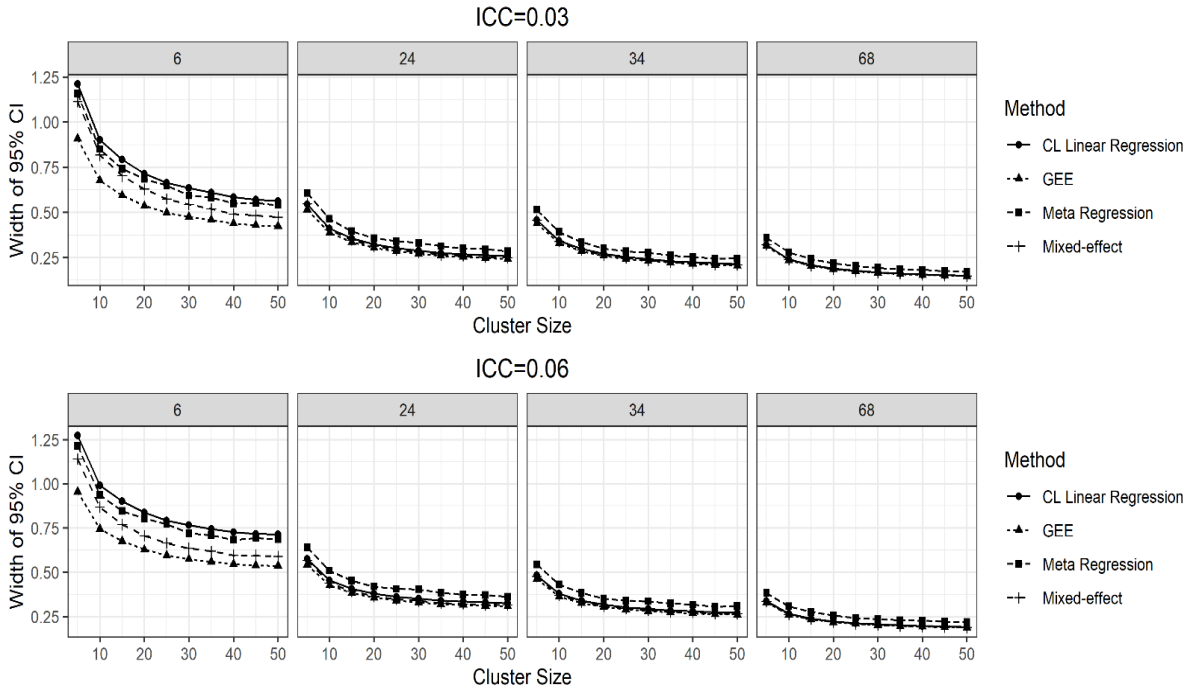

Figure 4: Width of 95% CI for testing null treatment effect, while the true treatment effect was 0.11, over 1000 simulations for ICC=0.03 & 0.06, and number of clusters 6, 24, 34 and 68

Appendix A: Supplemental material

Table A1: Type I error rate for ICC=0.03

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.056	0.135	0.091	0.054
	10	0.050	0.130	0.101	0.054
	15	0.048	0.119	0.104	0.049
	20	0.054	0.136	0.101	0.053
	25	0.056	0.127	0.102	0.047
	30	0.055	0.130	0.103	0.044
	35	0.066	0.143	0.111	0.049
	40	0.043	0.131	0.099	0.050
	45	0.048	0.128	0.102	0.060
	50	0.058	0.129	0.106	0.061
24	5	0.053	0.068	0.049	0.052
	10	0.056	0.073	0.054	0.054
	15	0.061	0.080	0.067	0.058
	20	0.050	0.062	0.045	0.051
	25	0.052	0.062	0.047	0.055
	30	0.052	0.060	0.044	0.051
	35	0.051	0.067	0.052	0.053
	40	0.048	0.065	0.043	0.051
	45	0.064	0.085	0.061	0.060
	50	0.047	0.072	0.047	0.051
34	5	0.055	0.061	0.046	0.052
	10	0.041	0.050	0.034	0.040
	15	0.052	0.062	0.047	0.053
	20	0.043	0.051	0.037	0.048
	25	0.051	0.063	0.046	0.050
	30	0.050	0.061	0.045	0.046
	35	0.057	0.073	0.055	0.058
	40	0.053	0.062	0.048	0.048
	45	0.039	0.053	0.034	0.040
	50	0.046	0.052	0.040	0.048
68	5	0.057	0.061	0.049	0.054
	10	0.040	0.043	0.035	0.040
	15	0.040	0.046	0.028	0.044
	20	0.041	0.046	0.036	0.041

	25	0.035	0.042	0.030	0.037
	30	0.051	0.053	0.042	0.053
	35	0.052	0.059	0.038	0.055
	40	0.051	0.055	0.043	0.054
	45	0.035	0.040	0.036	0.039
	50	0.045	0.047	0.036	0.043

Table A2: Type I error rates for ICC=0.06

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.056	0.135	0.091	0.054
	10	0.050	0.130	0.101	0.054
	15	0.048	0.119	0.104	0.049
	20	0.054	0.136	0.101	0.053
	25	0.056	0.127	0.102	0.047
	30	0.055	0.130	0.103	0.044
	35	0.066	0.143	0.111	0.049
	40	0.043	0.131	0.099	0.050
	45	0.048	0.128	0.102	0.060
	50	0.058	0.129	0.106	0.061
24	5	0.053	0.068	0.049	0.052
	10	0.056	0.073	0.054	0.054
	15	0.061	0.080	0.067	0.058
	20	0.050	0.062	0.045	0.051
	25	0.052	0.062	0.047	0.055
	30	0.052	0.060	0.044	0.051
	35	0.051	0.067	0.052	0.053
	40	0.048	0.065	0.043	0.051
	45	0.064	0.085	0.061	0.060
	50	0.047	0.072	0.047	0.051
34	5	0.055	0.061	0.046	0.052
	10	0.041	0.050	0.034	0.040
	15	0.052	0.062	0.047	0.053
	20	0.043	0.051	0.037	0.048
	25	0.051	0.063	0.046	0.050
	30	0.050	0.061	0.045	0.046
	35	0.057	0.073	0.055	0.058
	40	0.053	0.062	0.048	0.048
	45	0.039	0.053	0.034	0.040
	50	0.046	0.052	0.040	0.048
68	5	0.057	0.061	0.049	0.054
	10	0.040	0.043	0.035	0.040
	15	0.040	0.046	0.028	0.044
	20	0.041	0.046	0.036	0.041
	25	0.035	0.042	0.030	0.037

	30	0.051	0.053	0.042	0.053
	35	0.052	0.059	0.038	0.055
	40	0.051	0.055	0.043	0.054
	45	0.035	0.040	0.036	0.039
	50	0.045	0.047	0.036	0.043

Table A3: Empirical power for ICC=0.03

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.074	0.151	0.117	0.055
	10	0.086	0.164	0.139	0.071
	15	0.077	0.171	0.142	0.074
	20	0.098	0.215	0.163	0.091
	25	0.100	0.223	0.154	0.096
	30	0.116	0.206	0.160	0.096
	35	0.112	0.227	0.171	0.105
	40	0.117	0.240	0.184	0.115
	45	0.110	0.239	0.172	0.099
	50	0.125	0.255	0.180	0.114
24	5	0.125	0.144	0.109	0.103
	10	0.154	0.189	0.155	0.146
	15	0.233	0.265	0.212	0.228
	20	0.280	0.318	0.245	0.265
	25	0.312	0.350	0.275	0.299
	30	0.300	0.331	0.263	0.275
	35	0.330	0.368	0.296	0.332
	40	0.335	0.386	0.314	0.340
	45	0.384	0.409	0.321	0.362
	50	0.390	0.435	0.364	0.385
34	5	0.151	0.171	0.131	0.140
	10	0.240	0.267	0.211	0.235
	15	0.302	0.334	0.264	0.275
	20	0.354	0.396	0.308	0.336
	25	0.413	0.454	0.355	0.404
	30	0.413	0.443	0.366	0.410
	35	0.486	0.522	0.424	0.476
	40	0.488	0.519	0.429	0.466
	45	0.505	0.533	0.443	0.488
	50	0.523	0.558	0.438	0.516
68	5	0.292	0.306	0.242	0.272
	10	0.438	0.458	0.370	0.427
	15	0.565	0.581	0.481	0.569
	20	0.646	0.664	0.544	0.642
	25	0.691	0.708	0.603	0.689

	30	0.724	0.737	0.637	0.720
	35	0.765	0.774	0.689	0.758
	40	0.796	0.807	0.685	0.781
	45	0.822	0.829	0.713	0.817
	50	0.816	0.825	0.712	0.805

Table A4: Empirical power for ICC=0.06

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.082	0.158	0.114	0.063
	10	0.080	0.168	0.134	0.078
	15	0.077	0.152	0.140	0.070
	20	0.077	0.188	0.145	0.074
	25	0.087	0.201	0.137	0.091
	30	0.088	0.189	0.137	0.082
	35	0.102	0.193	0.149	0.085
	40	0.100	0.205	0.173	0.079
	45	0.095	0.195	0.145	0.084
	50	0.097	0.213	0.141	0.090
24	5	0.122	0.144	0.106	0.103
	10	0.131	0.157	0.129	0.132
	15	0.188	0.215	0.167	0.189
	20	0.228	0.254	0.195	0.209
	25	0.232	0.261	0.213	0.227
	30	0.224	0.256	0.189	0.216
	35	0.243	0.274	0.206	0.243
	40	0.241	0.273	0.223	0.233
	45	0.267	0.301	0.236	0.241
	50	0.262	0.301	0.247	0.267
34	5	0.139	0.168	0.122	0.135
	10	0.222	0.245	0.183	0.227
	15	0.237	0.268	0.208	0.229
	20	0.265	0.299	0.240	0.260
	25	0.319	0.345	0.269	0.301
	30	0.312	0.332	0.261	0.300
	35	0.350	0.382	0.299	0.345
	40	0.352	0.384	0.312	0.334
	45	0.339	0.372	0.308	0.340
	50	0.366	0.401	0.309	0.352
68	5	0.272	0.283	0.219	0.256
	10	0.385	0.400	0.317	0.378
	15	0.464	0.481	0.400	0.470
	20	0.515	0.534	0.425	0.514
	25	0.545	0.554	0.458	0.541

	30	0.564	0.582	0.482	0.548
	35	0.566	0.582	0.497	0.567
	40	0.596	0.616	0.506	0.598
	45	0.622	0.647	0.537	0.615
	50	0.634	0.650	0.544	0.625

Table A5: RMSEs for ICC=0.03

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.28	0.28	0.28	0.28
	10	0.21	0.21	0.22	0.21
	15	0.18	0.18	0.18	0.18
	20	0.17	0.17	0.18	0.17
	25	0.15	0.15	0.15	0.15
	30	0.15	0.15	0.15	0.15
	35	0.14	0.14	0.14	0.14
	40	0.14	0.14	0.14	0.14
	45	0.13	0.13	0.13	0.13
	50	0.13	0.13	0.14	0.13
24	5	0.14	0.14	0.14	0.14
	10	0.10	0.10	0.10	0.10
	15	0.09	0.09	0.09	0.09
	20	0.08	0.08	0.08	0.08
	25	0.08	0.08	0.08	0.08
	30	0.07	0.07	0.07	0.07
	35	0.07	0.07	0.07	0.07
	40	0.07	0.07	0.07	0.07
	45	0.07	0.07	0.07	0.07
	50	0.06	0.06	0.07	0.06
34	5	0.12	0.12	0.12	0.12
	10	0.08	0.08	0.08	0.08
	15	0.08	0.08	0.08	0.08
	20	0.07	0.07	0.07	0.07
	25	0.06	0.06	0.06	0.06
	30	0.06	0.06	0.06	0.06
	35	0.06	0.06	0.06	0.06
	40	0.06	0.06	0.06	0.06
	45	0.05	0.05	0.05	0.05
	50	0.05	0.05	0.05	0.05
68	5	0.08	0.08	0.08	0.08
	10	0.06	0.06	0.06	0.06
	15	0.05	0.05	0.05	0.05
	20	0.05	0.05	0.05	0.05
	25	0.04	0.04	0.04	0.04

	30	0.04	0.04	0.04	0.04
	35	0.04	0.04	0.04	0.04
	40	0.04	0.04	0.04	0.04
	45	0.04	0.04	0.04	0.04
	50	0.04	0.04	0.04	0.04

Table A6: RMSEs for ICC=0.06

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	0.29	0.29	0.30	0.29
	10	0.23	0.23	0.24	0.23
	15	0.20	0.20	0.21	0.20
	20	0.20	0.20	0.21	0.20
	25	0.18	0.18	0.18	0.18
	30	0.18	0.18	0.19	0.18
	35	0.17	0.17	0.18	0.17
	40	0.17	0.17	0.17	0.17
	45	0.16	0.16	0.17	0.16
	50	0.17	0.17	0.17	0.17
24	5	0.15	0.15	0.15	0.15
	10	0.11	0.11	0.11	0.11
	15	0.10	0.10	0.11	0.10
	20	0.09	0.09	0.09	0.09
	25	0.09	0.09	0.09	0.09
	30	0.09	0.09	0.09	0.09
	35	0.09	0.09	0.09	0.09
	40	0.08	0.08	0.08	0.08
	45	0.08	0.08	0.08	0.08
	50	0.08	0.08	0.08	0.08
34	5	0.12	0.12	0.12	0.12
	10	0.09	0.09	0.09	0.09
	15	0.09	0.09	0.09	0.09
	20	0.08	0.08	0.08	0.08
	25	0.08	0.08	0.08	0.08
	30	0.08	0.08	0.08	0.08
	35	0.07	0.07	0.07	0.07
	40	0.07	0.07	0.07	0.07
	45	0.07	0.07	0.07	0.07
	50	0.07	0.07	0.07	0.07
68	5	0.09	0.09	0.09	0.09
	10	0.06	0.06	0.06	0.06
	15	0.06	0.06	0.06	0.06
	20	0.05	0.05	0.05	0.05
	25	0.05	0.05	0.05	0.05

	30	0.05	0.05	0.05	0.05
	35	0.05	0.05	0.05	0.05
	40	0.05	0.05	0.05	0.05
	45	0.05	0.05	0.05	0.05
	50	0.05	0.05	0.05	0.05

Table A7: Width of 95% confidence intervals for ICC=0.03

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	1.21	0.91	1.16	1.11
	10	0.90	0.68	0.85	0.82
	15	0.79	0.59	0.74	0.70
	20	0.72	0.54	0.68	0.63
	25	0.66	0.50	0.65	0.57
	30	0.63	0.48	0.59	0.54
	35	0.61	0.46	0.58	0.52
	40	0.58	0.44	0.55	0.49
	45	0.57	0.43	0.55	0.48
	50	0.56	0.42	0.54	0.47
24	5	0.55	0.51	0.61	0.55
	10	0.41	0.39	0.46	0.40
	15	0.36	0.34	0.40	0.35
	20	0.32	0.30	0.36	0.31
	25	0.30	0.29	0.34	0.29
	30	0.29	0.27	0.33	0.28
	35	0.28	0.26	0.31	0.26
	40	0.27	0.25	0.30	0.26
	45	0.26	0.25	0.30	0.25
	50	0.26	0.24	0.29	0.25
34	5	0.46	0.44	0.52	0.46
	10	0.34	0.33	0.39	0.34
	15	0.30	0.29	0.34	0.29
	20	0.27	0.26	0.30	0.26
	25	0.25	0.24	0.28	0.25
	30	0.24	0.23	0.28	0.24
	35	0.23	0.22	0.26	0.22
	40	0.22	0.22	0.25	0.22
	45	0.22	0.21	0.24	0.21
	50	0.21	0.21	0.25	0.21
68	5	0.32	0.31	0.36	0.32
	10	0.24	0.24	0.28	0.24
	15	0.21	0.20	0.24	0.21
	20	0.19	0.19	0.22	0.19
	25	0.18	0.17	0.20	0.18

	30	0.17	0.17	0.19	0.17
	35	0.16	0.16	0.19	0.16
	40	0.16	0.15	0.18	0.16
	45	0.15	0.15	0.18	0.15
	50	0.15	0.15	0.17	0.15

Table A8: Widths of 95% confidence intervals for ICC=0.06

Number of clusters	Number of individuals per cluster	CL Linear Regression	GEE	Meta Regression	Mixed-effect
6	5	1.27	0.95	1.21	1.14
	10	0.99	0.74	0.94	0.87
	15	0.90	0.68	0.85	0.77
	20	0.84	0.63	0.80	0.71
	25	0.79	0.59	0.77	0.66
	30	0.77	0.57	0.72	0.64
	35	0.75	0.56	0.71	0.62
	40	0.73	0.55	0.68	0.60
	45	0.72	0.54	0.69	0.59
	50	0.71	0.54	0.69	0.59
24	5	0.58	0.54	0.64	0.57
	10	0.45	0.43	0.51	0.44
	15	0.41	0.38	0.45	0.39
	20	0.38	0.36	0.42	0.36
	25	0.36	0.34	0.41	0.35
	30	0.35	0.33	0.40	0.34
	35	0.34	0.32	0.38	0.32
	40	0.33	0.32	0.37	0.32
	45	0.33	0.31	0.37	0.32
	50	0.33	0.31	0.36	0.31
34	5	0.48	0.46	0.54	0.48
	10	0.38	0.36	0.43	0.37
	15	0.34	0.33	0.38	0.33
	20	0.32	0.30	0.35	0.31
	25	0.30	0.29	0.34	0.29
	30	0.29	0.28	0.34	0.28
	35	0.28	0.27	0.32	0.28
	40	0.28	0.27	0.32	0.27
	45	0.27	0.26	0.30	0.27
	50	0.27	0.26	0.31	0.26
68	5	0.34	0.33	0.38	0.33
	10	0.27	0.26	0.31	0.26
	15	0.24	0.23	0.28	0.23
	20	0.22	0.22	0.26	0.22
	25	0.21	0.21	0.24	0.21

	30	0.20	0.20	0.23	0.20
	35	0.20	0.20	0.23	0.20
	40	0.20	0.19	0.22	0.19
	45	0.19	0.19	0.22	0.19
	50	0.19	0.19	0.22	0.19

Chapter 6

Discussion and Conclusion

In this thesis, we investigated the several methodological and statistical challenges pertaining to stratified CRTs : (i) surveyed the literature to assess the current practice about reporting and analysis of data from stratified CRTs; (ii) assessed the sensitivity of methods for analyzing continuous data from stratified CRTs; (iii) empirically investigated the sensitivity of methods for analyzing count data from stratified CRTs; (iv) evaluated the performance of methods for analyzing continuous data from stratified CRTs. In this chapter, we summarize the key findings focusing on the research questions this thesis is based on. The implications and limitations of these research works are also highlighted here.

6.1 Addressing the research questions

6.1.1 What is the quality of reporting of stratified CRTs?

In Chapter 2, we conducted a systematic survey [1] to appraise the reporting and analysis of data from stratified CRTs. Overall, 185 stratified CRTs were included for data abstraction. Data were abstracted on several design characteristics including reporting of sample size, randomization, primary method of analysis and reporting of results. All of the included studies did not report the sample size calculation. Only ~60% of the 185 studies reported the ICC or CV, while ~25% studies reported the adjustment for lost to follow-up and <10% reported method used to calculate the sample size. More than 50% of the studies

did not report the method used for randomization, while 23% of the studies did not define all the strata.

6.1.2 Is the intervention effect assessed through proper adjustments – namely, clustering and stratification?

In order to correctly assess the intervention effect from stratified CRTs it is necessary to adjust for both stratification and clustering in the primary analysis. One of the objectives of this thesis was to summarize the evidence on that front. From our systematic survey [1] in Chapter 2, we found that only 38% of the studies adjusted the primary method for both stratification and clustering. Further, only 19% and 31% of the studies included stratification variables in the study flow chart and baseline characteristics table, respectively.

6.1.3 How robust is the methods for analyzing continuous data from stratified CRTs?

In Chapter 3, we conducted an empirical study [2] to examine the sensitivity of methods for analyzing continuous data from stratified CRTs. Five individual- and cluster-level methods including: standard linear regression, cluster-level linear regression, GEE, mixed-effect and meta-regression were compared to assess effect of the intervention classroom curriculum resources (CCR) on improving the peer attitude towards children who stutter using the data from the Mallick et al [3] study. The conclusion from all methods was similar that is, CCR has no effect compared to usual care on improving the peer attitude [2]. The direction of the estimated effect was similar for all methods except for meta-

regression. The magnitudes of the estimated differences were similar among the methods with or without adjustment for stratification.

6.1.4 How robust is the methods for analyzing over-dispersed count data with excess zeros from stratified CRTs?

In Chapter 4, we assessed the sensitivity of methods for analyzing count data from stratified CRTs [4], using the data from the ViD*OS* study [5]. The outcome, number of falls, was over-dispersed with excessive zeros. Eight methods based on Poisson and negative binomial distributions were compared. The overall conclusions from all methods were similar that the KT intervention had no effect on number of falls. However, these methods differ in terms of estimated RRs, and widths of 95% confidence intervals. The estimated RRs were higher and confidence intervals were wider for methods that did not take into account the inflated zeros.

6.1.5 What is the impact of ignoring the adjustment for stratification?

In Chapter 3 and 4 we assessed the impact of not adjusting for stratification in estimating the intervention effect when the outcomes of interest were continuous and count, respectively. The results from these empirical studies demonstrated that, the widths of the 95% confidence intervals were wider and p-values were higher in the absence of adjustment for stratification, even if the methods were adjusted for clustering.

6.1.6 How the varying number of clusters, cluster sizes, ICCs, and effect sizes impact the performance of methods for analyzing continuous data from stratified CRTs?

In Chapter 6, we conducted a simulation study to evaluate the performance of several methods for analyzing continuous data from stratified CRTs [6]. We considered a stratified CRT with one binary stratification variable. Data were generated for varying number of clusters, cluster sizes, ICCs, and effect sizes. Mixed-effects method, GEE, meta-regression and cluster-level linear regression were compared in terms of type I error rate, statistical power, root mean square error and widths of 95% confidence intervals.

The performance of all methods improved as the number of clusters and cluster sizes increased. GEE and meta-regression yielded ~10% type I error rate for small number of clusters [6]. Rejection rate for all methods for testing the null intervention effect increased as the effect size increased. All methods but meta-regression had approximately similar accuracy and precision. Performance of all methods worsened as the ICC increases. Meta-regression was the least efficient and least powerful method compared to GEE, mixed-effects method and cluster-level linear regression for assessing intervention effect from stratified CRTs [6].

6.2 Implications for research and researchers

From our systematic survey [1] it is evident, reporting of stratified CRTs require significant improvement. It is important to report the sample size calculation and the necessary parameters used to calculate the sample size including ICC, one- or two-sided test, and method used to calculate the sample size. Further, it is vital to define all the strata

and include the stratification variables in the study flow-chart or baseline characteristics table to highlight the stratification nature of the study. The primary method to examine the intervention effect should adjust for both stratification and clustering as failure to such adjustment will lead to erroneous conclusion [2,4,7]. Stratum-specific reporting of number of clusters, cluster sizes and effect size (if possible) would be beneficial for the readers.

The results from the empirical studies demonstrated the robustness of several methods for analyzing continuous and zero-inflated over-dispersed count data from stratified CRTs. The overall conclusion from all the methods were similar. But these methods differ in terms of effect size and precision. Also, these empirical studies confirmed that the failure to adjust for stratification yield wider confidence intervals and larger p-values, which in turn reinforce the need for such adjustment for assessing the effect of intervention.

We investigated the performance of several methods for analyzing the continuous data from stratified CRTs under varying number of clusters, cluster sizes, ICCs, and effect sizes, and found that meta-regression was the least powerful and least efficient method compared to GEE, mixed-model, and cluster-level linear regression methods. However, people have to be cautious about using cluster-level linear regression since it is based on cluster-level mean.

6.3 Major limitations and future work

There are several major limitations of this thesis. First, the systematic survey [1] we conducted was based on MEDLINE only. Also, one reviewer was involved in the study

selection and data abstraction. A large-scale systematic review including all the available databases would provide more complete picture about the reporting and analysis of data from stratified CRTs.

Second, our empirical studies on continuous [2] and count [4] data were based on very small and pilot stratified CRTs, respectively. Empirical study based on large stratified CRT can be further investigated to assess the robustness of methods.

Finally, in the simulation study [6] we only considered a stratified design with one stratification variable with two strata. This can be further extended involving more than one stratification variable and more than two strata. Also, we did not consider any adjustment for small number of clusters in the GEE method [8] in our simulation study, which can be further investigated. This type of simulation study can be further extended to binary or count data.

6.4 Conclusion

In this thesis, we conducted a systematic survey to summarize the evidence about the reporting and analysis of data from stratified CRTs. We identified the significant deficiency in reporting and analysis of data and highlighted some areas that need to be included in the reporting. We examined the robustness of methods for analyzing continuous and count data from stratified CRTs. Finally, we evaluated the performance of methods for analyzing continuous data from stratified CRTs. We believe, these research works will guide the researchers to correctly assess the effect of intervention as well as improve the reporting from stratified CRTs.

Reference

1. Borhan S, Papaioannou A, Ma J, Adachi J, and Thabane L. Analysis and reporting of data from stratified cluster randomized trials. *Trials*, Under review.
2. Borhan S, Mallick R, Pillay M, Kathard H, Thabane L. Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study. *Contemporary Clinical Trial Communications* 2019; 15:100405.
3. Mallick R, Kathard H, Borhan ASM, Pillay M, Thabane L. A Cluster randomised trial of a classroom communication resource program to change peer attitudes towards children who stutter among grade 7 students. *Trials* 2018; 19: 664.
4. Borhan S, Kennedy C, Ioannidis G, Papaioannou A, Adachi J, Thabane L. An empirical comparison of methods for analyzing over-dispersed zero-inflated count data from stratified cluster randomized trials. *Contemporary Clinical Trial Communications* 2020; 17:100539.
5. Kennedy et al. Successful knowledge translation intervention in long-term care: final results from the vitamin D and osteoporosis study (ViDOS) pilot cluster randomized controlled trial. *Trials* 2015; 16:214.
6. Borhan S, Ma J, Papaioannou A, Adachi J, Thabane L. Performance of methods for analyzing continuous data from stratified cluster randomized trials - a simulation study. *BMC Medical Research Methodology* 2020. Under Review.
7. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or

minimisation. *Stat Med* 2012;31:328-40.

8. Leyrat C, Morgan K, Leurent B, Kahan B. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology* 2018;47(1):321-31.