MONASH University

Improving the design, reporting and statistical methods used in cluster randomised crossover trials

Sarah J. Arnup
BSc(Hons), MPhil, MBiostat

A thesis submitted for the degree of
Doctor of Philosophy at Monash University in 2019
Department of Epidemiology and Preventive Medicine
School of Public Health and Preventive Medicine

Copyright Notice

Contents

Abstract v

Declaration vii

List of research outputs xv

Acknowledgments xix

List of abbreviations xxi

1 Introduction 1

1.1 Introduction 1
1.2 An overview of the parallel-group cluster randomised trial design 2
1.3 An overview of the individually randomised crossover trial design 4
1.4 An overview of the cluster randomised crossover design 4
1.5 Evaluation of universal interventions in intensive care research 6
1.6 A review of methods for analysing CRXO trials 7
1.7 A review of methods for sample size calculation for CRXO trials 8
1.8 Guidance for the design, statistical methods and reporting of CRXO trials . 10
1.8.1 Design and statistical methodology for CRXO trials 10
1.8.2 Reporting of CRXO trials 11
1.9 Rationale and outline of thesis 11
1.9.1 Aims and objectives 11
1.9.2 Outline of thesis 13

2 The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review 15

3 Appropriate methods were infrequently used in cluster randomised crossover trials 23

4 The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality 37

5 Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial 57

6 Sample size calculations for cluster randomised crossover trials in Australian and New Zealand intensive care research 75

7 Summary and conclusions 85

7.1 Introduction 85

7.2 Summary of thesis chapters 86

7.2.1 Chapters 2 and 3: A protocol and systematic review of CRXO trials 86

7.2.2 Chapter 4: A proposal for reporting items and an assessment of reporting quality for CRXO trials 87

7.2.3 Chapter 5: A graphical illustration of the components of variation from the CRXO design and a sample size tutorial 87

7.2.4 Chapter 6: Sample size calculations for CRXO trials in Australian and New Zealand intensive care research 88

7.3 Overall discussion 89

7.4 Future research 91

References 95

Appendix A Additional file 1 accompanying Chapter 2 - Systematic review data extraction form 101

Appendix B Supplementary data accompanying Chapter 3 - Supplementary tables S1 to S4, Systematic review data extraction form, Reference list of included trials 115

Appendix C Additional files accompanying Chapter 4 - Systematic review search strategies, Supplementary tables S1 and S2 153

Appendix D Blog article accompanying Chapter 4 - A call for cluster randomized cross-over trial reporting guidelines 161

Appendix E Additional files accompanying Chapter 5 - Sample size Stata do files for continous and binary outcomes 165

Appendix F Supplementary material accompanying Chapter 6 - Appendices I to IV, Stata ado file 171

Appendix G Sample size formulae for the risk ratio in stratified cluster randomised crossover trials 183

G.1 Model for the probability of an event in a stratified cluster randomised crossover trial 183

G.2 Estimating the intervention effect 184

G.2.1 Stratum-specific estimator of the intervention effect 184
G.2.2 Overall pooled estimator of the intervention effect 184
G.2.3 Variance of the intervention effect estimator $\hat{\beta}_s$ 185
G.3 Sample size formulae 186
G.3.1 Stratified sample size formula for the risk ratio in cluster randomised crossover trials 187
G.3.2 Unstratified sample size formula for the risk ratio in cluster randomised crossover trials 187
G.3.3 Unstratified sample size formula for risk ratio in parallel-group cluster randomised trials 187

Abstract

Introduction

Randomised controlled trials provide the most reliable estimates of the effects of health care interventions. The cluster randomised crossover (CRXO) design is a form of a cluster randomised trial where clusters are randomised to a sequence of interventions. The design has the potential to allow health care interventions to be evaluated using a randomised design, where cluster randomisation is required, but a parallel-group cluster randomised design is not feasible. The design, conduct, reporting, and use of statistical methodology for CRXO trials has not been evaluated.

Aim

The aims of this thesis are to: assess the design, reporting quality, and statistical methods used in CRXO trials that evaluate clinical and public health interventions; provide guidance to health researchers on the design and reporting of CRXO trials; increase health researchers' understanding of the methodological requirements of the design; and to extend the formulae for sample size calculation to stratified CRXO trials.

Methods

Several methods were used in this thesis. A systematic review was undertaken to characterise how the CRXO design is used in practice. Modification of existing reporting guidelines and identification of areas where reporting items may need to be developed was used to i) assess the reporting quality of published CRXO trials, and ii) propose a set of reporting items for CRXO trials. Graphical illustrations were used to show the intuition behind the key parameters in the CRXO design. Sample size considerations for potential CRXO trials in the intensive care setting were provided to illustrate the use of sample size formulae, and their dependency on the required correlation parameters. Formulae were derived to extend existing sample size methodology to stratified CRXO trials to assess the potential benefits of such stratification.

Results

The systematic review identified 91 trials, including 139 endpoint analyses. Of these, potentially appropriate methodology was used in nine sample size calculations, four of 127 individual-level analyses and 10 of 12 cluster-level analyses. Incomplete reporting of the design aspects unique to the CRXO design was common. Graphical illustrations showed how the parameters required by the sample size formula arise from the design. Stratifying the sample size formula led to a reduction in the required number of clusters to detect a constant risk ratio, when compared to the unstratified sample size formula. Example sample size calculations were sensitive to small changes in the required parameters.

Discussion

The key findings from this thesis are that the methods used for the sample size calculation and analysis of CRXO trials rarely accounted appropriately for the CRXO design, and that reporting of the trial aspects unique to CRXO trials was frequently incomplete. The use of inappropriate methodology can lead to incorrect conclusions regarding the effectiveness of the intervention. Possible explanations for the use of inappropriate methodology include: limited understanding amongst health researchers of the need for specialist methodology, lack of recognition of existing methodology, difficulty in specifying the parameters required by the sample size formula, and lack of availability of appropriate methodology. A CONSORT extension for CRXO trials is needed to improve reporting quality.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes five original papers published in peer reviewed journals. The core theme of the thesis is to facilitate improvement in the design, reporting and statistical methods used in cluster randomised crossover trials. The ideas, development and writing up of the manuscripts in the thesis were the principal responsibility of myself, the candidate, working within the Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine under the supervision of Professor Andrew Forbes, Associate Professor Joanne McKenzie, and Associate Professor David Pilcher.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. In the case of Chapters 2, 3, 4, 5 and 6 my contribution to the work is as follows:

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names, Nature and % of co-authors' contribution**	**Co-authors, Monash student?**
2	The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review	Published in peer reviewed journal: *Systematic Reviews*	80%. Conceived the study in collaboration with Prof Andrew Forbes and Ass Prof Joanne McKenzie. Led the design of the study and drafted the manuscript with contributions from Ass Prof McKenzie and Prof Forbes. Made subsequent revisions to the manuscript based on critical review from all co-authors.	**Andrew Forbes 4%** Contributed to the conception and design of the study and drafting of the manuscript, and provided critical review of the manuscript. **Brennan Kahan 4%** Commented on the design of the study and provided critical review of the manuscript. **Katy Morgan 4%** Commented on the design of the study and provided critical review of the manuscript. **Steve McDonald 4%** Contributed to the design of the search strategy. **Joanne McKenzie 4%** Contributed to the conception and design of the study and drafting of the manuscript, and provided critical review of the manuscript.	No for all

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names, Nature and % of co-authors' contribution**	**Co-authors, Monash student?**
3	Appropriate methods were infrequently used in cluster-randomized crossover trials	Published in peer reviewed journal: *Journal of Clinical Epidemiology*	80%. Conceived the study in collaboration with Prof Andrew Forbes and Ass Prof Joanne McKenzie. Led the conduct of the study, including performing searching, screening, data extraction and statistical analyses. Led the design of the study and drafted the manuscript with contributions from Ass Prof McKenzie and Prof Forbes. Made subsequent revisions to the manuscript based on critical review from all co-authors.	**Andrew Forbes 5%** Contributed to the conception and design of the study and drafting of the manuscript, screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript. **Brennan Kahan 5%** Screened studies for inclusion in the review, extracted data and provided critical review of the manuscript. **Katy Morgan 5%** Screened studies for inclusion in the review, extracted data and provided critical review of the manuscript. **Joanne McKenzie 5%** Contributed to the conception and design of the study and drafting of the manuscript, screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript.	No for all

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names, Nature and % of co-authors' contribution**	**Co-authors, Monash student?**
4	The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality	Published in peer reviewed journal: *Trials*	80%. Conceived the study in collaboration with Prof Andrew Forbes and Ass Prof Joanne McKenzie. Led the conduct of the study, including performing searching, screening, data extraction, statistical analyses, development of the proposed reporting items and flow diagram. Led the design of the study and drafted the manuscript with contributions from Ass Prof McKenzie and Prof Forbes. Made subsequent revisions to the manuscript based on critical review from all co-authors.	**Andrew Forbes 5%** Contributed to the conception and design of the study and drafting of the manuscript, screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript. **Brennan Kahan 5%** Screened studies for inclusion in the review, extracted data and provided critical review of the manuscript. **Katy Morgan 5%** Screened studies for inclusion in the review, extracted data and provided critical review of the manuscript. **Joanne McKenzie 5%** Contributed to the conception and design of the study and drafting of the manuscript, screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript.	No to all

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names, Nature and % of co-authors' contribution**	**Co-authors, Monash student?**
5	Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial	Published in peer reviewed journal: *Trials*	80%. Conceived the tutorial in collaboration with Prof Andrew Forbes and Ass Prof Joanne McKenzie. Led the design of the tutorial and draft the manuscript in collaboration with Prof Forbes, Ass Prof McKenzie and Prof Hemming. Made subsequent revisions to the manuscript based on critical review from all co-authors.	**Joanne McKenzie 7.5%** Contributed to the conception and design of the study and drafting of the manuscript, and provided critical review of the manuscript. **Karla Hemming 2.5%** Contributed to the design of the study and drafting of the manuscript, and provided critical review of the manuscript. **David Pilcher 2.5%** Contributed to the development of the example sample size calculation and provided critical review of the manuscript. **Andrew Forbes 7.5%** Contributed to the conception and design of the study and drafting of the manuscript, and provided critical review of the manuscript.	No to all

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names, Nature and % of co-authors' contribution**	**Co-authors, Monash student?**
6	Sample size calculations for cluster randomised crossover trials: explanation and examples in the Australian and New Zealand intensive care setting	Published in peer reviewed journal: *Critical Care and Resuscitation*	80%. Conceived the tutorial in collaboration with Prof Andrew Forbes, Ass Prof Joanne McKenzie and Ass Prof David Pilcher. Led the design of the tutorial and draft the manuscript in collaboration with Prof Forbes, Ass Prof McKenzie, Ass Prof Pilcher and Prof Bellomo. Made subsequent revisions to the manuscript based on critical review from all co-authors.	**Joanne McKenzie 7.5%** Contributed to the conception and design of the tutorial and drafting of the manuscript, and provided critical review of the manuscript. **David Pilcher 2.5%** Contributed to the development of the example sample size calculations and provided critical review of the manuscript. **Rinaldo Bellomo 2.5%** Contributed to the development of the example sample size calculations and provided critical review of the manuscript. **Andrew Forbes 7.5%** Contributed to the conception and design of the tutorial and drafting of the manuscript, and provided critical review of the manuscript.	No to all

I have renumbered sections of published papers in order to generate a consistent presentation within the thesis.

Candidate's Signature		**Date**

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor's Signature		**Date**

List of research outputs

Listed below are the candidate's first-author and co-authored publications and conference proceedings that are relevant to the period of candidature.

Publications relevant to the thesis

1. **Arnup SJ**, Forbes AB, Kahan BC, Morgan KE, McDonald S, McKenzie JE: **The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review.** *Syst Rev* 2014, **3**:86.

2. **Arnup SJ**, Forbes AB, Kahan BC, Morgan KE, McKenzie JE: **Appropriate statistical methods were infrequently used in cluster-randomized crossover trials.** *J Clin Epidemiol* 2016, **74**:40-50.

3. **Arnup SJ**, Forbes AB, Kahan BC, Morgan KE, McKenzie JE: **The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality.** *Trials* 2016, **17**(1):575.

4. **Arnup SJ**, McKenzie JE, Hemming K, Pilcher D, Forbes AB: **Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial.** *Trials* 2017, **18**(1):381.

5. **Arnup SJ**, McKenzie JE, Pilcher D, Bellomo R, Forbes AB: **Sample size calculations for cluster randomised crossover trials: explanation and examples in the Australian and New Zealand intensive care setting.** *Crit Care Resuscitation* 2018, **20**(2):117-123.

Blog publications relevant to thesis

1. **Arnup SJ**, McKenzie JE: **A call for cluster randomized cross-over trial reporting guidelines.** *BMC On Medicine* Available at: http://blogs.biomedcentral.com/onmedicine/2016/12/07/call-cluster-randomised-cross-trial-reporting-guidelines/

Additional peer-reviewed publications

1. Churilov L, **Arnup S**, Johns H, Leung T, Roberts S, Campbell BC, Davis SM, Donnan GA : **An improved method for simple, assumption-free ordinal analysis**

of the modified Rankin Scale using generalized odds ratios. *Int J Stroke* 2014, **9**(8):999-1005.

2. Lean WL, **Arnup S**, Danchin M, Steer AC: **Rapid diagnostic tests for group A streptococcal pharyngitis: a meta-analysis.** *Pediatrics* 2014, **134**(4):771-81.

3. Davidson AJ, Morton NS, **Arnup SJ**, de Graaff JC, Disma N, Withington DE, Frawley G, Hunt RW, Hardy P, Khotcholava M, von Ungern Sternberg BS, Wilton N, Tuo P, Salvo I, Ormond G, Stargatt R, Locatelli BG, McCann ME, General Anesthesia compared to Spinal anesthesia Consortium: **Apnea after awake regional and general anesthesia in infants: the General Anesthesia compared to Spinal anesthesia study–comparing apnea and neurodevelopmental outcomes, a randomized controlled trial.** *Anesthesiology* 2015, **123**(1):38-54.

4. Frawley G, Bell G, Disma N, Withington DE, de Graaff JC, Morton NS, McCann ME, **Arnup SJ**, Bagshaw O, Wolfler A, Bellinger D, Davidson AJ, General Anesthesia compared to Spinal anesthesia Consortium: **Predictors of failure of awake regional anesthesia for neonatal hernia repair: data from the General Anesthesia compared to Spinal anesthesia study–comparing apnea and neurodevelopmental outcomes.** *Anesthesiology* 2015, **123**(1):55-65.

5. McNab S, Duke T, South M, Babl FE, Lee KJ, **Arnup SJ**, Young S, Turner H, Davidson A: **140 mmol/L of sodium versus 77 mmol/L of sodium in maintenance intravenous fluid therapy for children in hospital (PIMS): a randomised controlled double-blind trial.** *Lancet* 2015, **385**(9974):1190-7.

6. McNab S, Babl FE, Lee KJ, **Arnup SJ**, Davidson A: **Isotonic fluid for intravenous hydration maintenance in children - Authors' reply.** *Lancet* 2015, **386**(9989):136.

7. Myers J, Gibbons K, **Arnup S**, Volders E, Naughton G: **Early childhood nutrition, active outdoor play and sources of information for families living in highly socially disadvantaged locations.** *J Paediatr Child Health* 2015, **51**(3) 287-93.

8. Davidson AJ, Disma N, de Graaff JC, Withington DE, Dorris L, Bell G, Stargatt R, Bellinger DC, Schuster T, **Arnup SJ**, Hardy P, Hunt RW, Takagi MJ, Giribaldi G, Hartmann PL, Salvo I, Morton NS, von Ungern Sternberg BS, Locatelli BG, Wilton N, Lynn A, Thomas JJ, Polaner D, Bagshaw O, Szmuk P, Absalom AR, Frawley G, Berde C, Ormond GD, Marmor J, McCann ME, GÀŠ˙consortium: **Neurodevelopmental outcome at 2 years of age after general anaesthesia and awake-regional anaesthesia in infancy (GAS): an international multicentre, randomised controlled trial.** *Lancet* 2016, **387**(10015):239-50.

9. Shoemark H, Harcourt E, **Arnup SJ**, Hunt RW: **Characterising the ambient sound environment for infants in intensive care wards.** *J Paediatr Child Health* 2016, **52**(4):436-40.

10. McCann ME, Withington DE, **Arnup SJ**, Davidson AJ, Disma N, Frawley G, Morton NS, Bell G, Hunt RW, Bellinger DC, Polaner DM, Leo A, Absalom AR, von Ungern-Sternberg BS, Izzo F, Szmuk P, Young V, Soriano SG, de Graaff JC, GȦŚĊonsortium: **Differences in blood pressure in infants after general anesthesia compared to awake regional anesthesia (GAS study-a prospective randomized trial).** *Anesth Analg* 2017, **125**(3):837-45.

11. Uebergang LK, **Arnup SJ**, Hiscock H, Care E, Quach J: **Sleep problems in the first year of elementary school: the role of sleep hygiene, gender and socioeconomic status.** *Sleep Health* 2017, **3**(3):142-7.

12. Liew PY, Stewart K, Khan D, **Arnup SJ**, Scheinberg A: **Intrathecal baclofen therapy in children: an analysis of individualized goals.** *Dev Med Child Neurol* 2018, **60**(4):367-373.

13. Disma N, Withington D, McCann ME, Hunt RW, **Arnup SJ**, Izzo F, de Graaff JC, Mattioli G, Morton N, Frawley G, Davidson A, Lynn A, Szmuk P, Thomas JJ, Ragg P, Pini Prato A, GAS Consortium: **Surgical practice and outcome in 711 neonates and infants undergoing hernia repair in a large multicenter RCT: Secondary results from the GAS Study.** *J Pediatr Surg* 2018, **53**(9):1643-1650.

14. Reid SM, Meehan EM, **Arnup SJ**, Reddihough DS: **Intellectual disability in cerebral palsy: a population-based retrospective study.** *Dev Med Child Neurol* 2018, **60**(7):687-694.

15. Hiscock H, Quach J, Paton K, Peat R, Gold L, **Arnup S**, Sia KL, Nicolaou E, Wake M: **Impact of a Behavioral Sleep Intervention on New School Entrants' Social Emotional Functioning and Sleep: A Translational Randomized Trial.** *Behav Sleep Med* 2018, **14**:1-15.

16. Gentry KR, **Arnup SJ**, Disma N, Dorris L, de Graaff JC, Hunyady A, Morton NS, Withington DE, McCann ME, Davidson AJ, Lynn AM, GAS Trial Consortium: **Enrollment challenges in multicenter, international studies: The example of the GAS trial.** *Paediatr Anaesth* 2019, **29**(1):51-58.

Conference proceedings

1. **Arnup SJ**, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Cluster randomised crossover trials infrequently use appropriate statistical methods: results from a systematic review. Oral presentation at the 36th International Society for Clinical Biostatistics conference 2015, Utrecht, the Netherlands.

2. **Arnup SJ**, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Cluster randomised crossover trials infrequently use appropriate statistical methods: results from a systematic review. Oral presentation at the Australian Clinical Trials Alliance 2015 International Clinical Trials Symposium, Sydney, Australia.

Acknowledgements

In the time taken to complete this thesis, Ross and I welcomed two beautiful sons into our lives. At times, the research has been all consuming, and other times has been squeezed into the margins of life between two small babies.

I offer my deepest thanks to my supervisors, Andrew Forbes, Jo McKenzie and Dave Pilcher. Andrew and Jo, over the past five years I've shared with you some of the most personal and emotional moments in my life, and I thank you for all the kindness, patience and wisdom that came along with academic supervision. I thank you for some very entertaining supervision meetings, which both kept me sane and drove me crazy. I've learnt a tremendous amount about trial design, methodology and the evaluation of both. I feel I am a far better statistician for it. Dave, I thank you for sharing your expertise on conducting trials in the ICU and for your incredibly fast and detailed response to every question I asked. It has been a pleasure to meet you and I hope the work in this thesis is well received!

Thank you to my numerous co-authors: Brennan Kahan, Katy Morgan, Steve McDonald, Karla Hemming, Rinaldo Bellomo. These papers have provided me with a wonderful opportunity to learn from your expertise.

I am grateful to Monash University for funding from the Monash Graduate Scholarship, and to VicBiostat for providing a top-up scholarship, which made this thesis possible.

I received technical assistance from the following generous people, thank you: to Sue Brennan for advice on using EndNote to manage systematic reviews, to Monica Taljaard for providing an unpublished data extraction form as a starting point for our data extraction form, particularly for the population details and intervention type sections, to Ross Dunn for his assistance in the development of the Access database and Steve McDonald for his assistance in the development of the search strategies.

Thank you to my colleagues at the School of Public Health and Preventive Medicine (SPHPM) and the Clinical Epidemiology and Biostatistics Unit (CEBU), for your ongoing enthusiasm and encouragement, chocolate, and the sharing of ideas and brainstorming. It has been wonderful to have been surrounded by so many capable statisticians and researchers while completing my thesis.

To my dearest friends, who knew just what to do when I didn't. Your acts of kindness made this thesis possible and I treasure those memories fondly. To my parents and siblings, you have all been continuously supportive and understanding, and have repeatedly encouraged me to find ways to keep going.

And I could not have completed this thesis without the love and support of my dearest

Ross, nor without the joy of seeing my children smile and hearing them laugh. What an adventure this has been!

List of abbreviations

ANOVA	ANalysis Of VAriance
ANZICS-APD	Australia and New Zealand Intensive Care Society - Adult Patient Database
CONSORT	CONsolidated Standards Of Reporting Trials
CRCT	parallel-group Cluster Randomised Controlled Trial
CRXO	Cluster Randomised CrossOver
GEE	Generalised Estimating Equation
GLM	Generalised Linear Model
ICC	IntraCluster Correlation
ICU	Intensive Care Unit
IQR	InterQuartile Range
IRCT	Individually Randomised Controlled Trial
LOS	Length Of Stay
MRSA	Methicillin-resistant *Staphylococcus aureus*
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomised Controlled Trial
SD	Standard Deviation
WPC	within-cluster Within-Period Correlation
BPC	within-cluster Between-Period Correlation

Chapter 1

Introduction

1.1 Introduction

Randomised controlled trials (RCTs) provide the most reliable estimates of the effects of health care interventions [1]. Such estimates are used to inform evidence-based clinical practice guidelines. There are different types of designs for randomised trials. The most common form is the parallel-group individually randomised design, where individuals are randomised to interventions [2]. However, there are situations where is it necessary, or preferable, to randomise groups of individuals, such as schools or hospitals, rather than the individual students or patients, to interventions [3][4]. The cluster randomised crossover (CRXO) design is a form of a cluster randomised trial where groups of individuals (clusters) are randomised to a sequence of interventions [5][6].

The aims of this thesis are to assess the design, reporting quality, and statistical methods used in CRXO trials that evaluate clinical and public health interventions; provide guidance to health researchers on the design and reporting of CRXO trials; increase health researchers' understanding of the CRXO design; and extend the formulae for sample size calculation to stratified CRXO trials. The outputs of this thesis provide a set of tools for researchers, which is hoped will facilitate better understanding, design and reporting of CRXO trials.

This chapter sets the context for the research presented in this thesis by providing an overview of: the parallel-group cluster randomised design and individual crossover design, leading to the CRXO design; the use of the CRXO design in intensive care research; the methods available for analysis and sample size calculation of CRXO trials; and the existing guidance for the design, statistical methodology and reporting of CRXO trials. The chapter then concludes with the aims and objectives of the research, and presents an outline of the thesis.

1.2 An overview of the parallel-group cluster randomised trial design

In cluster randomised trial designs, groups of individuals, such as hospitals or schools, are randomly allocated to the intervention groups instead of patients or students, as in an individually randomised trial. These groups form "clusters". There are circumstances where it is preferable, or necessary, to use cluster randomisation in place of individual randomisation [3][7]. Common circumstances are to reduce contamination and because of logistical, feasibility or cost reasons.

Contamination may arise if individuals from the same cluster are randomised to different intervention groups. For example, in a trial of a lifestyle modification intervention that is delivered by general practitioners, patients in the control group may learn of the intervention delivered to the experimental group. Other examples include trials where the intervention involves training health care providers. Health care providers may not be able to unlearn training and therefore may not be able to administer both the control and experimental intervention to patients, as would occur in an individually randomised trial.

Cluster randomisation may also be used for logistical, feasibility, or cost reasons. For example, in a trial of a intravenous fluid in intensive care, the administration of both a control and an experimental fluid on a single ward may increase the possibility of patients receiving a fluid different from their allocation. This risk could be mitigated by providing only one fluid to the ward. Cluster randomisation may reduce the costs of conducting a trial as compared to an individually randomised trial when the intervention requires the cluster to be equipped with expensive equipment or training, because only half of the clusters are allocated to the intervention.

The unique features of cluster randomisation lead to additional methodological and design considerations as compared with designs using individual randomisation. A critical methodological consideration is the correlation between outcomes from individuals within a cluster. The outcomes of individuals within a cluster tend to be more similar to each other than to the outcomes of individuals in different clusters. For example, differences in the case-mix of patients attending different hospitals, may lead to patients in the same hospital having more similar outcomes than patients in other hospitals. This similarity is often quantified by the intracluster correlation (ICC), although other statistics have also been proposed [8].

As a result of the similarity between the outcomes within a cluster, a cluster randomised trial will generally require a larger sample size compared with an individually randomised trial to estimate the intervention effect to the same precision [9]. Failure to account for the clustering during the sample size calculation is likely to lead to an underestimation of the sample size required to detect the desired intervention effect with the specified power. Failure to account for the clustering during analysis is likely to lead to overly precise estimates of the intervention effect and potentially incorrect inferences about the effectiveness of the intervention [3].

As with all trials, an essential design consideration in cluster randomised trials is the

potential for bias in the estimate of the intervention effect. Bias has been classified into five different types: selection bias; performance bias; detection bias; attrition bias; and reporting bias [10]. The potential for selection bias is higher in cluster randomised trials than in individually randomised trials [11][12][13].

Selection bias can arise in a randomised trial when there is a systematic difference in the baseline characteristics of the participants between intervention groups. In an individually randomised trial, assessment of eligibility and baseline characteristics prior to randomisation, adequate random sequence generation and concealment of the sequence removes the potential for selection bias. However, in a cluster randomised trial, the inclusion of individual participants involves the identification and recruitment of both clusters and individuals within clusters. These identification and recruitment processes can introduce selection bias into a cluster randomised trial.

When individual participants within a cluster are identified and recruited before the randomisation of clusters takes place, there is no potential for selection bias. However, if individual participants are identified after the clusters have been randomised, selection bias can arise. When consent is not sought from individual participants, there is minimal risk of selection bias when eligible individuals are identified in a standardised way, and when the researchers responsible for identifying the individuals are blind to the cluster allocation. However, if consent is sought from the individuals recruited into the trial, minimising the risk of selection bias requires that the researchers responsible for recruiting the individuals are blind to the cluster allocation, and the recruited individuals can not become aware of the cluster allocation prior to entry into the study.

Limiting the effects of attrition bias in cluster randomised trials presents additional issues as compared with individually randomised trials. An intention-to-treat analysis, where the outcome data from all participants are analysed in the groups to which the participants were originally randomised, regardless of any subsequent nonadherence or deviation from the trial protocol, including withdrawal, is recommended to minimise the effect of attrition bias [14]. In a cluster randomised trial, an intention-to-treat analysis is difficult to apply, because the total number individuals assigned to each intervention group is not always known. In addition, methods to limit the bias resulting from missing outcome data at both the level of the participant and cluster have not been established [12].

Cluster randomised trials also pose ethical challenges for informed consent that are not present in individually randomised trials [7][15][16][17][18]. In a cluster randomised trial, individual participants may be unable to provide consent for randomisation, participation in an intervention, or data collection; and may also be unable to opt out of receiving the intervention [7][16]. For example, when patients require urgent treatment in a critical care setting. In addition, using cluster randomisation to increase the number of participants included in a trial requires careful ethical consideration [19]. Guidance on the ethical conduct of cluster randomised trials has been published [18], yet work to establish acceptable procedures for consent, and other ethical issues, is ongoing [17].

1.3 An overview of the individually randomised crossover trial design

A commonly used alternative to the parallel-group individually randomised design is the crossover individually randomised design [20]. In crossover designs, individuals are randomised to sequences of interventions, rather than single interventions. The trial is divided into periods of time, and during each period, the individual receives the corresponding intervention in the assigned sequence. For example, a two-sequence two-period design has patients randomised to either A then B, or B then A. While the period of time itself might affect the outcome of each individual in that period, regardless of the assigned intervention, such period effects are removed when the sequence of interventions is balanced between individuals [21].

The advantage of a crossover trial as compared with a parallel-group trial is that interventions can be compared within each individual, rather than between individuals. By performing a "within-subject" comparison of the interventions, the between subject component of variability is removed from the comparison between interventions, and hence the precision of the estimate of the difference between interventions is increased when compared with a between-subject comparison.

The crossover design also presents unique design challenges from the potential for bias that are not present in the parallel-group design. The greatest risk of bias comes from carryover effects, such that the intervention given in one period continues to affect the outcome of individuals in subsequent periods, and the effect on the outcome in subsequent periods differs for each intervention. Examples of situations where an intervention poses a risk of carryover include: the target of the intervention is behaviour change of either the individual level participant (e.g. patient), or the person delivering the intervention (e.g. clinician); a pharmacological intervention that contains ingredients that are still active in subsequent periods; an intervention that leads to a psychological change in the individual that changes perception of the effectiveness of subsequent interventions; or an intervention that leads to an environmental change that alters the effect of subsequent interventions. Appropriate use of wash-out periods, where no intervention is given between active periods, can lessen the risk of carryover if the effect of the intervention is reversible.

A second source of bias arises from treatment-by-period interactions: where the effect of the same intervention on the outcome differs according to the period in which it was received. In a crossover design with two intervention and two periods it is not possible to detect or remove either carryover effects or treatment-by-period interactions. Indeed it is not possible to distinguish between the two sources of bias [21]. Therefore appropriate trial design is essential to minimise the risk of bias.

1.4 An overview of the cluster randomised crossover design

A variation of the parallel-group cluster randomised design, which incorporates the crossover element, is the cluster randomised crossover (CRXO) design. In the CRXO design, each

cluster receives each intervention at least once, which each intervention delivered in a separate period of time, leading to the formation of "cluster-periods". The sequence in which the interventions are delivered to each cluster is randomised to control for potential period effects [5][6]. Within each cluster, each cluster-period may contain a sample of different individuals (a cross-sectional design), a cohort of the same individuals who are followed over time (a cohort design), or a mixture of the same and different individuals [22].

The CRXO design is gaining popularity in settings where cluster randomisation is considered necessary, but the parallel-group cluster randomised design is considered infeasible or problematic. The CRXO design has been proposed in place of the parallel-group cluster randomised design when the number of clusters required by a parallel-group cluster randomised trial is prohibitively large [23][24]; and when the characteristics of the included clusters vary importantly between intervention groups because the number of clusters is small [5][25].

The efficiency of a CRXO trial relative to an individually randomised trial or a parallel-group cluster randomised trial depends on the relationship between the outcomes from individuals within and between each cluster-period. Like individuals within a cluster in a parallel-group cluster randomised trial, individuals within a cluster-period tend to have more similar outcomes than individuals in different clusters. This similarity is typically measured by the within-cluster within-period ICC [5][6][22][24][26]. This tendency for similar outcomes increases the uncertainty in the estimation of the effect of each intervention compared with outcomes that are independent.

If the environment of the cluster remains similar over time, then the outcomes of individuals within the same cluster, but in different cluster-periods, tend to be similar also. This tendency is typically measured by the within-cluster between-period ICC [5][24][26]. The within-cluster between-period ICC is typically less than the within-cluster between-period ICC, because individuals in the same cluster, but in different periods, are likely to be less similar than individuals in the same cluster, in the same period. By comparing the interventions within cluster, the cluster-specific component of variation is removed from the comparison, and the uncertainty of the difference between interventions is decreased when there is a positive within-cluster between-period ICC [5]. Therefore, the crossover element of the design can offset the loss of precision arising from cluster randomisation.

Like the individually randomised crossover design, a key requirement of the two-period CRXO design is that the effect of an intervention given in one period does not carry over into the next period [5][27]. The estimate of the intervention effect may be biased when the intervention given in one period carries over to subsequent periods, and the carryover effect differs according to sequence allocation. In CRXO designs where cluster-periods contain different individuals, the potential for carryover is limited because any carryover can only take place at the cluster level. However, in CRXO designs where the same individuals are followed over time, carryover can also take place at individual participant level, and therefore, the potential for carryover is similar to that in individually randomised crossover designs.

1.5 Evaluation of universal interventions in intensive care research

The research presented in this thesis has been motivated by collaboration with colleagues designing trials to evaluate universal intensive care interventions within the Australian and New Zealand Intensive Care Society Clinical Trials Group. In planning these trials, design, statistical and reporting issues have arisen that needed to be addressed. The research presented in this thesis aims to address some of these issues.

A major challenge in designing RCTs in the intensive care setting is that intervention effect sizes are often small. A review of observed effect sizes in 38 trials to reduce mortality in the intensive care setting found an average effect size of 1.4% [28]. The detection of such small effect sizes, with reasonable levels of statistical power (e.g. >80%), can lead to designs where it is not possible to recruit the required number of participants. Therefore, trials of universal interventions require a design that minimises the number of participants needed to achieve the desired power, and maximises the inclusion of eligible of participants.

An additional challenge to evaluating many universal interventions is that individual randomisation is often infeasible. For example, individual randomisation is not feasible for interventions such as infection control and "bundles of care", which involve multiple practice changes simultaneously. This is because individual randomisation in the intensive care setting poses a high risk for contamination between intervention groups. Individual randomisation is also not feasible for "whole of intensive care processes" such as admission and discharge policies, because varying procedure by individual patient is not practical. For these reasons, it is often more feasible to perform cluster randomised trials in the intensive care setting.

The use of cluster randomisation in intensive care trials allows the intervention to be delivered as if it were a standard operating practice, hence leading to the inclusion of more participants than an individually randomised design, and may increase the efficiency of data collection [23]. However, cluster randomisation can result in designs which require more participants than individual randomisation when there is even a small within-cluster ICC. For example, consider a 12 month trial in which 1200 patients per intensive care unit are eligible. To achieve 80% power to detect an absolute risk reduction from 8.7% to 7.2% (a difference of 1.5%), with a significance level of 0.05, nine intensive care units are required in a parallel-group individually randomised trial (10,800 participants). In a parallel-group cluster randomised trial, if the within-cluster ICC is just 0.01, then the required number of intensive care units increases to 113 (135,600 participants).

In response to the challenges of designing RCTs in the intensive care setting, Bellomo et al. [23] proposed that CRXO trials should be considered more frequently to evaluate low-risk interventions, such as oxygen therapy, ulcer prophylaxis therapy, intravenous fluids and nutrition, in RCTs conducted within the Australian and New Zealand Intensive Care Society Clinical Trials Group.

1.6 A review of methods for analysing CRXO trials

In the analysis of a parallel-group cluster randomised trial, it is commonly recognised that the analysis must account for the correlation within clusters to correctly estimate the uncertainty in the intervention effect (e.g., Eldridge and Kerry 2012 [3], Donner and Klar 2000 [29], Hays and Moulten 2017 [30], Campbell and Walters 2014 [31]). In a CRXO trial, the correlation structure of the outcome data is more complex than in a parallel-group cluster randomised trial. Failure to appropriately account for this additional complexity can lead to incorrect conclusions about the effectiveness of the intervention [32][33].

Similarly to parallel-group cluster randomised trials, CRXO trials can be analysed at the level of the individual or at the level of the cluster. Both methods have been evaluated for normally distributed continuous and binary outcomes. Other outcome types, including count and time-to-event data, have not been explored. Following, a summary of the literature evaluating statistical methods for analysing the CRXO design is provided. These methods have focussed on the two-period two-group cross-sectional CRXO design.

Turner et al. has proposed a simple analysis approach for CRXO trials using a cluster-level analysis for continuous outcomes collected at the individual-level, which can also be applied to other outcome types. In a cluster-level analysis, the available data on each individual from each cluster-period are aggregated into a single measure, and for each cluster, the difference between interventions is calculated. Aggregating the data in each cluster-period accounts for the within-cluster within-period ICC, and comparing the cluster-period summaries within cluster accounts for the within-cluster between-period ICC.

Numerical simulation has been used to evaluate cluster-level methods with summary measures of the difference in period means for continuous outcomes and the risk difference between periods for binary outcomes [6][24][32]. The methods have been evaluated with small to moderate number of clusters (6 to 30 clusters). The method generally performed well under a range of ICC values and cluster-period sizes, except for binary outcomes with low prevalences (6% or less) or large ICCs (greater than 0.05) [24]. Inverse-variance weighting of the summary measure has been shown to be the most efficient cluster-level method [6][24][32], however the less complex size-weighted and unweighted analyses provide comparable power unless the cluster sizes are very unbalanced [24]. Despite their simplicity and good performance, cluster-level analyses may not be preferable to individual-level methods when adjustment for individual-level covariates or testing of an interaction between covariate is desired [6][24].

For individual-level analysis of continuous outcomes, the use of mixed-effects models that include random or mixed effects for cluster and cluster-by-period effects has been shown to generally perform well in numerical simulations [6]. These simulations have explored designs with a small numbers of participants per cluster-period (20) and ICCs ranging from small to large (0.01 to 0.1). Models that treat the cluster effect as fixed and allow for negative within-cluster ICCs perform best, but may not be appropriate if the number of clusters is too large to achieve model convergence, or a negative within-cluster ICC is implausible.

The performance of logistic mixed models for binary outcomes have only recently received attention. Initial results indicate that nominal Type I error rates may not be achieved even with 50 or more clusters, with the performance becoming worse as the difference between the within-cluster within-period ICC and within-cluster between-period ICC increases [32].

Generalised estimating equations (GEEs) are also used to analyse clustered data [34]. In CRXO trials, accounting for both the within-cluster within-period and within-cluster between-period ICCs requires the inclusion of robust standard errors (possibly with adjustments for small sample variance estimation [35]), or extensions that model the patterned correlation structure within clusters across multiple periods explicitly [24][36].

A recently evaluated individual-level method uses GEEs to fit the marginal mean model and allows for both the within-cluster within-period and within-cluster between-period ICCs by employing matrix-adjusted estimating equations with a nested exchangeable correlation matrix, and employs finite-sample adjustment to the sandwich variance estimator [36]. For the limited scenarios considered, tests of the intervention effect have been identified that provide nominal Type I error and power. The scenarios did not include rare events (less than 6%), ICCs smaller than 0.025, or variable cluster sizes, which have been shown to have poor performance in cluster-level analyses for binary outcomes [24].

The scenarios investigated in the above numerical simulation studies were informed by characteristics of CRXO trials the methodologists had encountered in practice. As with all numerical simulation, the conclusions may not hold beyond the evaluated scenarios. Therefore, to assist methodologists in developing and evaluating statistical methodology in future numerical simulation studies, a component of this thesis is to collate all CRXO trials and describe their design characteristics.

The analysis of CRXO trials requires complex methodology. Choosing an appropriate analysis method requires an understanding of the correlation structure induced by the trial design. To assist health researchers in identifying analysis methodology that is appropriate for their trial design and outcome data, increasing researchers' understanding of the CRXO design is a key aim of this thesis.

1.7 A review of methods for sample size calculation for CRXO trials

Rutterford et al. have provided an extensive review of methods for sample size calculation for cluster randomised trials, which also covers the CRXO design [37]. For all methods used to estimate sample size, the specified power will only be obtained if the methods used to estimate the variance of the intervention effect in the sample size calculation match the analysis methodology. As for the analysis methodology, the sample size can be determined for either an individual-level or cluster-level analysis. For designs with equal cluster-period sizes and continuous outcomes, the sample size required for cluster-level and individual-level analyses are equivalent. This equivalence does not hold for designs with unequal

cluster-period sizes within or between clusters, or for binary outcomes.

The simplest CRXO design is the two-period two-group cross-sectional design, with cluster-periods of equal sizes both within and between clusters, and with the number of clusters balanced between randomisation sequences. For this design, sample size formulae or variance inflation factors, relative to a parallel-group individually randomised design, have been derived for both cluster-level analyses and individual-level analyses of binary and normally distributed continuous outcomes [22][26][38]. For cluster-level analyses of binary outcomes, these formulae have been extended to incorporate period effects and unequal cluster-period sizes [24]. For individual-level analyses of continuous and binary outcomes, sample size formulae have been based on methods that use GEEs to fit a marginal model and matrix-adjusted estimating equations to account for the correlation structure [36].

Extensions to the simple CRXO design include the cohort design and designs with more than two periods. An inflation factor for the cohort design with two periods, relative to an individually randomised trial, has been proposed [22], but does not account for variation in expected outcomes between cluster-periods within a cluster. Hooper et al. have provided formulae for both the cross-sectional and cohort design with more than two periods, assuming a constant correlation between the population mean of any two cluster-periods within a cluster [39]. A less restrictive method for designs with more than two periods allows for an exponential decay over time in the correlation between cluster-periods within each cluster [40]. These methods have been developed for continuous outcomes, but may be extended to binary outcomes.

Numerical simulation is an alternative to estimating the sample size from asymptotic formula, and has been recommended when the number of clusters is small [37]. However, simulation may be prohibitive for many health researchers due to the complexity and time involved to set up and conduct. Simulation-based software for assessment of power is available, but does not account for the within-cluster between-period ICC [25]. The theoretical power estimated from sample size formulae has been verified empirically with as few as four to eight clusters, suggesting simulation may be unnecessary [24][36][39]. It is important to note, however, that small number of clusters may also affect the type I error rate [39], and pose additional risks to chance imbalances, generalisability and appropriate options for analysis [41].

In planning trials within the Australian and New Zealand Intensive Care Society Clinical Trials Group, researchers have observed that the characteristics of potential clusters vary according to type of hospital (i.e. tertiary or private hospital) or geographical region (rural or metropolitan). This scenario has not been considered in the published sample size methodology for CRXO trials. Therefore, an aim of this thesis is to extend the existing formulae for sample size calculation to settings where trials are stratified by groups of clusters that vary according to factors that affect the outcome of interest.

1.8 Guidance for the design, statistical methods and reporting of CRXO trials

Appropriate trial design and statistical methods are essential for obtaining an unbiased estimate of an intervention. In addition, comprehensive reporting allows for an assessment of threats to the validity of the trial results, an assessment of the adequacy of the statistical methods, replication of trial methodologies, incorporation of the trial's results in synthesis products such as meta-analyses, and implementation of the evaluated intervention(s). Guidance on trial design and the use of statistical methods may increase researchers' understanding of the design and analysis requirements, thereby reducing the use of inappropriate designs and methodology. To date, there is only limited guidance available on the design and statistical methods for CRXO trials. Further, there are no published reporting guidelines specifically developed for CRXO trials.

The first step in determining whether the CRXO design requires a reporting guideline is to assess the quality of reporting in published trials [42]. Similarly, guidance on the use of the design and statistical methods is only required if these aspects are inadequate in published trials. Therefore, the first aim of the research presented in this thesis is to assess the design, reporting quality, and statistical methods used in CRXO trials. The guidance currently available for CRXO trials, and the subsequent components of this thesis, are now discussed.

1.8.1 Design and statistical methodology for CRXO trials

Parienti and Kuss provide some limited guidance on designing CRXO trials [5]. They provide a brief discussion of potential sources of bias in the estimate of the intervention effect that can arise from the design. Other authors note the need to reduce the risk of carryover during the design stage [6][22][25][26], and how cross-sectional and cohort CRXO designs may differ in the risk of participant dropout [22].

Several methodological publications provide examples of sample size calculation [24][26] and analyses of CRXO trials [5][6][24][32][40]. However, the focus of these articles is primarily about evaluating the performance of the proposed methods, and therefore such guidance may not be readily accessible to health researchers without a methodological background [43]. An online application is available to estimate the design effect, power or variance for a two-group cross-sectional CRXO trial [33].

There is little guidance available to inform the choice of ICCs when determining the required sample size. Approaches for choosing the within-cluster ICC in sample size calculations for parallel-group cluster randomised trials have been proposed [44][45][46][47], and similar considerations are likely to apply when choosing the within-cluster within-period ICC in a CRXO design. However, there is limited guidance for selecting the value of the within-cluster between-period ICC [26][38][48].

To address these limitations, providing guidance for health researchers on the design of CRXO trials, and increasing health researchers' understanding of the CRXO design, forms

a component of this thesis.

1.8.2 Reporting of CRXO trials

In deciding whether a trial design requires a reporting guideline, consideration should first be given to whether any existing guidelines are adequate, potentially with modification [42]. The most applicable guidelines for the CRXO design are the existing CONSORT (Consolidated Standards of Reporting Trials) statements.

The CONSORT statement for parallel-group randomised trials was developed in an attempt to improve the quality of reporting of randomised trials. Extensions have been published for the parallel-group cluster randomised design [49] and the stepped wedge design [19]. Items from these statements are directly applicable to the CRXO design (e.g. "Allowance for clustering" and "Allowance for the number of steps" in the sample size justification) or are easily modifiable (e.g. "Identification as a cluster randomised trial in title") for CRXO trials. However, the CRXO design has distinct characteristics when compared with the parallel-group cluster randomised design and the stepped-wedge design, such as the potential for carryover of the intervention effect to subsequent periods, which may necessitate the development of a specific guideline for the CRXO design. Therefore, an aim of this thesis is to identify aspects of reporting in CRXO trials that can be addressed by existing guidelines, and aspects which are unique to the CRXO design, and subsequently propose possible reporting items and areas where items may need to be developed.

1.9 Rationale and outline of thesis

1.9.1 Aims and objectives

The CRXO design has the potential to allow health care interventions to be evaluated using a randomised design, where cluster randomisation is required, but the parallel-group cluster randomised design is either not appropriate or not feasible. However, without appropriate design, conduct, and statistical methodology, the estimates of the effectiveness of an intervention obtained from a CRXO trial may be misleading. And without complete, transparent and clear reporting, it may not be possible for those using trial reports to assess the validity of the trial results. To date, the design, conduct, and reporting of CRXO trials has not been described.

Therefore, the aim of this thesis is to assess the design, reporting and statistical methods used in CRXO trials that evaluate clinical and public health interventions; and consequentially develop guidance on the design and reporting of CRXO trials for health researchers, increase health researchers' understanding of the design, and extend the existing sample size methodology to CRXO trials conducted in typical intensive care settings. The specific objectives of the thesis are to:

- Assess the design, reporting and statistical methods used in CRXO trials that evaluate clinical and public health interventions, by reviewing the:

  - characteristics of reported trials and evaluated interventions;
  - motivations for using the CRXO design;
  - values of the CRXO design parameters;
  - justification and methodology for the sample size calculation and analyses; and,
  - quality of reporting of the CRXO design aspects.

- Facilitate improvement in the reporting of CRXO trials, by:
  - proposing possible content for a reporting guideline;
  - identifying which aspects of reporting in CRXO trials can be addressed by existing guidelines, potentially with modification;
  - outlining possible options for reporting the aspects of CRXO trials that are unique to the design; and,
  - discussing areas where reporting items may need further development.

- Increase health researchers' understanding of the CRXO design, by illustrating graphically:
  - how individual responses in a CRXO trial can be understood in terms of components of variation;
  - how the similarity between groups of individuals within clusters can be quantified by the components of variation, hence leading to the definitions within-cluster within-period and within-cluster between-period ICCs;
  - the effects of the cluster randomisation and multiple period aspects of the design on the estimate of an intervention; and,
  - when the CRXO design is equivalent to the parallel-group cluster randomised and individually randomised crossover design.

- Provide guidance for determining the required sample size for CRXO trials, by presenting:
  - recommendations for selecting the parameters required by the sample size formulae;
  - estimates of the sample size parameters for intensive care trials; and,
  - worked examples of how to determine the required sample size in an intensive care setting.

- Extend the existing sample size formulae for CRXO trials to stratified trial designs, in order to:
  - accommodate for settings where groups of clusters (strata) vary according to factors that affect the outcome of interest;
  - express the intervention effect as a risk ratio when the outcome is binary; and,
  - assess the benefits of stratification for trials conducted within the Australia and New Zealand intensive care setting.

1.9.2 Outline of thesis

This thesis includes five published manuscripts. **Chapter 2** details a protocol for a systematic review of CRXO trials and aims to examine the characteristics of reported trials, motivations for using the CRXO design, the values of the CRXO design parameters, the justification and methodology for the sample size calculation and analyses, and the quality of reporting of the CRXO design aspects. This protocol has been published in *Systematic Reviews*. **Chapter 3** presents and evaluates the following aspects of the systematic review: characteristics of reported trials, motivations for using the CRXO design; the values of the CRXO design parameters; and the justification and methodology for the sample size calculation and analyses. This work has been published in *Journal of Clinical Epidemiology*. In **Chapter 4**, the reporting quality of CRXO trials is assessed, and items are proposed for a reporting guideline. This work was published in *Trials*. **Chapter 5** provides a graphical illustration of the effects of the cluster randomisation and multiple period aspects of the design, recommendations for selecting parameters for sample size calculation and examples of how to determine the sample size for a CRXO trial. This work was published in *Trials*. A sample size formula for detecting a risk ratio with stratified clusters is presented in **Chapter 6**, along with estimates of the required parameters and examples of its use in the intensive care setting, and was published in *Critical Care and Resuscitation*. **Chapter 7** presents a summary of the findings and suggestions for future research.

Chapter 2

The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review

This chapter presents a protocol for a systematic review of the design, reporting and statistical methods used in CRXO trials that evaluate clinical and public health interventions. This is the first extensive systematic review of such trials. The systematic review was necessary to understand how the CRXO design is used in practice, and to inform the subsequent chapters of this thesis.

The aims of the systematic review are to summarise 1) the characteristics of reported trials, 2) the motivations for using the CRXO design, 3) the values of the CRXO design parameters, 4) the justification and methodology provided for the published sample size calculation and analyses, and 5) the quality of reporting of the design aspects. The results of the systematic review are presented in Chapters 3 and 4.

The protocol presented in this chapter provides a detailed description of the systematic review methods. This includes the search strategies (across multiple databases), the processes used for screening the abstracts and full-text articles, and the methods for data extraction and analysis.

Chapter 2 is presented as a manuscript, which was published in *Systematic Reviews* in August 2014. The pages have been renumbered for the thesis, but the manuscript is otherwise unchanged.

The electronic data extraction form developed for this review, and referred to in the manuscript as "Additional file 1", is appended to this thesis in Appendix A.

Arnup et al. Systematic Reviews 2014, 3:86
http://www.systematicreviewsjournal.com/content/3/1/86

SYSTEMATIC REVIEWS

PROTOCOL Open Access

The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review

Sarah J Arnup[1*], Andrew B Forbes[1], Brennan C Kahan[2], Katy E Morgan[2], Steve McDonald[1] and Joanne E McKenzie[1]

Abstract

Background: The cluster randomized crossover (CRXO) design is gaining popularity in trial settings where individual randomization or parallel group cluster randomization is not feasible or practical. In a CRXO trial, not only are clusters of individuals rather than individuals themselves randomized to trial arms, but also each cluster participates in each arm of the trial at least once in separate periods of time.
We will review publications of clinical trials undertaken in humans that have used the CRXO design. The aim of this systematic review is to summarize, as reported: the motivations for using the CRXO design, the values of the CRXO design parameters, the justification and methodology for the sample size calculations and analyses, and the quality of reporting the CRXO design aspects.

Methods/Design: We will identify reports of CRXO trials by systematically searching MEDLINE, PubMed, Cochrane Methodology Register, EMBASE, and CINAHL Plus. In addition, we will search for methodological articles that describe the CRXO design and conduct citation searches to identify any further CRXO trials. The references of all eligible trials will also be searched.
We will screen the identified abstracts, and retrieve and assess for inclusion the full text for any potentially relevant articles. Data will be extracted from the full text independently by two reviewers. Descriptive summary statistics will be presented for the extracted data.

Discussion: This systematic review will inform both researchers addressing CRXO methodology and trialists considering implementing the design. The results will allow focused methodological research of the CRXO design, provide practical examples for researchers of how CRXO trials have been conducted, including any shortcomings, and highlight areas where reporting and conduct may be improved.

Keywords: Cluster randomized trial, Crossover, Intra-cluster correlation, Sample size, Design, Statistical analysis, Reporting

Background

The most commonly used experimental design to assess the effects of an intervention is the individually randomized parallel two-arm trial [1]. However, randomizing individuals is not always possible, and in many circumstances groups of people, or 'clusters', are instead randomly allocated to the intervention groups. Cluster randomization is commonly used in the following situations: when contamination may occur if individuals in the same cluster were randomized to different intervention groups, the intervention is targeted at the cluster level, or for logistical, feasibility, or ethical reasons [2].

Individuals within a cluster tend to have more similar outcomes than individuals across clusters. For example, due to case-mix differences of patients presenting to different hospitals, patients in the same hospital may have more similar outcomes than patients across different hospitals. As a result, a cluster randomized trial usually requires a larger sample size than an individually randomized trial in order to achieve the same power to detect the same difference between groups. Failure to account for the clustering during analysis can lead to overly precise estimates of the

* Correspondence: sarah.arnup@monash.edu
[1]School of Public Health and Preventive Medicine, Monash University, Level 6, The Alfred Centre, Melbourne, VIC 3004, Australia
Full list of author information is available at the end of the article

intervention effect and hence potentially incorrect inferences about the effectiveness of the intervention [2].

A variation of the parallel group cluster randomized design is the cluster randomized crossover design (CRXO). In the CRXO design each cluster receives each intervention at least once in separate periods of time [3,4]. During each time period the cluster may contain different individuals, the same individuals, or a mixture of both different and same individuals [5].

Analogous to trials where individuals are randomized and a crossover is included in the design to improve efficiency, incorporating a crossover into a parallel group cluster randomized design increases efficiency if the cluster environment remains similar between time periods [5]. The gains in efficiency of a CRXO trial over a parallel group cluster randomized trial depend upon the number of clusters, the size of the clusters, the number of time periods, and the similarity between individuals within the trial. The similarity in the outcomes of individuals within a cluster within a time period is typically measured by the within-cluster within-period intra-cluster correlation coefficient (ICC). The similarity between individuals within the same cluster, both within the same time period and across different time periods, is typically measured by the within-cluster between-period ICC [4,6].

To our knowledge, there have only been limited reviews of the CRXO trial design. These reviews have taken place in the introductory sections of methodological papers with the purpose of illustrating the design and highlighting the need for appropriate methods of analysis [3,4,6,7]. Turner *et al*. [3] reviewed eight trials [8-15] from 1985 to 2003 and noted that the majority of these trials did not allow for the within-cluster within-period and within-cluster between-period correlations in the analysis of outcomes. In the one trial [8] that did allow for these correlations in the analysis by using hierarchical modelling, Turner *et al*. [3] noted that no justification was given for the choice of analysis.

The CRXO design is gaining popularity in settings where cluster randomization is required, but the parallel group cluster randomized design is not practical because it leads to a prohibitively large sample size. However, no systematic review of the use of the CRXO design has been performed to date. Such a review will inform both researchers addressing CRXO methodology and trialists considering implementing this design.

Objectives

The purpose of this systematic review is to establish from CRXO publications: the motivations for using the CRXO design, the values of the CRXO design parameters, the justification and methodology for the sample size calculations and analyses, and the quality of reporting the CRXO design aspects.

Methods/Design

Search methods for identification of studies

We will search for reports of CRXO trials that were conducted in humans and reported in English up until April 2014. One author (SA) will search for articles indexed in MEDLINE, PubMed, Cochrane Methodology Register, EMBASE and CINAHL Plus. (The search strategies for Ovid MEDLINE and the additional databases are in Appendix 1). Ovid was chosen to search MEDLINE because proximity searches, which cannot be performed in PubMed, are an essential component of the search strategy. As PubMed contains some additional publications not found in MEDLINE, a modified but less sensitive search will be performed using PubMed.

To supplement the above searches, SA will search CRXO methodology articles for further references to CRXO trials. A citation search of all identified methodology articles will be performed in Web of Science. SA and JM will identify CRXO methodology articles from PubMed using the following search strategy: ((cluster[tiab] AND cross*over[tiab]) OR cluster-crossover[tiab]) AND (method*[tiab] OR design [tiab] OR calcul*[tiab] OR analy*[tiab]).

Finally the references of all eligible articles will be screened by SA for further CRXO trials. If the title of the article or the text of the referring article suggests a CRXO design was used in the trial then the full text will be screened for eligibility by two reviewers (SA and AF or JM). This process will continue until no further eligible articles are identified.

Inclusion criteria

We will include reports of CRXO trials with the following elements: the trial was undertaken in humans; the allocation of the intervention was to clusters of individuals rather than individuals themselves - the allocation does not have to be at random, since the statistical considerations remain the same irrespective of the method used to allocate clusters to the sequence of interventions; and each cluster received each intervention in a sequence over time (conventional crossover design), or at least some clusters crossed over from one intervention to another (such as two-treatment-four-sequence designs AA, AB, BA, and BB).

Study selection

Titles and abstracts of all articles identified through the electronic searches will be imported into EndNote (EndNote X6, Thomson Reuters, New York, USA) and duplicates removed. Each abstract and title will assessed by one of five reviewers and a further 50% of abstracts will be assessed independently by a second reviewer. Full text articles will be retrieved when both reviewers answer 'yes' or 'unclear' to all selection criteria. The full text will not be retrieved if both reviewers agree that at least one selection criteria was not met. The full text will be retrieved for the

Arnup et al. Systematic Reviews 2014, 3:86
http://www.systematicreviewsjournal.com/content/3/1/86

remaining articles where all selection criteria assessed as 'no' by one reviewer were assessed as 'yes' or 'unclear' by the other reviewer.

Two reviewers will assess the full text articles. Trials will be included in the review if both reviewers agree that all selection criteria are met. Trials will be excluded if both reviewers agree that at least one selection criteria was not met. For the remaining trials, the decision to include the trial in the review will be by consensus between the two reviewers or by referral to a third reviewer.

Data extraction and management

Two reviewers will independently extract data using an electronic data extraction form developed for this review (see Additional file 1). The data extraction form has been piloted by five reviewers in one to three studies each and adjusted accordingly.

We will extract data for each trial on: identification of the design in the title or abstract, justification for using the design, acknowledgement of the underlying assumptions of the design, demographic details (country, setting, unit of clustering, type of intervention, and control), characteristics, methods used in the trial (recruitment, randomization, allocation, and blinding), reporting of baseline characteristics of the trial design, and statistical analysis (methods to estimate intervention effects and adjustment for covariates). The extracted design characteristics will include: number of clusters, number of periods, number of cluster-periods (clusters × periods), number of individuals in the trial, number of interventions and the allocation of interventions to cluster periods, the variability of the number of individuals between cluster-periods, the reported measure of similarity between the outcomes of individuals within a cluster within a given period, and the reported measure of similarity between outcomes of individuals within a cluster between different periods.

The extracted data will include verbatim free text and categorization of the text into pre-specified options where possible. Any free text that does not fall into the pre-specified options will be categorized through discussion between reviewers. If data are not reported in the article or are incomplete,'not stated' will be recorded on the data extraction form. Trialists will not be contacted, since we intend to examine trialists' reported motivations. Differences in data extraction will be resolved through discussion until consensus is reached, or by referral to a third reviewer.

Analysis

The flow of information through the systematic review will be reported in accordance with the PRISMA statement [16].

We will calculate descriptive summary statistics using frequencies and percentages of responses to categorical data. Free text will be classified and frequencies and percentages of the categories will be presented in the analysis. For continuous data the range and mean with SD or median with IQR will be presented as appropriate.

Discussion

Our systematic review is designed to establish the motivations for using the CRXO design, the values of the CRXO design parameters, how both the sample size calculations and analyses account for the correlation structure and the incorporation of any covariates, and the quality of reporting the CRXO design, including the reporting of the correlation structure.

Strengths and limitations of our protocol

To our knowledge, this will be the first systematic review of CRXO trials with a rigorous and pre-specified methodology. We have pre-defined our screening and data extraction forms. Where possible, reviewers will classify article text according to pre-defined categories rather than categorize the free text after all the data has been captured. Pre-specifying the methodology and data collection reduces the risk of introducing bias into the review. The full text screening and data extraction will be performed by two reviewers. A subset of the abstract screening will be performed by two reviewers. The data abstraction form has been piloted on several articles by more than one reviewer.

A limitation of this review is the difficulty in identifying CRXO trials. Trials that use cluster randomization frequently do not use the word 'cluster' in the title or abstract, and it is often not apparent that the allocation of the intervention was at the cluster level unless the methods are read in the full text article [17]. In an attempt to limit missed studies, the search strategy encompasses units that are typically cluster randomized (such as schools or hospitals) and the references of all eligible articles will be searched. In addition, a search for CRXO methodology articles, and articles which cite them, will be undertaken to identify further trials.

CRXO designs may be employed in areas outside of clinical trials undertaken in humans, for example, variants of split-plot designs in agricultural sciences. There may be studies in behavioral, social, or educational sciences which will be missed by the search methodology employed in this review as our search is restricted to a limited number of databases. However, while the application of the CRXO design in these fields may be interesting from a methodological perspective, the focus of this systematic review is cluster randomization and crossover of interventions in human clinical trials in health.

We are interested in the design, methods, and motivations for using the CRXO design. Our ability to assess some of these elements may be limited because of missing and incomplete reporting in the trial publications. While contact with trial authors may help establish some of missing elements, we do not plan to contact authors since

we wish to reflect the information as reported. Decision-makers are generally reliant on only the information within publications, and therefore examining the quality and completeness of reporting is important. Knowledge of the adequacy of reporting is an essential step in developing reporting guidelines for such trials, if a need is found [18].

Both the stepped wedge design and the split-cluster design have similarities with the CRXO design. However, these designs were not considered in this review. A systematic review of the stepped wedge design was performed by Mdege *et al.* [19]. The split-cluster design does not have distinct time periods, so any similarity between the sub-clusters at a single point in time is likely to be different in nature to the similarity in clusters between time periods.

Implications of this research

Results from our systematic review will allow for focused methodological research of the CRXO design. The results will also provide practical examples for researchers of how CRXO trials have been conducted, including any shortcomings, and highlight areas where reporting and conduct may be improved.

Appendix 1: Search strategies

Ovid MEDLINE search

CROSS OVER TERMS

1. (cross-over or cross?over or "cross* over").tw.
2. (switch-over or switch?over or "switch* over" or switch-back or switch?back or "switch* back" or switched).tw.
3. ((change-over or change?over or "change* over") not ((change-over or change?over or "change* over") adj1 time)).tw.
4. (ab*ba* adj3 design*).tw.
5. exp Cross-Over Studies/
6. 1 or 2 or 3 or 4 or 5

CLUSTER ALLOCATION TERMS

7. ((unit$1 or school$1 or hospital$1 or cluster* or region $1 or ward* or practice* or communit* or population* or facility or facilities or practitioner*) adj15 random*).tw.
8. ((unit$1 or school$1 or hospital$1 or cluster* or region $1 or ward* or practice* or communit* or population* or facility or facilities or practitioner*) adj15 interven*).tw.
9. ((group* adj random*) or (group* adj interven*)).tw.
10. 7 or 8 or 9

HUMANS ONLY

11. Humans/
12. Animals/
13. 12 not 11

COMBINE CONCEPTS

14. 6 and 10
15. 14 not 13

PubMed search

CROSS OVER TERMS

1. "cross-over"[tiab] OR crossover[tiab] OR "cross over" [tiab] OR "crossed over"[tiab]
2. "switch-over"[tiab] OR switchover[tiab] OR "switch over"[tiab] OR "switch-back"[tiab] OR switchback[tiab] OR "switch back"[tiab] OR switched[tiab]
3. (change-over[tiab] OR changeover[tiab] OR "change over"[tiab] OR "changed over"[tiab] OR "changes over"[tiab]) not ("change-over time"[tiab] OR "changeover time"[tiab] OR "change over time"[tiab] OR "changed over time"[tiab] OR "changes over time"[tiab])
4. ab*ba[tiab]
5. Cross-Over Studies[mh]
6. #1 OR #2 OR #3 OR #4 OR #5

CLUSTER ALLOCATION TERMS

7. (cluster-randomi*[tiab] OR "cluster randomized" [tiab] OR "cluster randomised"[tiab] OR "cluster randomization"[tiab] OR "cluster randomization" [tiab])

HUMANS ONLY

8. (Animals[mh] NOT Humans[mh])

COMBINE CONCEPTS

9. #6 AND #7
10. #9 NOT 8
11. #10 NOT MEDLINE[sb]

EMBASE search via embase.com

CROSS OVER TERMS

1. (cross-over or crossover or "cross over" or "crosses over" or "crossed over" or "crossing over"):ti:ab
2. (switch-over or switchover or "switch over" or "switches over" or "switched over" or switch-back or "switchback" or "switch back" or "switches back" or "switched back" or switched):ti:ab
3. ((change-over or changeover or "change over" or "changes over" or "changed over") not ((change-over or changeover or "change over" or "changes over" or "changed over") near/1 time)):ti:ab

4. (abba near/3 design):ti:ab or (abba near/3 designs):ti:ab
5. “crossover procedure”/exp
6. #1 or #2 or #3 or #4 or #5

CLUSTER ALLOCATION TERMS

7. ((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) near/15 (random or randomly or randomise or randomize or randomised or randomized or randomises or randomizes or randomisation or randomization)):ti:ab
8. ((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) near/15 (intervene or intervention or interventions)):ti:ab
9. ((group or groups or grouped) near/1 (random or randomly or randomise or randomize or randomised or randomized or randomises or randomizes or randomisation or randomization)):ti:ab or ((group or groups or grouped) near/1 (intervene or intervention or interventions)):ti:ab
10. #7 or #8 or #9

HUMANS ONLY

11. ‘animal’ not ‘human’

COMBINE CONCEPTS

12. #6 and #10
13. #12 not #11
14. #13 not ‘medline’

CINAHL Plus search
CROSS OVER TERMS

1. TI (("cross-over" or "cross?over" or "cross* over")) OR AB (("cross-over" or "cross?over" or "cross* over"))
2. TI (("switch-over" or "switch?over" or "switch* over" or "switch-back" or "switch?back" or "switch* back" or switched)) OR AB (("switch-over" or "switch?over" or "switch* over" or "switch-back" or "switch?back" or "switch* back" or switched))
3. TI ((("change-over" or "change?over" or "change* over") not (("change-over" or "change?over" or "change* over") n1 time))) OR AB ((("change-over" or "change?over" or "change* over") not (("change-over" or "change?over" or "change* over") n1 time)))
4. TI (ab*ba* n3 design*) OR AB (ab*ba* n3 design*)
5. (MH "Crossover Design")
6. S1 or S2 or S3 or S4 or S5

CLUSTER ALLOCATION TERMS

7. TI (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 random*)) OR AB (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 random*))
8. TI (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 interven*)) OR AB (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 interven*))
9. TI (((group* n1 random*) or (group* n1 interven*))) OR AB (((group* n1 random*) or (group* n1 interven*)))
10. S7 or S8 or S9

HUMANS ONLY

11. (MH "Human")
12. (MH "Animals")
13. S12 not S11

COMBINE CONCEPTS

14. S6 and S10
15. S14 not S13
16. Exclude MEDLINE

Additional file

Additional file 1: This additional file contains the data that will be extracted from included studies.

Abbreviations
CRXO: Cluster Randomized CrossOver; ICC: Intra-cluster Correlation Coefficient.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SA drafted the manuscript, search strategies and data extraction form, and participated in the testing of the search strategy and data extraction form. AF conceived of the review, participated in the development and testing of the search strategy and data extraction form, and helped draft the manuscript. BK revised and tested the search strategy and data extraction form, and critically reviewed the manuscript. KM revised and tested the search strategy and data extraction form, and critically reviewed the manuscript. SM developed and tested the search strategy and critically reviewed the manuscript. JM participated in the development and testing of the search strategy and data extraction form, and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Monica Taljaard for providing an unpublished data extraction form as a starting point for our data extraction form, particularly for the population details and intervention type sections. Funding was provided to Katy Morgan through an NIHR (National Institute for Health Research) research methods fellowship. Joanne McKenzie is supported by an NHMRC (National Health and Medical Research Council) Australian Public Health Fellowship (grant number: 1072366).

Author details
[1]School of Public Health and Preventive Medicine, Monash University, Level 6, The Alfred Centre, Melbourne, VIC 3004, Australia. [2]Pragmatic Clinical Trials Unit, Queen Mary University of London, 58 Turner St, London E1 2AB, UK.

Received: 10 June 2014 Accepted: 28 July 2014
Published: 12 August 2014

References

1. Chan A-W, Altman DG: **Epidemiology and reporting of randomised trials published in PubMed journals.** *Lancet* 2005, **365:**1159–1162.
2. Eldridge S, Kerry S: *A Practical Guide to Cluster Randomised Trials in Health Services Research.* UK: John Wiley & Sons, Ltd; 2012.
3. Turner RM, White IR, Croudace T, Group PIPS: **Analysis of cluster randomized cross-over trial data: a comparison of methods.** *Stat Med* 2007, **26:**274–289.
4. Parienti JJ, Kuss O: **Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies.** *Contemp Clin Trials* 2007, **28:**316–323.
5. Rietbergen C, Moerbeek M: **The design of cluster randomized crossover trials.** *J Educ Behav Stat* 2011, **36:**472–490.
6. Giraudeau B, Ravaud P, Donner A: **Sample size calculation for cluster randomized cross-over trials.** *Stat Med* 2008, **27:**5578–5585.
7. Reich N, Milstone A: **Improving efficiency in cluster-randomized study design and implementation: taking advantage of a crossover.** *Open Access J Clin Trials* 2014, **6:**11–15.
8. Palmer RH, Louis TA, Hsu LN, Peterson HF, Rothrock JK, Strain R, Thompson MS, Wright EA: **A randomized controlled trial of quality assurance in sixteen ambulatory care practices.** *Med Care* 1985, **23:**751–770.
9. Richmond R, Heather N, Wodak A, Kehoe L, Webster I: **Controlled evaluation of a general practice-based brief intervention for excessive drinking.** *Addiction* 1995, **90:**119–132.
10. Dahler-Eriksen BS, Lauritzen T, Lassen JF, Lund ED, Brandslund I: **Near-patient test for C-reactive protein in general practice: assessment of clinical, organizational, and economic outcomes.** *Clin Chem* 1999, **45:**478–485.
11. Dearlove J, Illingworth S: **A controlled trial of parent initiated and conventional preschool health surveillance using personal child health records.** *Arch Dis Child* 1999, **80:**507–510.
12. Glazener CM, Ramsay CR, Campbell MK, Booth P, Duffty P, Lloyd DJ, McDonald A, Reid JA: **Neonatal examination and screening trial (NEST): a randomised, controlled, switchback trial of alternative policies for low risk infants.** *BMJ* 1999, **318:**627–631.
13. Flanagan JR, Doebbeling BN, Dawson J, Beekmann S: **Randomized study of online vaccine reminders in adult primary care.** *Proc AMIA Symp* 1999, **1999:**755–759.
14. Harris MA, Byles JE, Cockburn J, D'Este C: **A general practice-based recruitment strategy for colorectal cancer screening.** *Aust N Z J Public Health* 2000, **24:**441–443.
15. van Alem AP, Vrenken RH, de Vos R, Tijssen JG, Koster RW: **Use of automated external defibrillator by first responders in out of hospital cardiac arrest: prospective controlled trial.** *BMJ* 2003, **327:**1312.
16. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P: **Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.** *BMJ* 2009, **339:**b2535.
17. Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, Donner A: **Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards.** *BMC Med Res Methodol* 2010, **10:**15.
18. Moher D, Schulz KF, Simera I, Altman DG: **Guidance for developers of health research reporting guidelines.** *PLoS Med* 2010, **7**(2):e1000217.
19. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ: **Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation.** *J Clin Epidemiol* 2011, **64:**936–948.

doi:10.1186/2046-4053-3-86
Cite this article as: Arnup *et al.*: The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review. *Systematic Reviews* 2014 **3**:86.

Chapter 3

Appropriate methods were infrequently used in cluster randomised crossover trials

This chapter consists of the first of two papers that present the results of the systematic review. The protocol for the systematic review was provided in Chapter 2. The second paper is presented in Chapter 4.

In Chapter 3, the following aspects of the systematic review are presented 1) the characteristics of reported trials, 2) the motivations for using the CRXO design, 3) the values of the CRXO design parameters, and 4) the justification and methodology provided for the published sample size calculation and analyses. In addition, the implications of using inappropriate sample size and analysis methodology are discussed.

Chapter 3 is presented as a manuscript, which was published in *Journal of Clinical Epidemiology* in June 2016. The pages have been renumbered for the thesis, but the manuscript is otherwise unchanged.

Supplementary tables and additional files referred to in the manuscript as "Appendix" are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix B	Table S1 / Appendix	A summary of the changes in methods in the published manuscript from the protocol methods
	Table S2 / Appendix	The country where each included trial was conducted
	Table S3 / Appendix	The type of randomised cluster in each included trial
	Table S4 / Appendix	The justifications provided for use of the cluster randomised crossover design by type of cluster randomised
	Additional file 1 / Appendix	The revised electronic data extraction form used for the systematic review
	Additional file 2 / Appendix	The reference list for trials included in the systematic review

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 74 (2016) 40–50

Journal of Clinical Epidemiology

REVIEW ARTICLES

Appropriate statistical methods were infrequently used in cluster-randomized crossover trials

Sarah J. Arnup[a], Andrew B. Forbes[a], Brennan C. Kahan[b], Katy E. Morgan[c], Joanne E. McKenzie[a,*]

[a]*School of Public Health and Preventive Medicine, The Alfred Centre, Monash University, 99 Commercial Road, Melbourne, Victoria 3004, Australia*
[b]*Pragmatic Clinical Trials Unit, Queen Mary University of London, 58 Turner St, E1 2AB London, UK*
[c]*Medical Statistics Department, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*

Accepted 20 November 2015; Published online 26 November 2015

Abstract

Objective: To assess the design and statistical methods used in cluster-randomized crossover (CRXO) trials.

Study Design and Setting: We undertook a systematic review of CRXO trials. Searches of MEDLINE, EMBASE, and CINAHL Plus; and citation searches of CRXO methodological articles were conducted to December 2014. We extracted data on design characteristics and statistical methods for sample size, data analysis, and handling of missing data.

Results: Ninety-one trials including 139 end point analyses met the inclusion criteria. Trials had a median of nine clusters [interquartile range (IQR), 4–21] and median cluster-period size of 30 individuals (IQR, 14–77); 58 (69%) trials had two periods, and 27 trials (30%) included the same individuals in all periods. A rationale for the design was reported in only 25 trials (27%). A sample size justification was provided in 53 (58%) trials. Only nine (10%) trials accounted appropriately for the design in their sample size calculation. Ten of the 12 cluster-level analyses used a method that accounted for the clustering and multiple-period aspects of the design. In contrast, only 4 of the 127 individual-level analyses used a potentially appropriate method.

Conclusions: There is a need for improved application of appropriate analysis and sample size methods, and reporting, in CRXO trials.

Keywords: Cluster-randomized crossover trial; Crossover; Cluster; Sample size; Design; Statistical analysis

1. Introduction

The cluster-randomized crossover (CRXO) design is gaining popularity in settings where cluster randomization is required, but the parallel group cluster-randomized design is not feasible because the required number of clusters is prohibitively large [2,3]. In the CRXO design, hospitals, schools, or other groups of people (“clusters”) are randomly assigned to a sequence of interventions. Each cluster receives each intervention at least once in a separate period of time, leading to the formation of “cluster-periods” [4,5]. Within each cluster, each cluster-period may contain a repeated cross-section of different individuals, a cohort of the same individuals who are followed over time, or a mixture of the same and different individuals [6].

This design differs from the parallel group cluster-randomized design and the individually randomized crossover design. In the parallel group cluster-randomized design [7], each cluster is assigned only a single intervention, rather than a sequence of interventions. Each cluster therefore contains a single cross-section of different individuals. In the individually randomized crossover design [8], a cohort of individuals, rather than a series of clusters of individuals, are randomly assigned to a sequence of interventions. We refer the reader to Hooper and Bourke [9] for examples of other cluster designs conducted over multiple periods.

In both the individually randomized crossover design and CRXO design, randomization of the intervention

Funding: S.J.A. was supported in part by a Monash University Graduate Scholarship and a National Health and Medical Research Council of Australia Centre of Research Excellence grant (1035261) to the Victorian Centre for Biostatistics (ViCBiostat). K.E.M. was funded through a National Institute for Health Research (NIHR) research methods fellowship (MET-12-16). J.E.M. was supported by a National Health and Medical Research Council (NHMRC) Australian Public Health Fellowship (1072366).

Conflicts of interest: None.

* Corresponding author. Tel.: +61-3-9903-0380; fax: +61-3-9903-0556.

E-mail address: joanne.mckenzie@monash.edu (J.E. McKenzie).

http://dx.doi.org/10.1016/j.jclinepi.2015.11.013

What is new?

Key findings

- Reporting of the rationale for using cluster-randomized crossover (CRXO) trials was uncommon, despite this being a recommended reporting item for cluster-randomized trials [1]. Sample size calculations were commonly not reported, and only a minority of CRXO trials used sample size methods that appropriately accounted for the design.
- Only rarely did the used statistical methods account for the design, that is, adjust for the clustering and multiple period aspects.

What this adds to what was known?

- This is the first systematic review of CRXO trials. The results of this review provide a comprehensive assessment of the design characteristics, statistical methods for sample size, data analysis, and handling of missing data in CRXO trials.

What is the implication and what should change now?

- Trialists need to account for both the cluster randomization and multiple period aspects of the design in sample size calculations and statistical analyses. Methods and assumptions need to be clearly reported and justified.
- The development of reporting guidelines for CRXO trials is needed to facilitate clearer and complete reporting.

sequence serves to control for period effects (i.e., changes that occur over time that are unrelated to the intervention); and a key requirement of the two period design is that the effect of an intervention given in one period does not carry over into the next period [5,8]. In CRXO designs where cluster-periods contain different individuals, the potential for carryover is limited because any carryover can only take place at the cluster level. However, in CRXO designs where the same individuals are followed over time, carryover can also take place at individual subject level, and therefore, its potential is similar to that in individually randomized crossover designs.

The efficiency of a CRXO trial relative to an individually randomized trial or a parallel group cluster-randomized trial depends on the relationship between the outcomes from individuals within and between each cluster-period [6]. Individuals within a cluster tend to have more similar outcomes than individuals across clusters [7]. For example, because of differences in case-mix between patients presenting to different hospitals, patients in the same hospital may have more similar outcomes than patients in other hospitals. Likewise, individuals within a cluster-period tend to have more similar outcomes than individuals in different clusters. This similarity is typically measured by the within-cluster within-period intracluster correlation (ICC) [3–6,10]. This tendency for similar outcomes increases the uncertainty in the estimation of the effect of each intervention compared with outcomes that are independent.

If the environment of the cluster remains similar over time, then the outcomes of individuals within each cluster across different cluster-periods tend to be similar also. This tendency is typically measured by the within-cluster between-period ICC [3,5,6,10]. By comparing the interventions within cluster, the cluster-specific variation is removed from the comparison, and the uncertainty of the difference between interventions is decreased when there is a positive within-cluster between-period ICC [5]. Therefore, the crossover element of the design can offset the loss of precision arising from cluster randomization.

In the analysis of data from a parallel group cluster-randomized trial, it is recognized that the analysis must account for the correlation within clusters to correctly estimate the uncertainty in the intervention effect, for example, by including the cluster unit of randomization as a random effect in a generalized linear model (GLM) (e.g., Eldridge 2012). However, it is unclear whether trialists recognize that both the within-cluster within-period and the within-cluster between-period ICCs must be appropriately incorporated into sample size calculations and analyses to yield appropriate sample sizes and intervention effects with the correct standard errors in CRXO trials.

There have only been limited reviews examining the application and use of analytical methods for CRXO trials. These reviews have taken place in the introductory sections of methodological articles with the purpose of illustrating the design and highlighting the need for appropriate methods of analysis [4,5,10,11]. Therefore, we used systematic review methodology to examine the settings, design characteristics, justifications for using the design, quality of reporting, and sample size and analysis methods of trials that have used the CRXO design [12]. In this article, we focus on the design characteristics; statistical methods for sample size and data analysis, and the appropriateness of those methods; and the completeness of reporting of the statistical methods.

We begin with a brief review of recommended sample size and analysis methods for CRXO trials in Section 2. In Section 3, we outline the systematic review methods. Results are presented in Section 4 and discussed in Section 5.

2. Brief review of sample size and analysis methods for CRXO trials

Only limited methodological research has been published to guide trialists in performing sample size

calculations for CRXO trials. Giraudeau et al. and Donner et al. derived an inflation factor for a two-period two-group CRXO trial relative to an individually randomized trial, when the cluster-period sizes are assumed to be equal, and this was extended to incorporate period effects and unequal cluster-period sizes by Forbes et al. [3,10,13].

CRXO trials can be analyzed at the level of the individual or at the level of the cluster. The statistical challenges for the analysis of CRXO trials differ according to whether an individual-level or cluster-level analysis method is chosen; however, the target parameter of the analysis remains the intervention effect at the individual level [4,14].

For individual-level analyses, the use of mixed-effects models with continuous outcomes that include random or fixed effects for cluster and cluster-by-period effects has been shown to generally perform well in numerical simulations [4]. The performance of logistic mixed models for binary outcomes has only recently received attention with initial results indicating that poor performance may occur even with 50 or more clusters depending on the ICC values (Morgan et al., unpublished result). Other outcome types, including count and time-to-event data, have not been explored. Generalized estimating equations (GEEs) are also used to analyze clustered data [15]. In CRXO trials, accounting for both ICCs requires the inclusion of robust standard errors (possibly with adjustments for small sample variance estimation [16]), or extensions which model the patterned correlation structure within clusters across multiple periods explicitly [3].

Turner et al. [4] recommended a simple approach using a cluster-level analysis for continuous outcomes collected at the individual-level; however, this method can also be applied to other outcome types. In a cluster-level analysis, the available data on each individual from each cluster-period are aggregated into a single measure, and for each cluster, the relevant difference between interventions is constructed. Aggregating the data in each cluster-period accounts for the within-cluster within-period ICC, and comparing the cluster-period summaries within cluster accounts for the within-cluster between-period ICC. This method has been evaluated by both Turner et al. and Forbes et al. and performs well [3,4].

3. Systematic review methods

The protocol for the review has been published [12]. We provide a brief overview of the methods, along with deviations from the planned methods.

3.1. Literature search

The following sources were searched (to December 2014) for CRXO trials: MEDLINE, PubMed, EMBASE, and CINAHL Plus. In addition, CRXO methodology articles were searched to identify further references to CRXO trials. We searched for methodology articles in PubMed using the following search strategy: ((cluster[tiab] AND cross*over[tiab]) OR cluster-crossover[tiab]) AND (method*[tiab] OR design [tiab] OR calcul*[tiab] OR analy*[tiab]). A citation search of all identified methodology articles was performed in Web of Science. Finally, the references of all eligible articles were screened for CRXO trials. No restriction was applied to the publication date.

3.2. Trial inclusion criteria

Trials and protocols for trials that met the following inclusion criteria were included in the review: the trial was undertaken in humans; the trial was reported in English; the allocation of the intervention was to clusters of individuals rather than individuals themselves; each cluster received each intervention in a sequence over time (conventional crossover design) or at least some clusters crossed over from one intervention to another (such as two-treatment-four-sequence designs AA, AB, BA, and BB); at least some clusters crossed each way between at least two interventions (e.g., one cluster received AB and one cluster received BA); and the intervention given in the one period was not deliberately intended by design to affect individuals in subsequent periods (e.g., interventions intended to change the prescribing behavior of health care provider, where patients form the cluster). Two criteria were added to the planned criteria; see Table S1/Appendix at www.jclinepi.com for further details on changes to the published protocol.

3.3. Selection of trials for inclusion in the review

One author (S.J.A.) screened all titles and abstracts using the predefined eligibility criteria, and 50% of the titles and abstracts were screened independently by at least one coauthor. Full articles were then screened by one author (S.J.A.) using the predefined eligibility criteria. All eligible articles were double screened along with 20% of articles that were initially determined to be ineligible by S.J.A. Differences were resolved by discussion or by referral to a third author. No ineligible articles were subsequently found to be eligible.

3.4. Data extraction

The data extraction form incorporated items from the CONSORT extension to cluster-randomized trials [1] and systematic reviews of the design, reporting, and methodological aspects of stepped wedge [17], individual crossover [18], and parallel group cluster-randomized trials [19,20]. The data extraction form was piloted on five trials by each author. This resulted in modifications and clarifications to the form (see Additional File 1/Appendix at www.jclinepi.com). Data were entered into a database (Microsoft Access 2010, Redmond, WA, USA). The extracted information included identification of the design in the title or abstract, justification for using the design, acknowledgment

of the underlying assumptions of the design, demographic details (country, setting, unit of clustering, type of intervention, and control), design characteristics, methods used in the trial (recruitment, randomization, allocation, and blinding), reporting of baseline characteristics of the trial design, and statistical analysis (methods to estimate intervention effects and adjustment for covariates). The extracted design characteristics included the following: number of clusters, number of periods, number of cluster-periods (clusters × periods), number of individuals in the trial, number of interventions and the allocation of interventions to cluster-periods, the variability of the number of individuals between cluster-periods, the reported measure of similarity between the outcomes of individuals within a cluster within a given period, and the reported measure of similarity between outcomes of individuals within a cluster between different periods.

We wished to collect information on the range of statistical methods used within the trials, so implemented the following process to select outcomes and their associated statistical methods. We collected information on the primary outcome, where we defined primary using the following hierarchy: the first eligible primary outcome in the protocol document or first published article for the study if there is no protocol document; the outcome used for the sample size calculation; or the first outcome listed in the methods section of the abstract. We then collected information on multiple secondary outcomes, selected using the following process: the outcome was reported in the abstract and of a different data type to the primary outcome; the outcome was reported in the abstract and of the same data type as the previously included outcomes but analyzed by a different method; the outcome was reported in the article and of different data type or analyzed by a different method as the previously included outcomes.

One author (S.J.A.) extracted data from all trials, and data from 20% of the trials were independently double data extracted by the coauthors. Three of the five authors (S.J.A., J.E.M., and A.B.F.) reviewed the discrepancies arising from the double data extraction and discussed processes for further reviewing items where there was inconsistency. We rereviewed the following items for all trials: contamination as a justification for using the design; use of a washout period; blinding of the deliverers of the intervention; and mean cluster-period size. The sample size outcome scale was reviewed in all trials that used a count analysis. The scale of the outcome measure used in the sample size calculation was reviewed in all trials that were initially classified as count or binary. Any binary outcome that can occur multiple times per person in the period of measurement was classified as a count outcome. We classified binary outcomes that are associated with a period separately to binary outcomes measured over a fixed period because the statistical and sample size issues for these outcomes need to be considered separately. S.J.A. reviewed the following fields again in 10 randomly selected trials: method of recruitment, allocation and blinding; reporting of baseline characteristics; sample size outcome scale; and use of covariates in analysis. A.B.F. and J.E.M. reviewed the mean cluster-period size again in 20 randomly selected trials.

3.5. Defining the appropriateness of sample size and analysis methods used in the CRXO trials

We classified the sample size and analysis methods in each trial as either ‘‘potentially appropriate’’ or ‘‘inappropriate.’’ Given there is limited methodological research investigating the performance of sample size and analysis methods (Section 2), it was not possible to classify the methods as definitely ‘‘appropriate.’’ We attempted to replicate the sample size calculation to assist in classifying the sample size methodology. Any trial that reported a method which attempted to adjust for both the cluster randomization and multiple period design aspects, or equivalently the within-cluster within-period and within-cluster between-period ICCs, was considered to use ‘‘potentially appropriate’’ methodology. Methods that aggregated the data in each cluster-period were judged to have accounted for the cluster randomization design aspect. Trials using either GEE or GLM methodology that reported applying robust standard errors were considered ‘‘potentially appropriate’’ if the number of clusters in a trial was at least 30, based on recommendations for parallel group cluster-randomized trials (Hayes, 2009; pg 223) [14]. However, the performance of robust standard errors has not been extensively studied in CRXO trials, and in reality, a higher number of clusters may be required (Morgan et al., unpublished result) [7]. Trials that reported using the statistical packages R (R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.) or SAS (Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA) to fit GEEs were assumed to have applied robust standard errors because robust standard errors are fitted by default in these packages. Methods that did not adjust for both the cluster randomization and multiple period design aspects, but explicitly stated that the within-cluster within-period and within-cluster between-period ICCs were assumed to be equal were considered also to be ‘‘potentially appropriate.’’

3.6. Analysis

We present descriptive summary statistics using frequencies and percentages of responses to categorical data. Free text was classified, and frequencies and percentages of the categories are presented. For continuous data, the range and mean/standard deviation or median/interquartile range (IQR) are presented as appropriate. The individual trial data can be made available on request to the corresponding author.

4. Results

4.1. Results of the search

Fig. 1 shows the flow diagram of the CRXO trial selection process. Of the 3,425 records identified through database searching, 170 were duplicates and 3,046 were ineligible based on screening of abstracts, leaving 209 full-text articles to assess for eligibility. Of these 209 articles, 98 were assessed as eligible. A further four articles were identified through the methodology article reference and citation search and four articles from the references of eligible articles. In total, 106 articles from 91 unique trials were included in the review (see Additional File 2/ Appendix at www.jclinepi.com). Seventy-nine trials had

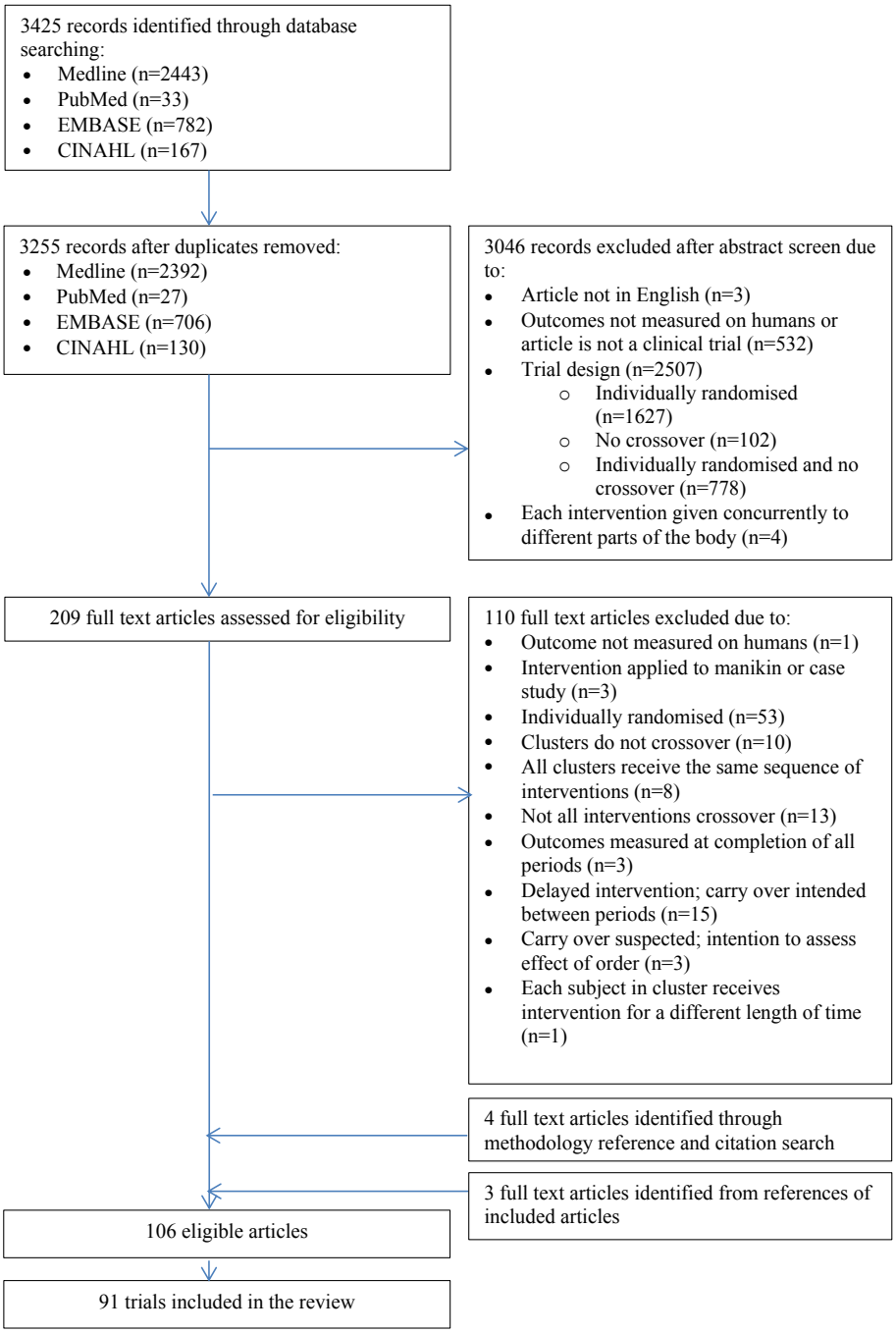


Fig. 1. Flow of articles through the systematic review.

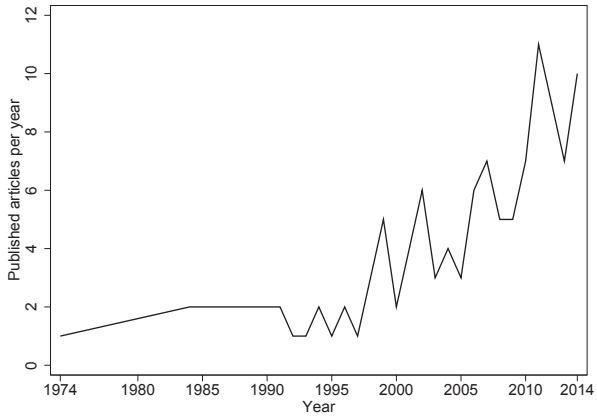


Fig. 2. Number of cluster-randomized crossover publications per 5-year period.

only one associated article (eight of which were protocols), nine trials had two associated articles, and three trials had three associated articles.

4.2. Characteristics of the trials

The earliest identified trial was published in 1974. Half of the 91 trials were published after 2006, with nine trials published in 2014 (Fig. 2). Most trials were conducted in a developed country (n = 86, 95%) and were conducted within only one country (n = 86, 95%) (Table S2/ Appendix at www.jclinepi.com).

The types of clusters varied. Almost half of the trials (n = 45, 49%) randomized hospitals or wards within hospitals, and of these, 19 (21%) randomized intensive care units. Thirteen trials (14%) each randomized individual health care providers and schools or classes (Table S3/ Appendix at www.jclinepi.com).

The trials investigated a wide range of diseases and conditions and health care delivery models. Twenty trials (22%) investigated infection control, 11 (12%) investigated infectious diseases, 11 (12%) investigated cardiovascular disease, and 10 (11%) examined the delivery of health services (Table 1).

4.3. Design characteristics of the trials

The most common trial design included two interventions (n = 81, 89%) (Table 2). In 49 (54%) trials, the interventions were delivered directly to the individuals within the clusters. In 22 (24%) trials, the intervention was targeted at the health care provider rather than the individuals under their care, and in 14 (15%) trials, the intervention was targeted at the organization of the health care provider or health service delivery (Table 1).

Trials had a median of nine clusters (IQR, 4–21; range, 2–268) and had a median cluster-period size of 30 (IQR, 14–77; range, 2–1,319). Most trials (69%) used two periods. Trials randomizing hospitals used fewer clusters (median: 6; IQR, 2–10; range, 2–46), but were larger in size (median cluster-period size: 57; IQR, 21–197; range, 5–1,319) than trials randomizing schools and health care

Table 1. Characteristics of the cluster-randomized crossover trials

General characteristics	*N* (%)
Disease or domain under study	
Infection control	20 (22)
Infectious disease	11 (12)
Cardiovascular disease	11 (12)
Health services delivery	10 (11)
General and public health	6 (7)
Medical training	5 (5)
Communication of health information	4 (4)
Pregnancy, childbirth, and early childhood	3 (3)
Mental health and behavioral conditions	3 (3)
Respiratory disease	3 (3)
Blood sample contamination	3 (3)
Cognition	3 (3)
Central nervous system and musculoskeletal disease	2 (2)
Urogenital disease	2 (2)
Oral health	2 (2)
Nutritional and metabolic disorders	1 (1)
Digestive disorders	1 (1)
Pain management	1 (1)
Type of intervention	
Intervention targeting the individual	49 (54)
Intervention targeting health care provider	22 (24)
Quality improvement intervention	14 (15)
Intervention resulting in change to the participant environment	6 (7)
Justification for design	
Justification for both cluster randomization and crossover	25 (27)
Justification for neither cluster randomization or crossover	42 (46)
Justification for cluster randomization	36 (40)
Justification for crossover	38 (42)
Consent	
Individual or those acting on their behalf	32 (35)
Cluster	30 (33)
Opt out	4 (4)
Varied by site[a]	1 (1)
Unclear	24 (26)
Type of consent given by cluster-level decision maker[b] (n = 34)	
Participation and data collection	6 (18)
Participation with individual consent for data collection	23 (68)
Not stated who gave consent for data collection	5 (15)
Was the randomization sequence randomly generated?[c]	
No	3 (3)
Yes—sufficient information to replicate	35 (38)
Yes—insufficient information to replicate	41 (45)
Unclear	12 (13)
Covariates were used in the randomization	30 (33)

[a] Consent was sought from the individuals or was obtained at the cluster level, varying by randomization site.

[b] Includes consent given by cluster-level decision maker and opt out consent.

[c] A classification of no indicates that the treatment sequences were intentionally assigned to each cluster. The randomization procedure was judged to be insufficient if reported that the allocation of treatment sequences was randomized, but no further detail on the randomization was reported. If further details (e.g., toss of coin, computer randomization program) were provided, the procedure was judged to be sufficient.

Table 2. Design characteristics of the cluster-randomized crossover trials by type of cluster randomized

Design characteristics	All cluster types	Type of cluster randomized			
	Total *N* = 91; *n* (%)	Hospital *N* = 45; *n* (%)	School *N* = 13; *n* (%)	Health care provider *N* = 13; *n* (%)	Other cluster types[a] *N* = 20; *n* (%)
Number of interventions					
2	81 (89)	38 (84)	12 (92)	12 (92)	19 (95)
3	9 (10)	6 (13)	1 (8)	1 (8)	1 (5)
4	1 (1)	1 (2)	0	0	0
Number of clusters					
Median (IQR)	9 (4–21)	6 (2–10)	17 (9–22)	23 (18–34)	14 (5–40)
Range	2–268	2–46	4–46	3–64	2–268
Unclear	4 (4)	2 (4)	2 (15)	0	0
Number of periods[b]					
2	58 (69)	27 (60)	10 (83)	8 (73)	13 (81)
3	9 (11)	7 (16)	1 (8)	1 (9)	0
4+	17 (20)	11 (24)	1 (8)	2 (18)	3 (19)
Unclear	7 (8)	0	1 (8)	2 (15)	4 (20)
Cluster-period size					
Median (IQR)	30 (14–77)	57 (21–194)	23 (17–43)	20 (10–56)	12 (6–27)
Range	2–1,319	5–1,319	10–152	2–82	2–77
Unclear	21 (23)	9 (20)	1 (8)	2 (15)	8 (40)
Same participants in all periods[c]	27 (30)	3 (7)	12 (93)	3 (23)	9 (45)
Washout period or reason for not including washout period explained	45 (49)	27 (60)	7 (54)	3 (23)	8 (40)

Abbreviation: IQR, interquartile range.

[a] Other cluster types include the following: aged care facilities, dementia facilities, primary care practices, and outpatient facilities; households and geographic regions; and worksite departments, emergency responder teams, and individual patients (units receiving treatment were individual teeth or muscles).

[b] Percentages of nonmissing data presented.

[c] Data were dichotomized into same participants in all periods or no participants in multiple periods and some, but not all participants, in multiple periods.

providers. The same participants were included in all periods in 27 (30%) trials, with this occurring more commonly in trials randomizing schools (Table 2).

In 45 (49%) trials, a washout period was either incorporated into the design or the reason for not using one was explained. Trials randomizing health care providers used a washout less frequently (3/13, 23%) than trials randomizing hospitals (27/45, 60%) or schools (7/13, 54%) (Table 2).

Of the 91 included trials, 76 (84%) trials stated that clusters were assigned to intervention sequences at random. Only 35 (38%) of these trials provided sufficient detail to replicate the randomization. Thirty trials (33%) used restricted randomization to balance covariates between the intervention sequences. Consent to participate was sought from the individuals or those acting on their behalf in 32 trials (35%). In 30 (33%) trials, consent to participate was obtained at the cluster level, whereas in 24 (26%) trials, it was unclear how consent was sought (Table 1).

4.4. Justification for design

Only 25 (27%) trials provided justification for both the cluster randomization and crossover aspects of the design, whereas 42 (46%) trials provided no justification for either design aspect (Table 1). A justification was provided for the cluster randomization aspect in 36 trials (40%), and a justification for the crossover aspect was provided in 38 trials (42%). Justification for cluster randomization was given in only 3 of 13 (23%) of the trials that randomized health care providers, where interventions were primarily targeted at the cluster level (9 of 13 trials). In contrast, justification for cluster randomization was given in 22 (49%) of hospital trials, where the interventions were more frequently targeted at the individual level (62% of hospital trials) (Table S4/Appendix at www.jclinepi.com).

The main reasons cited for cluster randomization (n = 36) were to avoid contamination between individuals within the cluster by either the individuals themselves (11 trials, 31%) or by those delivering the intervention (11 trials, 31%) and because it was not practical to individually randomize (11 trials, 31%). The main reason cited for crossing over the intervention within a cluster (n = 38) was to attempt to eliminate differences in cluster-level characteristics between clusters (27 trials, 71% of reasons cited) (Table S4/Appendix at www.jclinepi.com).

4.5. Statistical methods for sample size estimation

Of the 91 trials, 53 provided some detail of a sample size calculation and 35 (38%) did not report a sample size calculation. None of the trials reported using unequal cluster-period sizes (Table 3).

Only nine trials (10%) used methods that appropriately accounted for both the within-cluster within-period and within-cluster between-period ICCs. Eleven trials (12%) used a method appropriate for a parallel group cluster-

Table 3. Sample size methods used in the cluster-randomized crossover trials

Sample size characteristics	n (%)
Trial reported a sample size calculation?	*N* = 91
No	35 (38)
No—justification for not reporting calculation provided	3 (3)
Yes—sufficient information to replicate calculation	39 (43)
Yes—insufficient information to replicate calculation	14 (15)
Sample size methods	*N* = 53
Method included covariates in the sample size calculation?[a]	1 (2)
Use of unequal cluster-period sizes?[b]	0
	N = 91
Methods appropriate for individually randomized parallel group design (outcomes assumed to be independent)	31 (34)
Methods appropriate for parallel group cluster-randomized design	11 (12)
Cluster accounted for, method unclear	1
Sample size inflated by design effect to account for within-cluster correlation	10
Methods appropriate for CRXO design	9 (10)
Paired cluster-level means	1
Sample size inflated by design effect that accounted for within-cluster within-period ICC and within-cluster between-period ICC	4
Sample size for individual crossover design inflated by design effect to account for within-cluster ICC	1
Sample size for stepped wedge design with verification using simulation	1
Sample size inflated by a best guess to account for CRXO design	1
Sample size estimated using simulation for CRXO design	1
Methods unclear	2 (2)
Trial reported parameters used to account for correlation between outcomes in sample size calculation	
Methods appropriate for parallel group cluster-randomized design (*n* = 11)	10 (91)
Methods appropriate for CRXO design (*n* = 9)	5 (56)

Abbreviations: CRXO, cluster-randomized crossover; ICC, intracluster correlation.

[a] Compared with no and unclear combined.

[b] Compared with equal cluster-period sizes. Cluster-period sizes were judged to be equal if the sample size could be reproduced by assuming equal sizes.

randomized trial, therefore ignoring the crossover aspect of the design. The remaining 33 trials either assumed the observations were independent (31 trials) or the reporting of sample size methodology was insufficient to make any assessment (two trials).

Trials that used a sample size calculation for a parallel group cluster-randomized design almost always reported the value used to account for the nonindependence of outcomes within each cluster (10 of 11 trials). Seven of these values were based on a best guess, two were taken from published research, and one trial quoted the maximum correlation between clusters that would guarantee 80% power for their fixed sample size. Of the nine trials accounting for the CRXO design, five reported the values used to account for the nonindependence of outcomes within and between periods within a cluster. In a further one trial: values were not reported; the within-cluster within-period ICC and within-cluster between-period ICC were assumed to cancel such that the sample size calculation assumed the outcomes were independent; the estimated reduction in power through simulation was reported, without reporting the values used within the simulation; and the sample size was inflated by a best guess value.

4.6. Statistical methods for data analysis

Across the 91 trials, 175 outcomes (median = 2 outcomes per trial; IQR, 1–2) met our inclusion criteria, from which we assessed the associated analytical methods. We excluded 36 of the analytical methods from further assessment because: the level of analysis was not clear (n = 12); the intervention effect was estimated separately in each cluster or the intervention effect was estimated between clusters separately in each period (n = 10); no comparison between intervention groups was made (n = 9); no information on the method used for comparison was provided (n = 3); or only descriptive statistics only were reported (n = 2). Of the remaining 139 analyses, 127 (91%) were performed at the individual level and 12 (9%) were performed at the cluster level. Across the 139 analyses, we deemed 14 (10%) to be potentially appropriate. We now detail the methods used by the level at which the analysis was undertaken (Table 4).

Of the 12 cluster-level analyses, 10 (83%) used a method that accounted for the correlation between cluster-period summaries within each cluster and therefore appropriately accounted for both the cluster randomization and crossover aspects of the design. In one analysis, the methodology was judged to be inappropriate and the methodology in the remaining analysis was unclear.

In the 12 cluster-level analyses, the observations were collapsed within cluster-periods to a cluster-period mean in five analyses, to a rate or count in another five analyses, and a log-incidence rate in one analysis. In the remaining analysis, the expected rate in each cluster-period was obtained from a GLM fitted with a log-link function and

Table 4. Statistical analysis methods used in the cluster-randomized crossover trials

Method	*N*	Potentially appropriate?	Cluster-level paired test	Permutation test	Fixed cluster effect[a]	Random cluster effect[a]	GEE	Robust standard errors[b]	Method did not account for cluster or cluster-period effect	Unclear if method accounted for cluster or cluster-period effect
Cluster level (*N* = 12)		10 (83%)								
GLM[c]	2	1	NA	NA	0	1	NA	1	0	0
ANOVA[d]	8	7	7	NA	0	0	NA	0	0	1
Nonparametric methods[e]	2	2	0	2	0	0	NA	0	0	0
Individual level (*N* = 127)		4 (3%)								
GLM[c]	64	1	NA	NA	8	24	NA	3	15	14
GEE[f]	11	3	NA	NA	NA	NA	4	7	NA	NA
ANOVA[d]	13	0	0	NA	2	0	NA	0	7	4
Other parametric models[g]	19	0	0	NA	4	0	NA	0	13	2
Nonparametric methods[e]	20	0	0	0	0	0	NA	0	19	1

Abbreviations: GLM, generalized linear model; ANOVA, analysis of variance; GEE, generalized estimating equation.

"NA" in the table indicates that the methodology in the column is not applicable to the analysis method in the row.

[a] Including trials where the either the cluster-period effect was not accounted for in the method or it was unclear if the cluster-period effect was accounted for.

[b] For GLM, a random effect for cluster was also used.

[c] Models including linear, logistic, poisson, binomial-identity link, ordinal, proportional hazards, and time series regression.

[d] Including ANOVA, repeated-measures ANOVA, and t-test.

[e] Including Kruskal–Wallis, Fisher's exact, Kaplan–Meier curve with log-rank test, Wilcoxon rank sum, exact test for incidence rates, and permutation tests.

[f] Links including normal-identity, binomial-logit, binomial-identity, and poisson-log.

[g] Including chi-squared, Mantel Haenszel chi-squared, McNemars test, and Wallenstein method.

individual-level covariates. The cluster-period summaries were compared using the following methods: paired t-test (seven analyses); permutation test (two analyses); and fitting a GLM with log-link function in two analyses: one applied robust standard errors to the estimate of the intervention effect and one included a random effect for cluster in the model. In the remaining analysis, it was unclear whether a paired or unpaired t-test was used.

In contrast, 4 of the 127 individual-level analyses potentially accounted appropriately for both the cluster randomization and crossover aspects of the design. Fifty-four analyses did not account for either the cluster randomization or crossover aspects of the design. The cluster unit was accounted for in 52 analyses: 35 analyses fitted a GLM that included a term for the cluster (including eight fixed effects and 27 random effects, three also with robust standard errors); a GEE approach was used in 11 analyses (seven of which were judged to be using robust standard errors); four analyses used a Mantel Haenszel stratified chi-squared test; and two analyses fitted an analysis of variance model with cluster as a fixed term. In the 10 analyses that applied robust standard errors, four included at least 30 clusters and were therefore classified as potentially appropriate. In 21 analyses, insufficient information was provided to determine if either the cluster randomization or crossover aspects were accounted for. No trials accounted for the crossover as recommended by Turner et al. and Parienti et al. which involved including cluster-by-period random effect terms [4,5].

4.7. Reporting and handling of missing data

Of 64 (70%) trials that reported missing data, 8 (13%) reported using a method to handle the missing data, including: use of random effects models, adjusted for covariates believed to be associated with missingness; multiple imputation to replace missing outcomes and covariates; and testing the sensitivity of the results to the missing data by substituting the missing values with extreme values.

5. Discussion

We undertook a systematic review to assess the design and statistical methods used in CRXO trials. CRXO trials have become more common over time and have been conducted within a variety of settings to assess a range of interventions. The methods used for the sample size calculation and analysis suggest that there was limited understanding of the effect of the cluster randomization and multiple period aspects of the design. There is a need for improved reporting of CRXO trials: justifications for using the design were rarely reported.

A key requirement of the CRXO design is that the intervention effect does not carry over from one period to the

next [5]. This can be achieved by many methods, for example using different participants in each period; providing an adequate washout time between periods; and blinding of the trialists involved in the delivery of the intervention and collection of outcome data. Although it is possible to undertake a statistical test of the interaction between treatment and period, it is not possible from this test to distinguish carryover effects from treatment by period interactions, in two-period two-intervention CRXO trials [21], and altering the analysis based on the results of this test leads to a biased estimate of the intervention effect and inflated type I error [22]. Regardless of the method used to reduce the risk of carryover, clear reporting of the method used is required to allow readers to assess the risk of bias to the intervention effect arising from the potential carryover.

Although the CRXO design is usually more efficient than parallel group cluster randomization, an individually randomized trial is usually more efficient than CRXO [3]. In over half of the CRXO trials included in our review, the intervention was delivered to the individual participants, and therefore, individual randomization may have been possible. Careful consideration should always be given to whether a more complex design is necessary, particularly when the design requires more participants [1].

5.1. Implications of the sample size and analysis methodology

5.1.1. Implications of statistical methods for sample size estimation

Over half (58%) of the trials that provided a sample size estimate reported using a method designed for an individually randomized trial using simple randomization. This approach is only appropriate for a CRXO trial when both the within-cluster within-period ICC and within-cluster between-period ICC are zero or the within-cluster between-period ICC is equal to within-cluster within-period ICC. These are strong and optimistic assumptions [3], and if violated, would lead to an underpowered study.

In trials where the sample size was estimated for a parallel group cluster-randomized design, the within-cluster between-period ICC is effectively treated as zero. This approach leads to a conservative sample size, and as a result, more participants will be included in the trial than are needed to obtain the desired power.

Following a suggestion by Donner et al., one trial assumed the within-cluster between-period ICC was half the within-cluster within-period ICC, which leads to a sample size estimate that is larger than the requirement for an individually randomized design but smaller than the requirement for an parallel group cluster-randomized trial [13]. In the absence of a priori knowledge of the within-cluster between-period ICC, this may be reasonable approach.

Other methods used to account for the CRXO design included inflating the sample size by an arbitrary amount; performing a power calculation using numerical simulation with a model that included both ICCs; basing the estimate on a paired t-test of the cluster-level means; and inflating the estimate from an individual-level crossover design with the same participants in both periods by an inflation factor for a parallel group cluster-randomized design. The first three methods may be appropriate if representative values of the sample size parameters were used. Further research is required to assess the appropriateness of the last method.

5.1.2. Implications of choice of statistical methods for cluster-level analyses

We found that when data were aggregated in each cluster period, the analyses used were usually appropriate. In addition, as noted by Turner et al., such cluster-level analyses are intuitive and easily understandable to health researchers [4]. Therefore, use of such an analysis approach would often seem reasonable.

5.1.3. Implications of choice of statistical methods for individual-level analyses

Analogous to the issues with sample size, individual-level analyses that assume observations within clusters and within cluster-periods are independent, or account only for cluster-level variation, can estimate standard errors that are too small. These analyses can result in inflated type I error rates and potentially lead to false-positive claims regarding the effectiveness of interventions.

5.1.4. Influence of statistical methodology articles on subsequent CRXO trials

The influence of statistical methodology articles on the use of appropriate sample size and analysis methods in CRXO trials would seem limited. Methods to perform sample size calculations for CRXO trials have been available since 2004 [10,13]; yet, the number of articles including a sample size calculation has remained around 60–70% since 2000, and only two CRXO trials cited a methods article for a CRXO sample size calculation. Methods for analyzing binary data in dental split-mouth trials have been available since 2004 [13], and in the context of CRXO trials, since 2007 [5]. Only four CRXO trials cited a methods article for analyzing CRXO trials, all citing Turner et al. [4]. It is unclear whether trialists recognize the need to use specialist methods when designing and analyzing these trials. Regardless, however, there is a clear need for development of accessible guidance for health researchers for the design, conduct, and analysis of CRXO trials.

5.2. Strengths and limitations

A potential limitation of our review was our ability to locate all CRXO trials. Locating CRXO trials is difficult because there is no validated search strategy, and the

language used to describe the design is inconsistent. Furthermore, many trialists may be unaware that they have used a CRXO design and so fail to use key words in the abstract that describe the clustering or crossover aspects. To optimize our yield of CRXO trials, we used a broad search strategy, searched references of all eligible CRXO trials, and undertook citation searches to CRXO trial methodology articles. Although our yield of CRXO trials may be incomplete, it represents the most comprehensive review of this trial design to date. Furthermore, it may be argued that CRXO trials that are better reported, and thus easier to locate, are also more likely to use appropriate statistical methods. Therefore, results from our review may present an optimistic view of the design and statistical methods used in CRXO trials.

We maintained consistency in the review by having one author perform all screening and data extraction and verified the results by having a subsample of all reviewed trials independently assessed by at least one other reviewer.

6. Conclusions

The CRXO design has been used in a wide range of settings for the past four decades. However, the statistical methods used in the sample size determination and analysis rarely account appropriately for the design aspects. The justifications for using the design are rarely reported. It is unclear whether trialists recognize the need for specialist methods in designing and analyzing these trials. There is an urgent need for accessible guidance for health researchers on the design, conduct, analysis, and reporting of the CRXO design.

Acknowledgments

The authors acknowledge Ross Dunn for his assistance in the development of the Access database and Steve McDonald for his assistance in the development of the search strategies. The authors thank Monica Taljaard for providing an unpublished data extraction form as a starting point for our data extraction form, particularly for the population details and intervention type sections.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.11.013.

References

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

[2] Bellomo R, Forbes A, Akram M, Bailey M, Pilcher DV, Cooper DJ. Why we must cluster and cross over. Crit Care resuscitation 2013;15:155−7.

[3] Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. Clin Trials 2015;12:34−44.

[4] Turner RM, White IR, Croudace T, Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. Stat Med 2007;26:274−89.

[5] Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. Contemp Clin Trials 2007;28:316−23.

[6] Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. J Educ Behav Stat 2011;36:472−90.

[7] Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. John Wiley & Sons, Ltd, Chichester, West Sussex, UK; 2012.

[8] Louis TA, Lavori PW, Bailar JC 3rd, Polansky M. Crossover and self-controlled designs in clinical research. N Engl J Med 1984;310:24−31.

[9] Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. BMJ 2015;350:h2925.

[10] Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Stat Med 2008;27:5578−85.

[11] Reich NGM, Aaron M. Improving efficiency in cluster-randomized study design and implementation: taking advantage of a crossover. Open Access J Clin Trials 2014;6:11−5.

[12] Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McDonald S, McKenzie JE. The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review. Syst Rev 2014;3:86.

[13] Donner A, Klar N, Zou G. Methods for the statistical analysis of binary data in split-cluster designs. Biometrics 2004;60:919−25.

[14] Hayes RJ, Moulton LH. Cluster randomised trials. Chapman and Hall/CRC, Boca Raton, FL, USA; 2009.

[15] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13−22.

[16] Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, Wolfson M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics 2007;63:935−41.

[17] Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol 2011;64:936−48.

[18] Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. Trials 2009;10:27.

[19] Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. J Clin Epidemiol 2015;68:716−23.

[20] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ 2011;343:d5886.

[21] Jones B, Kenwood MG. Design and analysis of cross-over trials. 2nd ed. London: Chapman & Hall/CRC; 2003.

[22] Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. Stat Med 1989;8:1421−32.

Chapter 4

The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality

Chapter 4 is the second of two papers that present the results of the systematic review. The protocol for the systematic review was provided in Chapter 2.

The joint aims of the research presented in this chapter are to 1) assess the reporting quality of published CRXO trials, 2) propose a set of reporting items for CRXO trials, and 3) inform the development of a reporting guideline.

The CONSORT 2012 cluster trials extension forms the basis for the proposed reporting items in this chapter. Aspects of CRXO design that are not addressed by the existing guidelines are indicated, and possible items to report on these aspects are proposed.

The results of the systematic review suggest that a reporting guideline for the CRXO design may be beneficial in improving the completeness of reporting the design, statistical methods, and results. Until such guidance is developed, the proposed items identified in this paper provide interim guidance.

Chapter 4 is presented as a manuscript, which was published in *Trials* in December 2016. The pages have been renumbered for the thesis, but the manuscript is otherwise unchanged. A blog article was invited by *BMC On Medicine* to summarise and promote the published manuscript. The published blog is included Appendix D.

Supplementary tables and additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix C	Additional file 1	Systematic review search strategy
	Additional file 2: Table S1	The country where each included trial was conducted
	Additional file 3: Table S2	The type of randomised cluster in each included trial

Arnup et al. Trials (2016) 17:575
DOI 10.1186/s13063-016-1685-6

Trials

RESEARCH Open Access

The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality

Sarah J. Arnup[1], Andrew B. Forbes[1], Brennan C. Kahan[2], Katy E. Morgan[3] and Joanne E. McKenzie[1*]

Abstract

Background: The cluster randomised crossover (CRXO) design is gaining popularity in trial settings where individual randomisation or parallel group cluster randomisation is not feasible or practical. Our aim is to stimulate discussion on the content of a reporting guideline for CRXO trials and to assess the reporting quality of published CRXO trials.

Methods: We undertook a systematic review of CRXO trials. Searches of MEDLINE, EMBASE, and CINAHL Plus as well as citation searches of CRXO methodological articles were conducted to December 2014. Reporting quality was assessed against both modified items from 2010 CONSORT and 2012 cluster trials extension and other proposed quality measures.

Results: Of the 3425 records identified through database searching, 83 trials met the inclusion criteria. Trials were infrequently identified as "cluster randomis(z)ed crossover" in title ($n = 7$, 8%) or abstract ($n = 21$, 25%), and a rationale for the design was infrequently provided ($n = 20$, 24%). Design parameters such as the number of clusters and number of periods were well reported. Discussion of carryover took place in only 17 trials (20%). Sample size methods were only reported in 58% ($n = 48$) of trials. A range of approaches were used to report baseline characteristics. The analysis method was not adequately reported in 23% ($n = 19$) of trials. The observed within-cluster within-period intracluster correlation and within-cluster between-period intracluster correlation for the primary outcome data were not reported in any trial. The potential for selection, performance, and detection bias could be evaluated in 30%, 81%, and 70% of trials, respectively.

Conclusions: There is a clear need to improve the quality of reporting in CRXO trials. Given the unique features of a CRXO trial, it is important to develop a CONSORT extension. Consensus amongst trialists on the content of such a guideline is essential.

Keywords: Cluster randomised crossover trial, Crossover, Cluster, Reporting quality

* Correspondence: joanne.mckenzie@monash.edu
[1]School of Public Health and Preventive Medicine, Monash University, The Alfred Centre, Melbourne, Victoria 3004, Australia
Full list of author information is available at the end of the article

Background

The cluster randomised crossover (CRXO) trial design has been used to evaluate a range of interventions, in a variety of settings [1]. In a CRXO trial hospital, schools or other groups of individuals ("clusters") are randomly assigned to a sequence of interventions. Each cluster receives each intervention at least once in a separate period of time, leading to the formation of "cluster-periods" [2, 3]. The design has potentially greater efficiency than a parallel group cluster randomised trial because the interventions are compared within each cluster [4].

Every trial design has specific features that need to be considered in the design, analysis, and reporting stages. In the case of the CRXO trial, a critical consideration is the correlation between participants within clusters through time. Individuals within a cluster tend to have more similar outcomes than individuals in different clusters. This similarity is typically measured by the within-cluster within-period intracluster correlation (ICC). Furthermore, the similarity between two individuals within a cluster is likely to dissipate as time increases between the measurement of the two individuals. The similarity between two individuals within a cluster, but in different time periods, is typically measured by the within-cluster between-period ICC [2–6].

Complete, transparent, and clear reporting of clinical trials is essential for those using trial reports. Comprehensive reporting allows for an assessment of threats to the validity of the trial results, an assessment of the adequacy of the statistical methods, replication of trial methodologies, incorporation of the trial's results in synthesis products such as meta-analyses, and implementation of the evaluated intervention(s). To assess the validity of the trial's results, the methodology should be reported in enough detail to allow for the evaluation of selection, performance, and detection biases [7].

The quality of reporting in randomised trials remains unacceptably low despite the introduction of reporting guidelines [8, 9]. The CONSORT (Consolidated Standards of Reporting Trials) statement reporting guideline for parallel group randomised trials was developed in an attempt to improve the quality of reporting of randomised trials. The CONSORT statement was first published in 1996, and has since been twice revised, first in 2001 [10] and then in 2010 [11]. The 2010 CONSORT statement includes 25 recommended items covering design, conduct, analysis, and other aspects. Extensions to the parallel group CONSORT statement have been published for some alternative designs; however, no extension currently exists for CRXO trials.

While a CONSORT extension is not available for CRXO trials, items from the 2012 cluster trials extension [12] and several items that have been proposed for reporting stepped wedge trials [13] are directly applicable (e.g. "Allowance for clustering" and "Allowance for the number of steps" in the sample size justification) or are easily modifiable (e.g. "Identification as a cluster randomised trial in title") for CRXO trials. However, the CRXO design has distinct characteristics when compared with the parallel group cluster randomised design and the stepped - wedge design, such as the adverse potential for carryover of the intervention effect to subsequent periods. Therefore, a separate reporting guideline for this trial design may be of value.

Assessing the quality of reporting is the suggested initial step in developing reporting guidelines [9]. Because no published reporting guidelines exist for CRXO trials, in this article we propose possible reporting items, and indicate areas where items may need to be developed, as a means to (1) facilitate discussion on possible items that could be considered for inclusion in a future reporting guideline, and (2) assess the quality of reporting in CRXO trials and thus determine if there is a need for a separate guideline.

To assess the quality of reporting in CRXO trials, we undertook a systematic review that collected information on a range of aspects including trialists' motivations for using the CRXO design, the design characteristics of CRXO trials, the statistical methods for sample size and data, and the quality of reporting of CRXO design aspects. In a previous publication, we evaluated the appropriateness of the statistical analysis and sample size methods [1]. In this article we evaluate the quality of reporting in CRXO trials.

In "Proposed reporting items for CRXO trials" we discuss potential modifications to the reporting items of the CONSORT 2012 cluster trials extension for CRXO trials, and propose areas where items may need to be developed. In "Systematic review methods" we outline the systematic review methods. The quality of reporting of CRXO trials is presented in the "Results" section. We discuss our findings and conclusions in the "Discussion" section.

Proposed reporting items for CRXO trials

In this section we suggest, and provide rationale for, possible modifications to reporting items of the CONSORT 2012 cluster trials extension for CRXO trials, and propose areas where items may need to be developed to address the unique design and analysis characteristics of CRXO trials. All CONSORT 2012 cluster trials extension items, proposed modifications, and other indicators of reporting quality are shown in Table 1.

Title and abstract (Items 1a, b)

The primary reasons for including a description of the trial design in the title and abstract are to ensure appropriate indexing in electronic databases [12, 14] and to alert the readers to the design so that they are less likely

Table 1 Quality of reporting of cluster randomised crossover trials as assessed against items from a modified 2012 CONSORT statement extension for cluster randomised trials and selected items from the 2010 CONSORT statement

Section	CONSORT Item no.	CONSORT 2012 extension for cluster trial design for Item no.	Reporting quality assessment measure	Reported? (*N* = 83)
Title and Abstract				
Identification of design in title	1a	Identification as a cluster randomised trial in the title	Identification as a CRXO trial in the title	7 (8%)
Reporting in abstract	1b	See Table 2 [14]	Identification as a CRXO trial in the abstract	21 (25%)
Background and objectives				
Rationale for design	2a	Rationale for using a cluster design	Rationale for using a cluster design AND a crossover of interventions at the cluster level	20 (24%)
Hypothesis and objectives	2b	Whether objectives pertain to the cluster level, the individual participant level or both	No modification proposed	Not assessed
Trial design				
Description of trial design	3a	Definition of cluster and description of how the design features apply to the clusters	Schematic representation of design (recommended especially for designs with >2 periods or interventions)	23 (28%)
			Definition of the cluster	77 (93%)
			Clear differentiation between cluster-period and cluster.	Not assessed
			Number of clusters	79 (95%)
			Number of periods	76 (92%)
			Duration of each time period or when the cross over will occur	Not assessed
			Cohort, repeated cross-sectional, or mixture of designs participants in each period	83 (100%)
			Discussion of the potential for carryover to occur	17 (20%)
			Reporting of use of washout period	83 (100%)
Participants				
Eligibility criteria	4a	Eligibility criteria for clusters	No modification proposed	Not assessed
Interventions				
Description of interventions	5	Whether interventions pertain to the cluster level, the individual participant level or both	No modification proposed	Not assessed
Outcomes				
Description of outcome measures	6a	Whether outcome measures pertain to the cluster level, the individual participant level or both	No modification proposed	Not assessed
Sample size	7a	Method of calculation, number of clusters(s) (and whether equal or unequal cluster sizes are assumed), cluster size, a coefficient of intracluster correlation (ICC or *k*), and an indication of its uncertainty	Was the method for sample size calculation reported, or justification for no sample size calculation provided?	48 (58%)
			Reference to the method used for the sample size calculation	Not assessed

Table 1 Quality of reporting of cluster randomised crossover trials as assessed against items from a modified 2012 CONSORT statement extension for cluster randomised trials and selected items from the 2010 CONSORT statement *(Continued)*

			Justification for number of clusters	33 (40%)
			Justification for number of periods	9 (11%)
			Equal or unequal number of periods per cluster	Not assessed
			Equal or unequal cluster-period sizes	42 (51%)
			A value for the within-cluster within-period ICC or variance components or other measure of correlations within data or justification for not including	13 (16%)
			A value for the within-cluster between-period ICC or variance components or other measure of correlations within data or justification for not including	4 (5%)
			A reference or explanation for the choice of ICCs or other measure of correlations	5 (6%)
			Reported whether the sample size methodology accounted for repeated measurements on the same individual	Not assessed
Sequence generation				
Method used to generate allocation sequence	8a	Method used to generate the random allocation sequence	No modification proposed	36 (43%)
Type of randomisation	8b	Details of stratification or matching if used	Does the article report whether stratified randomisation used?	83 (100%)
Allocation concealment mechanism				
Method used to implement the allocation sequence	9	Specification that allocation was based on clusters rather than individuals and whether allocation concealment (if any) was at the cluster level, the individual participant level, or both	Does the article report whether the people allocating the intervention sequence to the clusters know the allocation sequence?	40 (48%)
			Does the article report whether people recruiting/identifying participants knew which intervention sequence has been assigned to the cluster? ($n = 57$)[a]	44 (77%)
			Does the article report whether the people recruiting/identifying participants could have influenced which people were recruited/identified for inclusion in the study? ($n = 57$)[a]	54 (95%)
Implementation				
Method used to include clusters in trial	10a	Who generated the random allocation sequence, who enrolled clusters, and who assigned clusters to interventions	No modification proposed	Not assessed
Method used to include individuals in clusters	10b	Mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration, random sampling)	No modification proposed	Not assessed
Method of obtaining consent	10c	From whom consent was sought (representatives of the cluster, or individual cluster members, or both), and whether consent was sought before or after randomisation	From whom was consent sought?	60 (72%)
			Was consent sought before or after randomisation of the cluster when consent was sought from individuals? ($n = 30$)	16 (53%)

Table 1 Quality of reporting of cluster randomised crossover trials as assessed against items from a modified 2012 CONSORT statement extension for cluster randomised trials and selected items from the 2010 CONSORT statement *(Continued)*

Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how.	Were the participants aware of the intervention assigned to the cluster?	67 (81%)
			Were the researchers who delivered the intervention, i.e. caregiver, aware of the intervention assigned to the cluster?	82 (99%)
			If the outcome was self-reported ($n = 14$), was the participant aware of the intervention assigned to the cluster?	13 (93%)
			If the outcome was assessed by another person ($n = 69$), was the outcome assessor aware of the intervention assigned to the cluster?	45 (65%)
Statistical methods	12a	How clustering was taken into account	Justification for statistical analysis methods	Not assessed
			Reported whether the analysis was performed at the cluster or individual level.	78 (94%)
			Where there are more than two periods, reported whether a single correlation is assumed for the within-cluster between-period correlation	0 (0%)
			Was it possible to determine the method for accounting for both the cluster randomisation and multiple period aspects?	64 (77%)
			Was it possible to determine the method for accounting for the cluster randomisation aspect?	70 (84%)
			Was it possible to determine the method for accounting for the multiple period design aspect?	70 (84%)
Results				
Participant flow				
Number of clusters and participants	13a	For each group, the numbers of clusters that were randomly assigned, received intended treatment, and were analysed for the primary outcome	For each group, reported the number of clusters that were randomly assigned, received intended treatment in each period, and were analysed for the primary outcome	Not assessed
			For each group, reported the number of individuals that were randomly assigned, received the intended intervention in each period, and were analysed for the primary outcome	Not assessed
Losses and exclusions	13b	For each group, losses and exclusions for both clusters and individual cluster members	For each group, losses and exclusions for clusters, cluster-periods, and individual participants	Not assessed
Baseline data	15	Baseline characteristics for the individual and cluster levels as applicable for each group	Presentation of baseline characteristics data in table	
			No baseline characteristics table in article	24 (29%)
			Reported by total only	8 (10%)
			Reported by randomisation sequence with or without total	7 (8%)
			Reported by cluster only	2 (2%)
			Reported by intervention with or without total	37 (45%)
			Reported by cluster and period	2 (2%)

Table 1 Quality of reporting of cluster randomised crossover trials as assessed against items from a modified 2012 CONSORT statement extension for cluster randomised trials and selected items from the 2010 CONSORT statement *(Continued)*

			Reported by intervention and period	1 (1%)
			Reported by intervention, period, and cluster	2 (2%)
Number analysed	16	For each group, number of clusters included in each analysis	For each group, number of clusters, cluster-periods, and participants included in each analysis, stating reasons for exclusions	Not assessed
Outcomes and estimation	17a	Results at the individual or cluster level as applicable and a \coefficient of intracluster correlation (ICC or *k*) for each primary outcome	A coefficient for the within-cluster within-period correlation and within-cluster between-period correlation, or other measure (such as variance components), for each primary outcome	0 (0%)
Generalisability	21	Generalisability to clusters and/or individual participants (as relevant)	No modification proposed	Not assessed

[a] $n = 26$, no recruitment took place

to misinterpret the trial results [14]. A proposed modification is therefore to identify a trial as a "cluster randomised crossover trial" in title and abstract.

Background and objectives (Item 2a)

Providing a rationale for the trial design in the background informs the reader why the chosen design is best suited to address the research question. The cluster randomisation aspect of the CRXO design typically increases the required number of participants when compared to an individually randomised trial, potentially exposing more participants to harm than necessary if an individually randomised design was feasible [12]. In addition, both the crossover and cluster randomisation aspects of the CRXO design pose trial design, analysis, and implementation challenges. Hence, the choice to use the CRXO design in place of a simpler alternative such as a parallel group cluster randomised trial or individually randomised trial requires justification. Therefore, for a similar reason as proposed in the CONSORT 2012 cluster trials extension [12], we propose that the rationale for the use of cluster randomisation and for the crossover of interventions at the cluster level is included in the background.

Trial design (Item 3a)

Reporting the trial design allows the reader to replicate the design in future trials and assess whether the implemented sample size and analysis methods were appropriate for the design. We suggest that the following items might be considered important for clearly describing the design of a CRXO trial. Several of these items have been adapted from recommendations for reporting stepped-wedge trials [13]:

- Report the total number of randomised clusters in the trial.
- Report the total number of planned time periods for each cluster in the trial.
- Report the duration of each time period, for example, the duration of time or number of participants included in each cluster-period before the intervention is crossed over.
- Report whether the same, different, or a mix of same and different participants were included in each cluster-period. These designs are described as cohort, repeated cross-sectional, or mixture designs, respectively.
- For complex designs (i.e. designs with more than two interventions and two periods), consider including a schematic representation of the trial design depicting which interventions were allocated to each cluster in each period. For a simple design, the participant flow diagram (Item 13) may suffice.

We propose two new items for CRXO trials:

- Report the potential of the effect of the intervention given in one cluster-period to carry over to subsequent cluster-periods.
- Report methods for managing the risk of carryover, if necessary.

In addition to the above reporting items, we also suggest that articles clearly distinguish between the cluster and the cluster-period.

Sample size (Item 7a)

Reporting how a sample size calculation has been performed is important for replicability, transparency [12], and scientific and ethical reasons [15]. Reporting of sample size elements of cluster randomised trials has been shown to be incomplete [15]. The sample size calculation for CRXO trials should account for the predicted correlations arising from the design [5, 6]. In addition, the assumed sample size parameters and methodology should be reported. For CRXO trials, we suggest that the following sample size items might be considered important:

- Provide a reference for the sample size methodology or a description of the method when the method is not published.
- Report how the sample size methodology accounts for both the cluster randomisation (e.g. the within-cluster within-period ICC) and the multiple period aspects of the design (e.g. the within-cluster between-period ICC).
- Report how the sample size methodology accounts for whether the same, different, or a mix of the same and different participants will be included in each cluster-period.
- Report the number of clusters, number of periods, and number of participants per cluster-period, noting which are assumed and which are determined by the sample size calculation.
- Report whether a variable or constant number of periods per cluster and participants per cluster-period is assumed.
- Report the parameter values used to account for cluster randomisation and multiple periods.
- Provide a justification for the choice of parameter values and state any constraints on the number of clusters, number of periods, or number of participants per cluster-period.

Statistical methods (Item 12a)

The primary reasons for reporting the statistical methods are to allow for replication and for the reader

to evaluate whether the methods are appropriate for the design [11]. For CRXO trials we suggest that the following items might be considered:

- Provide a reference for the statistical methodology or a description of the method when the method is not published.
- Report whether the analysis was performed at the individual or cluster level.
- Report how both the cluster randomisation and the multiple period aspects of the design were accounted for.
- When there were more than two periods, report whether a constant within-cluster between-period ICC was assumed, and, if a constant within-cluster between-period ICC is not assumed, report what assumption or methodology was used.
- Describe how missing data will be managed at both the individual level and the cluster level [16].

Participant flow (Items 13a, b)

The CONSORT 2012 cluster trials extension [12] notes the importance of providing information on the flow of clusters through the trial (enrolment, allocation, follow-up, analysis) in addition to the flow of participants. A CRXO trial has the added complexity of cluster-periods nested within each cluster and potentially repeated measurements on some participants within each cluster. An additional consideration for reporting is the level at which the analysis is undertaken. The CONSORT 2012 cluster trials extension [12] recommends that if the analysis is aggregated at the cluster level, it is appropriate to show only the flow of clusters through the trial, while for analyses that do not use aggregated data, the flow of individual-level data should also be presented. However, we consider that it is important to show the flow of participants even when the analysis is aggregated at the cluster level, since aggregate-level analyses depend on the individual-level data. To facilitate discussion on presenting the flow of the number of clusters, cluster-period, and participants through a CRXO, we outline possible modifications to the flow diagram in the CONSORT 2012 cluster trials extension [12] and present a possible flow diagram in Fig. 1, although the exact form of the diagram is likely to depend on the trial.

Baseline data (Item 15)

The main reasons for reporting baseline characteristics are to describe the characteristics of the included population and permit an assessment of the success of the randomisation process. There is additional complexity in a CRXO design, because participants may be recruited to clusters over multiple time periods. Two key considerations when reporting participant and cluster characteristics in a CRXO trial are that (1) randomisation only ensures that, on average, cluster-level characteristics are balanced at baseline (assuming adequate sequence generation and allocation concealment), while individual-level characteristics may be influenced by selection processes; and (2) participants are often recruited at multiple time points, allowing presentation of cluster and individual characteristics at a single time point or summarised across multiple time points.

The cluster-level and individual-level characteristics can be either time invariant or time varying. For example, in a CRXO trial where hospital wards are randomised, the ward type (e.g. surgical, general medical) will remain constant for the duration of the trial. However, other cluster-level characteristics may vary, such as the type of clinicians working on the ward, due to staff changes (e.g. medical students and registrars moving in and out the ward). At the individual level, the characteristics of the individuals are likely to be time varying when new individuals are recruited across the periods (repeated cross-sectional design). However, if the trial is a cohort design, then individual-level characteristics such as sex will remain time invariant, while others, such as weight, may change over the duration of the trial.

We propose that the baseline characteristics are tabulated for each sequence and for each intervention within each sequence. A possible table for a two-period two-intervention CRXO trial is shown in Table 2. This table allows for a number of comparisons to be made for both time-invariant and time-varying characteristics. To facilitate discussion on presenting baseline characteristics in a CRXO trial, we outline a number potential comparisons, many of which have been used in published trials (Table 1), and exemplify these comparisons through Table 2.

Time-invariant characteristics

The time-invariant characteristics are described as follows:

1. Compare time-invariant characteristics of clusters allocated to sequence AB (Group 1 + Group 2) with clusters allocated to sequence BA (Group 3 + Group 4).
2. Compare time-invariant characteristics of participants recruited to sequence AB (Group 1 + Group 2) with participants recruited to sequence BA (Group 3 + Group 4) (cohort design).

Comparison 1 allows the success of the randomisation process to be evaluated. Comparison 2 allows for the process of recruiting participants into clusters to be evaluated. When the number of clusters is small, then chance imbalances between sequences may occur.

3. Compare time-invariant characteristics of clusters in all periods allocated to intervention A (Group 1 + Group

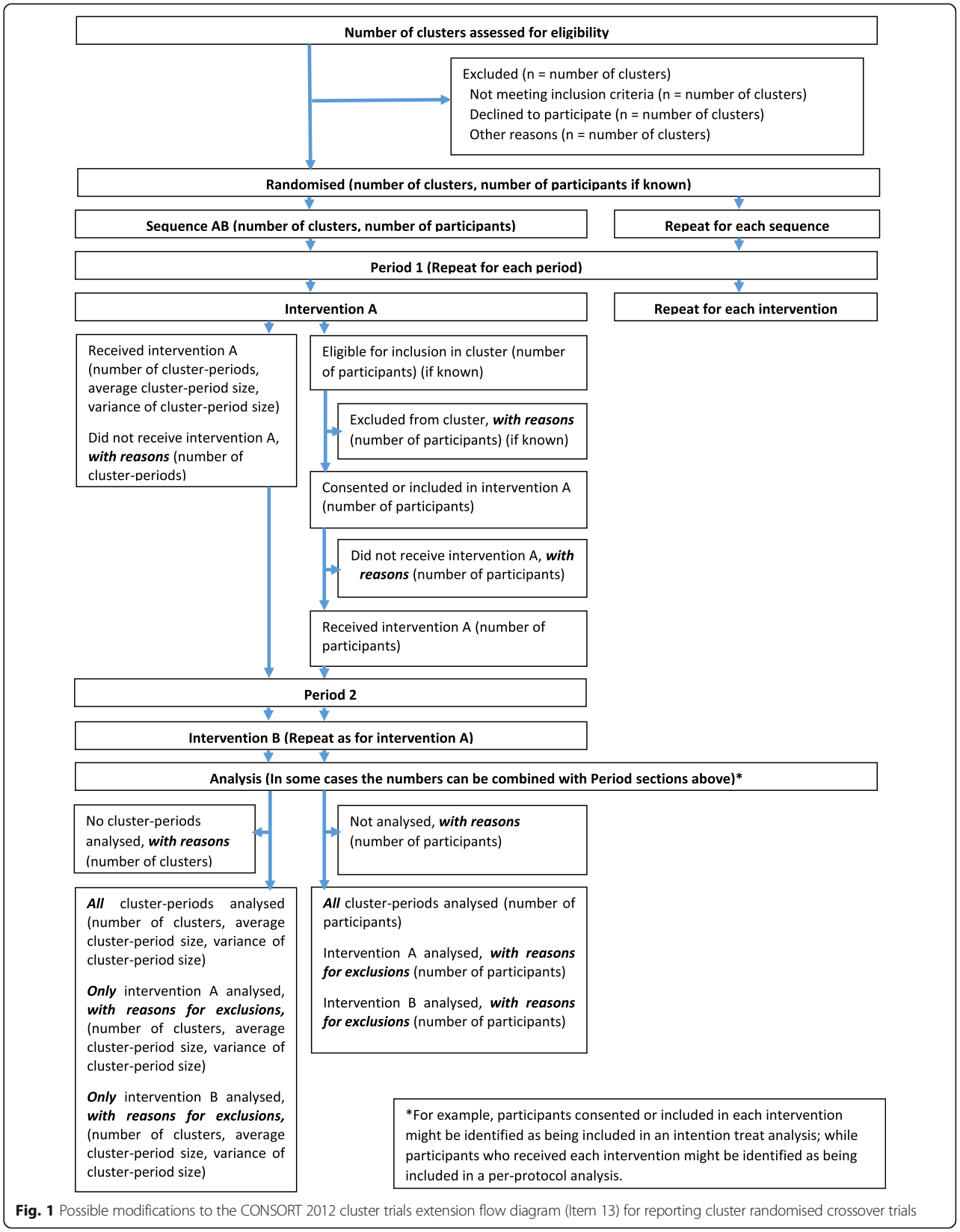


Fig. 1 Possible modifications to the CONSORT 2012 cluster trials extension flow diagram (Item 13) for reporting cluster randomised crossover trials

Table 2 Possible presentation of baseline characteristics in two-period two-intervention cluster randomised crossover trial

Characteristic	Intervention sequence AB		Intervention sequence BA	
	Period 1–Intervention A Group 1	Period 2–Intervention B Group 2	Period 1–Intervention B Group 3	Period 2–Intervention A Group 4
Time-invariant characteristics				
Time-invariant cluster characteristic	Such as proportion of each ward type: Cardiac: 25% Intensive Care: 40% Neurology: 35%	Such as proportion of each ward type: Cardiac: 25% Intensive Care: 40% Neurology: 35%	Such as proportion of each ward type: Cardiac: 35% Intensive Care: 45% Neurology: 20%	Such as proportion of each ward type: Cardiac: 35% Intensive Care: 45% Neurology: 20%
Time-invariant participant characteristic (cohort design only)	Such as patient sex: 59% male	Such as patient sex: 59% male	Such as patient sex: 48% male	Such as patient sex: 48% male
Time-varying characteristics				
Time-varying cluster characteristic	Nurse-to-patient ratio over 24 h, Median (IQR): 2.1 (2.0 – 2.2)	Nurse-to-patient ratio over 24 h, Median (IQR): 2.0 (1.9 – 2.1)	Nurse-to-patient ratio over 24 h, Median (IQR): 2.3 (2.1 – 2.4)	Nurse-to-patient ratio over 24 h, Median (IQR): 2.2 (2.1 – 2.4)
Time-varying participant characteristic (cohort and repeated cross-sectional design)	Such as patient weight (kg), Mean (SD): 83.4 (14.2)	Such as patient weight (kg), Mean (SD): 78.9 (15.6)	Such as patient weight (kg), Mean (SD): 81.2 (13.2)	Such as patient weight (kg), Mean (SD): 80.4 (11.2)

4) with clusters allocated to intervention B (Group 2 + Group 3).

4. Compare time-invariant characteristics of participants recruited to intervention A (Group 1 + Group 4) with participants recruited to intervention B (Group 2 + Group 3) (cohort design).

Comparisons 3 and 4 are equivalent to comparisons 1 and 2 when there is no loss of clusters or participants over time.

Time-varying characteristics

The time-varying characteristics are described as follows:

5. Compare time-varying characteristics of clusters allocated to sequence AB (Group 1 + Group 2) with clusters allocated to sequence BA (Group 3 + Group 4).
6. Compare time-varying characteristics of participants recruited to sequence AB (Group 1 + Group 2) with participants recruited to sequence BA (Group 3 + Group 4) (cohort and repeated cross-sectional design).

Comparisons 5 and 6 are likely to be of limited value. Presenting cluster-level characteristics summarised over multiple time periods can obscure whether randomisation was successful if systematic changes have occurred within the clusters. Likewise, presenting individual-level characteristics summarised over multiple time periods can obscure whether systematic changes have occurred in the recruitment of participants within the clusters.

7. Compare time-varying characteristics of clusters in all periods allocated to intervention A (Group 1 + Group 4) with clusters allocated to intervention B (Group 2 + Group 3).
8. Compare time-varying characteristics of participants recruited to intervention A (Group 1 + Group 4) with participants recruited to intervention B (Group 2 + Group 3) (cohort and repeated cross-sectional design).

As for comparisons 5 and 6, comparisons 7 and 8 also summarise cluster-level and individual-level characteristics over multiple time periods.

9. Compare time-varying characteristics of clusters allocated to intervention A with clusters allocated to intervention B, in the first period only (Group 1 vs Group 3).
10. Compare time-varying characteristics of participants recruited to intervention A with participants recruited to intervention B, in the first period only (Group 1 vs Group 3) (cohort and repeated cross-sectional design).

The considerations for comparisons 1 and 2 apply also to comparisons 9 and 10. However, comparisons 9 and 10 do not allow any evaluation of change in the characteristics over time and do not consider all participant data.

11. Compare characteristics of clusters allocated to intervention A with clusters allocated to intervention B, separately for each sequence (Group 1 vs Group 2 AND Group 3 vs Group 4).
12. Compare characteristics of participants recruited to intervention A with participants recruited to intervention B, separately for each sequence (Group

1 vs Group 2 AND Group 3 vs Group 4) (cohort and repeated cross-sectional design).
13. Compare characteristics of clusters allocated to intervention A with clusters allocated to intervention B, separately for each period (Group 1 vs Group 3 AND Group 2 vs Group 4).
14. Compare characteristics of participants recruited to intervention A with participants recruited to intervention B, separately for each period (Group 1 vs Group 3 AND Group 2 vs Group 4) (ohort and repeated cross-sectional design).

Presentation of cluster-level and individual-level characteristics separately by period in each intervention (comparisons 11–14) allows for assessment for any systematic change in characteristics over time or any potential interaction between intervention and time. Such changes will be obscured by presenting characteristics summarised over multiple time periods.

Number analysed (Item 16)

Reporting the number of clusters, cluster-periods, and participants that contribute to each analysis of each outcome is essential to interpreting the results. To facilitate discussion on presenting the numbers analysed in CRXO trials, we outline a potential approach:

- Present the number of clusters, cluster-periods, and participants analysed for the primary outcome as per the participant flow diagram (Fig. 1).

In addition, for each secondary analysis and outcome, either state that the same clusters, cluster-periods, and individuals are included as in the primary analysis, or where the number analysed differs from Fig. 1:

- Report the number of clusters that contribute to the analysis across all periods, also separately by intervention sequence, and give reasons for the exclusion of any whole clusters.
- Report the number of clusters that contribute to the analysis for only some periods, also separately by intervention and intervention sequence. Give reasons for the exclusion of any cluster-periods and state whether the remaining clusters-periods from that cluster were included.
- Report the number of participants included in the analysis, by intervention and intervention sequence, including the reasons for any exclusions at the individual level.

Outcomes and estimation (Item 17a)

The importance of providing estimates of within-cluster correlation in cluster randomised trials for the purpose of describing the clustering and future sample size estimation is well recognised [12]. For similar reasons, in CRXO trials it is important to provide estimates of within-cluster between-period ICCs, in addition to the within-cluster within-period ICCs for each outcome. Alternatively, if mixed models are used, the reporting of variance components can be provided from which the ICCs can be calculated.

Methods

The protocol for the review has been published [17]. Here we provide only a brief overview of the methods, along with deviations from the planned methods, and outline the measures used to assess reporting quality.

Literature search

In brief, MEDLINE, PubMed, EMBASE, and CINAHL Plus were searched until December 2014 for English language articles of CRXO trials. In addition to searching for CRXO trials, we searched PubMed for CRXO methodology articles to identify further references to CRXO trials. A citation search of all identified methodology articles was performed in Web of Science. Finally, the references of all eligible articles were screened for CRXO trials. No restriction was applied to the publication date. The search strategies for CRXO trials and CRXO methodology articles are outlined in Arnup et al. [17] and provided in Additional file 1.

Trial inclusion criteria

Trials that met the following inclusion criteria were included in the review: the trial was undertaken in humans; the allocation of the intervention was to clusters of individuals rather than individuals themselves; each cluster received each intervention in a sequence over time (conventional crossover design), or at least some clusters crossed over from one intervention to another (such as two-treatment-four-sequence designs AA, AB, BA, and BB); at least some clusters crossed each way between at least two interventions (e.g. one cluster received AB and one cluster received BA and therefore excludes pre-post designs); and the intervention given in the one period was not deliberately intended by design to affect individuals in subsequent periods (e.g. interventions intended to change the prescribing behaviour of health care providers). The latter two criteria were added while undertaking the review. Protocols were included in the review; however, for this article the focus is on the quality of reporting of trial reports, and hence protocols have been excluded.

Selection of trials for inclusion in the review

One author (SA) assessed all titles and abstracts using the eligibility criteria, and 50% of the titles and abstracts were screened independently by at least one co-author.

All full-text articles were then assessed by one author (SA) using the eligibility criteria. Of these, all eligible articles were double screened, along with 20% of articles that were initially determined to be ineligible. Differences in inclusion decisions were resolved by discussion or by referral to a third author. No ineligible articles were subsequently found to be eligible.

Data extraction and management

One author (SA) extracted data from all trials, and data from 20% of the trials were independently double data extracted by the co-authors. Three of the five authors (SA, JM, AF) reviewed the discrepancies arising from the double data extraction and discussed processes for further reviewing items where there was inconsistency. The processes and items where further review was undertaken are described in Arnup et. al. [17]. The data extraction form was piloted on five trials by each author. Data were entered into a database (Microsoft Access 2010, Redmond, Washington, USA).

To examine the reporting quality of the CRXO trials, we extracted reported information from the trials on selected 2012 cluster trial CONSORT extension items [12], with modification so that they were suitable to assess CRXO trials (Items 1a,b and Item 2a of the 2012 cluster trials extension [12]; hereafter we only refer to the item number). Where the CONSORT extension may not have adequately covered the unique characteristics of CRXO trials, we extracted information on indicators of the reporting quality for that item (Item 3a, Item 7a, Item 8b, Item 9, Item 10c, Item 12a, Item 15, Item 17a). We refer to these measures as indicators because further discussion between trialists using the CRXO design is required to determine if the measure adequately assesses reporting quality.

We did not extract information on CONSORT items where the reporting considerations did not differ from a parallel group cluster trial or individually randomised trial, e.g. description of the interventions and outcome measures (Item 2b, Item 3b, Items 4a, b, Item 5, Items 6a, b, Item 7b, Items 10a, b, Item 11b, Item 12b, Items 14a, b, Item 17b, Items 18–25.). However, there were two exceptions where we did extract information on the following 2010 CONSORT items [11]: Item 8a "Method used to generate random allocation sequence" because the item is required to evaluate the potential for selection bias, and Item 11a "Who was blinded after assignment to intervention?" because the item was required to evaluate the risk of detection bias (see the next paragraph). For Item 13 "Participant flow" and Item 16 "Number analysed", we present a discussion of possible reporting approaches only.

In addition to assessing the quality of reporting of CRXO trials against reporting items and indicators, we assessed whether the reported information was sufficient to judge the risk of selection, performance, and detection bias. The *Cochrane Handbook for Systematic Reviews of Interventions*, Section 8.4, defines selection bias as systematic differences between baseline characteristics of the groups that are compared; performance bias as systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest; and detection bias as systematic differences between groups in how outcomes are determined. (See Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from http://handbook.cochrane.org/.)

The full list of extracted data is available in Arnup et al. [17]. The extracted data specific to reporting quality in CRXO trials were: identification of the design in the title or abstract; justification for using the design; selected design characteristics including schematic representation of the design, definition of the cluster, number of clusters, number of periods, type of design (cohort, repeated cross-sectional, or mixture of designs), and management of the risk of carryover of intervention effects between periods; reporting of sample size calculation details including justification for the number of clusters, justification for the number of periods, equal or unequal cluster-period sizes, the assumed measure of similarity between the outcomes of individuals within a cluster within a given period and justification for assumption, and the reported measure of similarity between outcomes of individuals within a cluster between different periods and justification for assumption; methods used in the trial including recruitment, consent, randomisation, allocation, and blinding; statistical analysis including level of analysis, the method accounting for the similarity between the outcomes of individuals within a cluster within a given period and the similarity between outcomes of individuals within a cluster between different periods, and the reported measure of similarity between outcomes of individuals within a cluster between different periods; losses and exclusions of participants; and reporting of baseline characteristics.

Data coding

In the section, we provide details on how we judged each reporting quality measure.

We classified the following items as reported if a clear statement addressing the item was provided in the trial report: Items 1a, b, Item 2a, Item 10c, Item 12a, Item 15, and Item 17a.

For items that were not explicitly reported in the trial report, we reviewed the reported methods to determine whether enough information was provided to classify the following items as reported: same, different, or a mix of

participants in each period (Item 3a); equal or unequal cluster sizes in the sample size calculation (Item 7a); use of restricted randomisation (Item 8b); items addressing allocation concealment (Item 9); and blinding (Item 11) methods.

We classified the method of random allocation (Item 8a) as "reported" if the article included details on how the random allocation was achieved or clearly stated that the allocation was not random. The method of random allocation in articles that stated that the allocation was "at random", with no further detail, was classified as "not reported".

We classified the information reported in each trial as either sufficient or insufficient to assess the risk of selection bias, performance bias, and detection bias. To assess whether sufficient information was reported to judge the risk of selection bias, we required that enough detail was reported to assess (1) whether the researcher allocating the cluster to the intervention sequence was blind to future allocation assignments (Item 9); (2) whether the people recruiting or identifying participants knew which intervention sequence had been assigned to the cluster (Item 9); (3) whether the researcher recruiting or identifying participants could influence which individuals were included in the trial (Item 9); and if the individual participant (or other person on their behalf) provided consent, whether they had knowledge, prior to consenting, of the intervention assigned to the cluster (Item 10c). To assess whether sufficient information was reported to judge the risk of performance bias, we required that enough detail was reported to assess if either the participants or those delivering the intervention (e.g. health care professionals) were aware of the intervention (Item 11a). To assess whether sufficient information was reported to judge the risk of detection bias, we required that enough detail was reported to determine if the assessment of outcomes was self-reported or measured by another person and whether the assessor was aware of the intervention assigned to the cluster (Item 11a).

Data analysis

We present descriptive summary statistics using frequencies and percentages of responses to categorical data. Free text was classified and frequencies and percentages of the categories are presented. The extracted data from individual trials can be made available upon request to the corresponding author.

Results

Results of the search

Figure 2 shows the flow diagram of the CRXO trial selection process for the systematic review. Of the 3425 records identified through database searching, 170 were duplicates and 3046 were ineligible based on screening of abstracts, leaving 209 full-text articles to assess for eligibility. Of these 209 articles, 99 were assessed as eligible. A further four articles were identified through the methodology article reference and citation search, and three articles from the references of eligible articles. In this article we further exclude eight trials where only a protocol was available. In total, 98 articles from 83 trials were included in this paper. Seventy-one trials had only one associated article, nine trials had two associated articles, and three trials had three associated articles.

Characteristics of the trials

Most trials were conducted in a developed country ($n = 79$, 95%) and were undertaken within a single country ($n = 80$, 96%) (Additional file 2: Table S1). The types of clusters varied, with almost half of the trials ($n = 40$, 48%) randomising hospitals or wards within hospitals, 13 (16%) randomising individual health care providers, and 11 (13%) randomising schools or classes (Additional file 3: Table S2).

The trials investigated a wide range of diseases and conditions and health care delivery models. Nineteen trials (23%) investigated infection control, ten (12%) investigated cardiovascular disease, nine (11%) examined the delivery of health services, and eight (10%) investigated infectious diseases (Table 3).

The most common trial design involved two interventions ($n = 74$, 89%). The majority of trials (70%) used two periods. Trials had a median of eight clusters (IQR: 3 – 21, range: 2– 268) and a median cluster-period size of 27 (IQR: 14–77, range: 2–1319) (Table 2).

In 42 trials (51%) the interventions were delivered directly to the individuals within the clusters. In 21 trials (25%) the intervention was targeted at the health care provider rather than the individuals under their care, and in 14 trials (17%) the intervention was targeted at the organisation of the health care provider or health services delivery (Table 2).

Quality of reporting in CRXO trials as assessed against proposed or modified reporting items and other indicators

Trials were infrequently identified as "Cluster Randomised Crossover" trials in the title (8%) or in the abstract (25%). A rationale for both the cluster randomisation and crossover aspects of the design was provided in 20 trials (24%). Most design characteristics were reported; however, trials infrequently used a schematic to illustrate the design (28%), even in designs with either more than two periods or more than two interventions (30%, $n = 7/23$) (Table 1).

The reporting of the methods used to generate the allocation sequence and assign the allocation sequence to clusters was incomplete in 43% ($n = 36$) and 48% ($n = 40$) of trials, respectively. In 20% ($n = 17$) of the trials, the risk of a carryover of the intervention effect from one period to subsequent periods was discussed (Table 1).

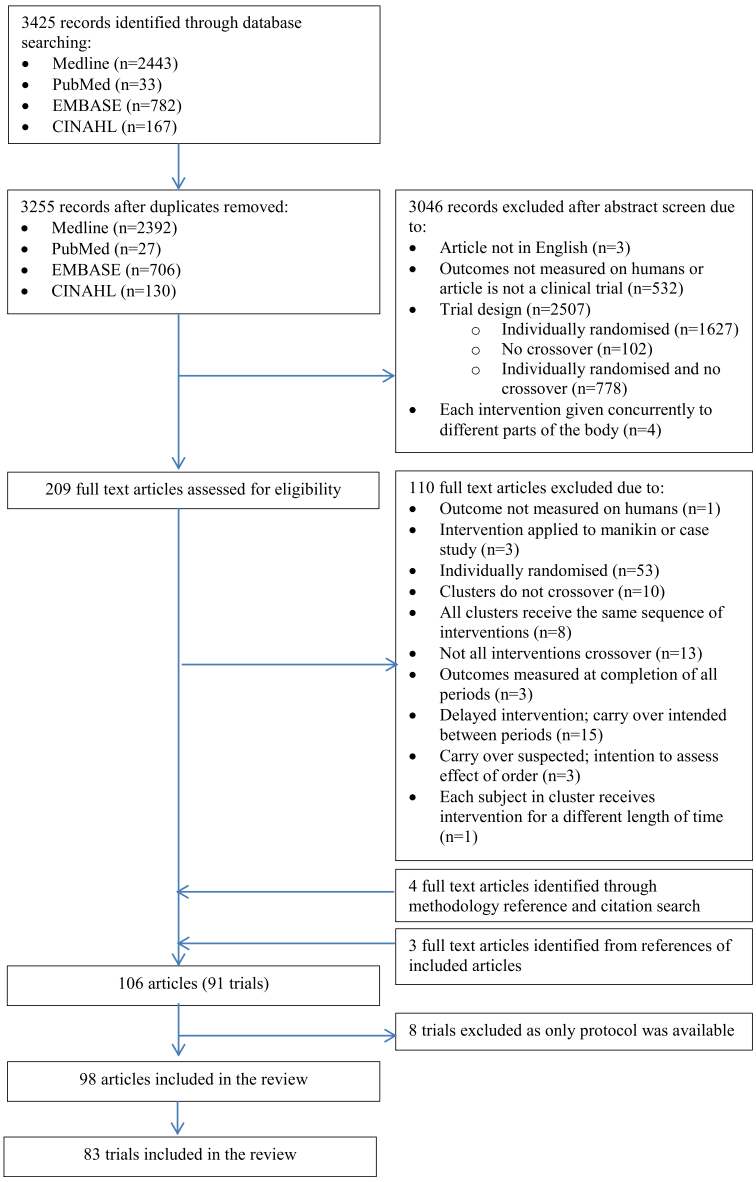


Fig. 2 Flow of articles through the systematic review

Reporting of the methods and parameters to calculate the sample size was often missing or incomplete. Only 48 trials (58%) provided a sample size calculation or justification for not performing a sample size calculation. Thirty-three trials (40%) provided justification for the number of clusters, and only nine trials (11%) provided justification for the number of periods. Of the 83 trials, only 13 (16%) reported the within-cluster within-period ICC, and only four (5%) reported the within-cluster between-period ICC (or corresponding variance components) that was assumed in the sample size calculation (Table 1).

The construction of the baseline characteristics tables of the individual participants varied. In most trials, the characteristics were reported by intervention group (45%); some trials reported by intervention sequence (8%). In 24 trials (29%), no baseline characteristics table was presented (Table 1).

Most trials gave sufficient detail to determine whether the analysis was performed at the level of the individual or the level of the cluster ($n = 78$, 94%). However, in 19 trials (23%) it could not be determined how or whether the analysis accounted for the cluster randomisation or multiple period aspects of the design. No trial reported a measure of both intracluster correlations or variance components induced by the cluster randomisation and multiple period aspects of the

Table 3 Characteristics of the cluster randomised crossover trials

Disease or domain under study	*n* (%) (*N* = 83)
Infection control	19 (23%)
Cardiovascular disease	10 (12%)
Health services delivery	9 (11%)
Infectious disease	8 (10%)
General and public health	5 (6%)
Medical training	5 (6%)
Communication of health information	4 (5%)
Pregnancy, childbirth, and early childhood	3 (4%)
Mental health and behavioural conditions	3 (4%)
Respiratory disease	3 (4%)
Blood sample contamination	3 (4%)
Cognition	3 (4%)
Central nervous system and musculoskeletal disease	2 (2%)
Oral health	2 (2%)
Nutritional and metabolic disorders	1 (1%)
Urogenital disease	1 (1%)
Digestive disorders	1 (1%)
Pain management	1 (1%)
Type of intervention	
Intervention targeting the individual	42 (51%)
Intervention targeting health care provider	21 (25%)
Quality improvement intervention	14 (17%)
Intervention resulting in change to the participant environment	6 (7%)
Number of interventions	
2	74 (89%)
3	8 (10%)
4	1 (1%)
Number of clusters - Median [IQR]; Range	8 [3–21]; 2 – 268
Unclear	4 (5%)
Number of periods[a]	
2	53 (70%)
3	8 (11%)
4+	15 (18%)
Unclear	7 (8%)
Cluster-period size - Median [IQR]; Range	27 [14–77]; 2 – 1319
Unclear	17 (20%)

[a]Percentages of non-missing data presented

design (Table 1). One trial reported the within-cluster ICC from an analysis that included only a random effect for cluster, therefore assuming that the within-cluster between-period ICC was equal to the within-cluster within-period ICC.

Quality of reporting in CRXO trials to allow assessment of bias

Selection bias

Twenty-five trials (30%) provided sufficient information to assess the risk of selection bias (Table 4). In 43 trials (52%) we were unable to judge the risk of selection bias because we could not determine whether the researchers responsible for allocating the intervention sequence to the clusters were aware, or not, of the intervention sequence; in 13 trials (17%) we were unable to judge the risk of selection bias because it was not clear whether the researchers recruiting participants were aware, or not, of the cluster's intervention sequence; in three trials (4%) we were unable to judge the risk of selection bias because we could not judge whether the researchers responsible for recruiting/identifying participants were able to influence recruitment; and in 14 of the 30 trials (47%) where individual consent was sought, we could not assess the risk of selection bias because we could not judge whether the participant was aware of the intervention assigned to the cluster prior to giving consent.

Performance bias

Sixty-seven trials (81%) provided sufficient information to assess performance bias (Table 4). In the 16 trials (19%) that did not provide sufficient detail to assess the risk of performance bias, we could not judge whether the intervention was concealed, or not, from the participants. In one trial (1%) we also could not judge whether the intervention was concealed, or not, at cluster level.

Detection bias

Fifty-eight trials (70%) provided sufficient information to assess detection bias (Table 4). Of the 14 trials (17%) in which the primary outcome was self-report, we could not judge if the participant was aware, or not, of the intervention assigned to the cluster-period in one trial (7%). Of the 69 trials (83%) where the primary outcome was not self-report, we could not judge if the assessor was aware, or not, of the intervention in 24 trials (35%).

Discussion

We proposed possible reporting items for CRXO trials as a basis for further discussion and to examine reporting quality. The items were either modified from those

Table 4 Quality of reporting in CRXO trials of material required to assess selection, performance, and detection biases

Bias	Sufficient information to evaluate risk of bias (*N* = 83)
Selection bias	25 (30%)
Performance bias	67 (81%)
Detection bias	58 (70%)

in the 2010 CONSORT [11] and 2012 cluster trial extension statements [12] or were proposed reporting indicators. Incomplete reporting of the design aspects that are unique to the CRXO design was found to be common in the published trials included in the systematic review.

The frequency of reporting of sample size calculations was similar in CRXO trials compared with other randomised trial designs, including individually randomised trials, parallel group cluster randomised trials, individual crossover trials, and stepped - wedge trials [8, 15, 18–21]. Reporting of the ICCs assumed in the sample size calculation was poorer in CRXO trials compared with parallel group cluster randomised trials, with the within-cluster within-period ICC and within-cluster between-period ICC assumed in the sample size calculation only reported in 5% of CRXO trials compared with 35% in parallel group cluster randomised trials [15]. Furthermore, no CRXO trial reported both the ICC observed in the analysis and the ICC assumed in the sample size calculation, compared with 11% of parallel group cluster randomised trials [15].

The completeness of reporting risk of bias domains for CRXO trials was better than previously observed estimates for the domains: method of random sequence generation, method of allocation concealment, and blinding [8, 18, 19, 21]. This more complete reporting may reflect our generous assessment of complete reporting for these domains, or the greater number of trials in this review that were published after the publication of the 2010 CONSORT statement [11] and 2012 cluster trials extension [12]. However, these domains were still incompletely reported in around half of CRXO trials.

Complete reporting of individually randomised crossover trials allows identification of the potential for carryover and of the methods used to manage potential carryover, including the use of washout periods. While we were able to judge if a washout period had been used in all CRXO trials, discussion of the potential for carryover only occurred in 20% of trials included in this systematic review. This estimate was similar to that observed in a study examining the reporting of individually randomised crossover trials (29%) [19]. Previous estimates for the reporting of the use of a washout period include 70% [19] and 99% [20].

Assessing the quality of reporting of published CRXO trials is a recommended initial step in developing reporting guidelines [9]. This should be undertaken in combination with reviewing relevant existing guidelines to determine whether it is most appropriate to amend an existing guideline or develop a new guideline. The CRXO design has unique features, and reporting guidance for these features is currently not addressed by items in existing guidelines [11–13]. Therefore, it was necessary to concurrently propose reporting items and assess the quality of reporting against these items.

The results of the present study suggest a need for improved reporting of CRXO trials, and given the lack of specific guidance for this design, a CONSORT extension would be of value. Recommended next steps would include setting up a consensus process, including participants with relevant expertise, to decide upon the specific items and their wording [9]. However, in the absence of specific guidance for this design, our suggested modifications may usefully inform reporting of CRXO trials until formal guidelines are developed.

Strengths and limitations

Our review represents the most comprehensive review of this trial design to date, despite some potential limitations in the methods used to locate CRXO trials, which have been previously discussed [1]. In brief, it may be argued that better reported trials are easier to locate, and thus, our results may present an optimistic view of the reporting quality in CRXO trials.

Our conclusions of the reporting quality in CRXO trials may also depend on our chosen reporting quality measures. However, our reporting quality items were predefined, and were based on items modified from the 2010 CONSORT statement [11] and 2012 cluster trials extension [12] wherever possible. However, the next step would be to undertake a more rigorous process to refine and agree upon the reporting items using a consensus process such as the Delphi method.

Conclusions

We have proposed possible reporting items for CRXO trials as a basis for further discussion and to examine reporting quality. We found that incomplete reporting of the design aspects that are unique to the CRXO design was common in the published trials included in this systematic review. Given these results, it is important that a CONSORT extension is developed. Consensus amongst trialists on the content of such a guideline is essential.

Additional files

Additional file 1: Search strategies. Appendix containing the search strategies to identify CRXO trials. (DOCX 12 kb)

Additional file 2: Table S1. Country where the trial was conducted. Table containing additional demographic data of the included trials in the review. (DOCX 12 kb)

Additional file 3: Table S2. Type of randomised cluster. Table containing additional demographic data of the included trials in the review. (DOCX 12 kb)

Abbreviations

CONSORT: CONsolidated Standards Of Reporting Trials; CRXO: Cluster randomised crossover; ICC: Intracluster correlation; IQR: Interquartile range; SD: Standard deviation

Acknowledgements
We acknowledge Ross Dunn for his assistance in the development of the Access database and Steve McDonald for his assistance in the development of the search strategies.

Funding
SA was supported in part by a Monash University Graduate Scholarship and a National Health and Medical Research Council of Australia Centre of Research Excellence grant (1035261) to the Victorian Centre for Biostatistics (ViCBiostat). Funding was provided to KM through a National Institute for Health Research (NIHR) research methods fellowship (MET-12-16). JM was supported by a National Health and Medical Research Council (NHMRC) Australian Public Health Fellowship (1072366).

Availability of data and materials
The datasets created and analysed during the current study are available from the corresponding author on request.

Authors' contributions
SA conceived the study in collaboration with AF and JM; led the design and conduct of the study, including performing searching, screening, data extraction, and statistical analyses; and led the drafting of the manuscript. AF and JM contributed to the conception and design of the study and drafting of the manuscript, screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript. BK and KM screened studies for inclusion in the review, extracted data, and provided critical review of the manuscript. All authors approved the final manuscript.

Authors' information
This section is not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
This section is not applicable.

Ethics approval and consent to participate
This section is not applicable.

Author details
[1]School of Public Health and Preventive Medicine, Monash University, The Alfred Centre, Melbourne, Victoria 3004, Australia. [2]Pragmatic Clinical Trials Unit, Queen Mary University of London, 58 Turner St, London E1 2AB, UK. [3]Medical Statistics Department, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Received: 23 March 2016 Accepted: 4 November 2016
Published online: 06 December 2016

References

1. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. J Clin Epidemiol. 2016;74:40–50.
2. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. Stat Med. 2007;26(2):274–89.
3. Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. Contemp Clin Trials. 2007;28(3):316–23.
4. Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. J Educ Behav Stat. 2011;36(4):472–90.
5. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Stat Med. 2008;27(27):5578–85.
6. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. Clin Trials. 2015;12(1):34–44.
7. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. Oxford: The Cochrane Collaboration; 2011. Available from http://handbook.cochrane.org/.
8. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ. 2010;340:c723.
9. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010;7(2), e1000217.
10. Moher D, Schulz KF, Altman D, Group C. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285(15):1987–91.
11. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):834–40.
12. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345, e5661.
13. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ. 2015;350:h391.
14. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Schulz KF, Group C. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med. 2008;5(1), e20.
15. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. J Clin Epidemiol. 2015;68(6):716–23.
16. Giraudeau B, Ravaud P. Preventing bias in cluster randomised trials. PLoS Med. 2009;6(5), e1000065.
17. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McDonald S, McKenzie JE. The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review. Syst Rev. 2014;3:86.
18. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011;343:d5886.
19. Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. Trials. 2009;10:27.
20. Li T, Yu T, Hawkins BS, Dickersin K. Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis. PLoS One. 2015;10(8), e0133023.
21. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol. 2011;64(9):936–48.

Chapter 5

Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial

The results of the systematic review, presented in Chapter 3, showed that health researchers infrequently use appropriate methods for the sample size calculation and analysis of outcome data in CRXO trials. Possible explanations for the use of inappropriate sample size methodology are 1) limited understanding amongst health researchers of the effect of the CRXO design aspects on sample size calculation, 2) lack of recognition of the availability of sample size formulae, 3) lack of availability of the parameters required by the formulae, and 4) lack of appropriate sample size methodology.

The aim of the research presented in this chapter is to address the first three explanations by providing statisticians and health researchers with an understanding of the methodological requirements of the CRXO design, by providing 1) graphical illustration of the effect of the cluster randomisation and multiple period aspects of the design on the correlation between individual responses in a CRXO trial, 2) worked examples of sample size calculation in the intensive care setting using previously published sample size formulae, and 3) guidance on how to select the parameters required by the sample size formulae.

Chapter 5 is presented as a manuscript, which was published in *Trials* in August 2017. The pages have been renumbered for the thesis, but the manuscript is otherwise unchanged. In this paper, the terminology used for "intracluster correlation" is "correlation".

Supplementary tables and additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix E	Additional file 1	Continous outcomes sample size Stata do file
	Additional file 2	Binary outcomes sample size Stata do file

Arnup et al. Trials (2017) 18:381
DOI 10.1186/s13063-017-2113-2

Trials

METHODOLOGY Open Access

Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial

Sarah J. Arnup[1], Joanne E. McKenzie[1], Karla Hemming[2], David Pilcher[3,4,5] and Andrew B. Forbes[1*]

Abstract

Background: In a cluster randomised crossover (CRXO) design, a sequence of interventions is assigned to a group, or 'cluster' of individuals. Each cluster receives each intervention in a separate period of time, forming 'cluster-periods'. Sample size calculations for CRXO trials need to account for both the cluster randomisation and crossover aspects of the design. Formulae are available for the two-period, two-intervention, cross-sectional CRXO design, however implementation of these formulae is known to be suboptimal. The aims of this tutorial are to illustrate the intuition behind the design; and provide guidance on performing sample size calculations.

Methods: Graphical illustrations are used to describe the effect of the cluster randomisation and crossover aspects of the design on the correlation between individual responses in a CRXO trial. Sample size calculations for binary and continuous outcomes are illustrated using parameters estimated from the Australia and New Zealand Intensive Care Society – Adult Patient Database (ANZICS-APD) for patient mortality and length(s) of stay (LOS).

Results: The similarity between individual responses in a CRXO trial can be understood in terms of three components of variation: variation in cluster mean response; variation in the cluster-period mean response; and variation between individual responses within a cluster-period; or equivalently in terms of the correlation between individual responses in the same cluster-period (within-cluster within-period correlation, WPC), and between individual responses in the same cluster, but in different periods (within-cluster between-period correlation, BPC).
The BPC lies between zero and the WPC. When the WPC and BPC are equal the precision gained by crossover aspect of the CRXO design equals the precision lost by cluster randomisation. When the BPC is zero there is no advantage in a CRXO over a parallel-group cluster randomised trial. Sample size calculations illustrate that small changes in the specification of the WPC or BPC can increase the required number of clusters.

Conclusions: By illustrating how the parameters required for sample size calculations arise from the CRXO design and by providing guidance on both how to choose values for the parameters and perform the sample size calculations, the implementation of the sample size formulae for CRXO trials may improve.

Keywords: Cluster randomised, Crossover, Sample size, Intracluster correlation, Within-period correlation, Between-period correlation, Components of variability

* Correspondence: Andrew.Forbes@monash.edu
[1]School of Public Health and Preventive Medicine, Monash University, The Alfred Centre, Melbourne, VIC 3004, Australia
Full list of author information is available at the end of the article

Background

Individually randomised trials are considered the 'gold standard' for evaluating medical interventions [1]. However, situations arise where is it necessary, or preferable, to randomise clusters of individuals, such as hospitals or schools, rather than the individual patients or students, to interventions [2, 3]. A cluster randomised trial will generally require a larger sample size compared with an individually randomised trial to estimate the intervention effect to the same precision [4].

In a two-period, two-intervention, cluster randomised crossover (CRXO) design, each cluster receives each of the two interventions in a separate period of time, leading to the formation of two 'cluster-periods'. In a cross-sectional design, each cluster-period consists of different individuals, while in a cohort design, each cluster-period consists of the same individuals. The order in which the interventions are delivered to each cluster is randomised to control for potential period effects [5, 6]. Like in an individually randomised trial, this adaption has the benefit of reducing the required number of participants [7]. The key to understanding the CRXO design is to recognise how both the cluster randomisation and crossover aspects of the design lead to variation between individual responses in a trial; and how these aspects of the design give rise to similarities in the responses of groups of individuals.

Sample size formula have been published for the two-period, two-intervention, cross-sectional CRXO design [8–10]. These formulae require a-priori specification of two correlations: the similarity between two individuals in the same cluster-period, typically measured by the within-cluster within-period correlation (WPC); and the similarity between two individuals in the same cluster, but in different cluster-periods, typically measured by the within-cluster between-period correlation (BPC). However, there is little guidance for informing the value of the BPC, nor on the sensitivity of the sample size to the chosen values of both correlations [11, 12].

A 2015 systematic review of CRXO trials found that both the cluster randomisation and crossover aspects of the design of the CRXO was appropriately accounted for in only 10% of sample size calculations and 10% of analyses [13]. This suggests that the CRXO design is not well understood.

The aims of this tutorial are to illustrate the intuition behind the CRXO design; to provide guidance on how to a-priori specify the WPC and BPC; and perform sample size calculations for two-period, two-intervention, cross-sectional CRXO trials.

In the 'Understanding the CRXO design' section, we describe how the cluster randomisation and crossover aspects of the design leads to variation between individual responses in a two-period, two-intervention, cross-sectional CRXO design, using intensive care unit (ICU) length(s) of stay (LOS) as an example. In the 'Performing a sample size calculation' section, we outline how to perform sample size calculations and discuss how to specify values of the WPC and BPC for sample size calculations. In the 'Common mistakes when performing a sample size analyses' section, we outline common mistakes made by trialists when performing sample size calculations for CRXO trials and the likely consequences of those mistakes. We conclude with a general discussion, considering extensions and larger designs.

Understanding the CRXO design

In this section we illustrate graphically how the cluster randomisation and crossover aspects of the CRXO design leads to variation in the responses of individuals in a CRXO trial, and how these aspects of the design can be used to measure the similarity between individuals using the WPC and BPC.

We illustrate the sources of variation and measures of similarity that arise in the two-period, two-intervention, cross-sectional CRXO design by considering a hypothetical CRXO trial conducted in 20 ICUs over a 2-year period. We consider the ICU LOS of all patients admitted to these 20 ICUs, and assume (for ease of exposition) that the number of patients in each ICU is infinitely large (or at least very large). As LOS is non-normally distributed and right skewed, we use the logarithmic transform of ICU LOS throughout our illustration.

Each ICU is randomly assigned to administer one of two interventions to all patients admitted during the first year (period 1). In the subsequent year, each ICU administers the alternate intervention (period 2). All patients admitted to a single ICU over the 2-year period can be thought of as belonging to a *cluster*. Within each ICU (cluster), the patients admitted during a 1-year period can be thought of as belonging to a separate *cluster-period*. Therefore, in each ICU (cluster) there are two cluster-periods.

The allocation of interventions to patients in the stratified, multicentre, parallel-group, individually randomised trial (IRCT) design, the parallel-group cluster randomised trial (CRCT) design, and the CRXO design are shown in Fig. 1. In each design, each intervention is given for one 12-month period. In the IRCT design half the patients in each centre (ICU) receive each intervention. In the CRCT design, all patients in a single ICU are assigned the same intervention.

Variation in the length of stay between patients

To illustrate the sources of variation and measures of similarity that arise in the CRXO design, we assume that the true difference between interventions is zero. In the hypothetical situation where we have an infinite number

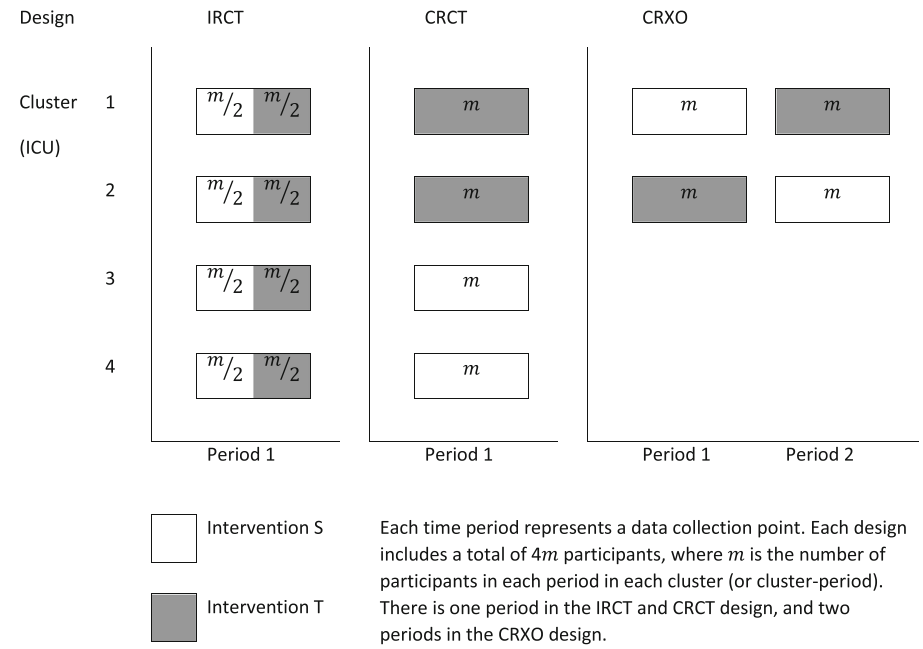


Fig. 1 Schematic illustration of the stratified, multicentre, parallel-group, individually randomised trial (IRCT), parallel-group cluster randomised trial (CRCT), and cluster randomised crossover (CRXO) design with the same total number of participants

of patients, the overall mean LOS for all patients in the trial will be equal to the true overall mean LOS for all patients who could be admitted to the 20 ICUs. The variation in LOS arises from both patient and ICU factors. In a CRXO design, the ICU (cluster) and the time period of admission (cluster-period) are both factors that could affect the patient's LOS and, therefore, explain some of the variation seen in patient LOS. For example, each ICU may have a different case mix of patients, different operating policies and procedures, and different staff. And within an ICU, changes to staff or policy over time could lead to differences in LOS between time periods. The following sections describe how the ICU and time period of admission can explain part of the variation in the LOS between patients.

Variation in the length of stay between ICUs

Each ICU has a true mean LOS for the infinite number of patients who could be hypothetically admitted to that ICU. When there is true variability between ICUs, the true mean LOS for each ICU will differ from the mean of all true ICU mean LOS. In the hypothetical situation where we have an infinite number of patients, the overall mean LOS for all patients and the mean of all true ICU mean LOS will be equal to the same true overall mean LOS.

Figure 2a, b, e and f show four scenarios that each illustrate variation in the true mean LOS across ICUs (*red circles*). The true mean LOS in each ICU may be similar and, therefore, close to the true overall mean LOS (*black line*) (Fig. 2a); or the true mean LOS of each ICU may be more dispersed about the true overall mean (Fig. 2b). The difference in the spread of true ICU mean LOS between Fig. 2a and b indicates greater variability in the true ICU mean LOS across ICUs in Fig. 2b than in Fig. 2a. The same comparison can be made between Fig. 2e and f.

Variation in the length of stay between time periods in an ICU

Within each ICU, there is also a true mean LOS for the infinite number of patients who could be hypothetically admitted in each 1-year period (i.e. each cluster-period). Figure 2a, b, e and f show also that there is variation in the difference between the true cluster-period mean LOS (*green circles*) and the true ICU mean LOS (*red circles*). The true cluster-period mean LOS may be similar to the true ICU mean LOS Fig. 2a); or the true mean LOS of each cluster-period may be more dispersed about the true ICU mean (Fig. 2e). The difference in the spread of the true cluster-period mean LOS between Fig. 2a and e indicates greater variability in true cluster-period mean LOS within ICUs in Fig. 2e than in Fig. 2a. The same comparison can be made between Fig. 2b and f.

Variation in length of stay between patients in a cluster-period

While there is a true mean LOS for all patients admitted in each cluster-period, the individual patients within each cluster-period will show variation in their LOS due to other patient factors (e.g. severity of their condition).

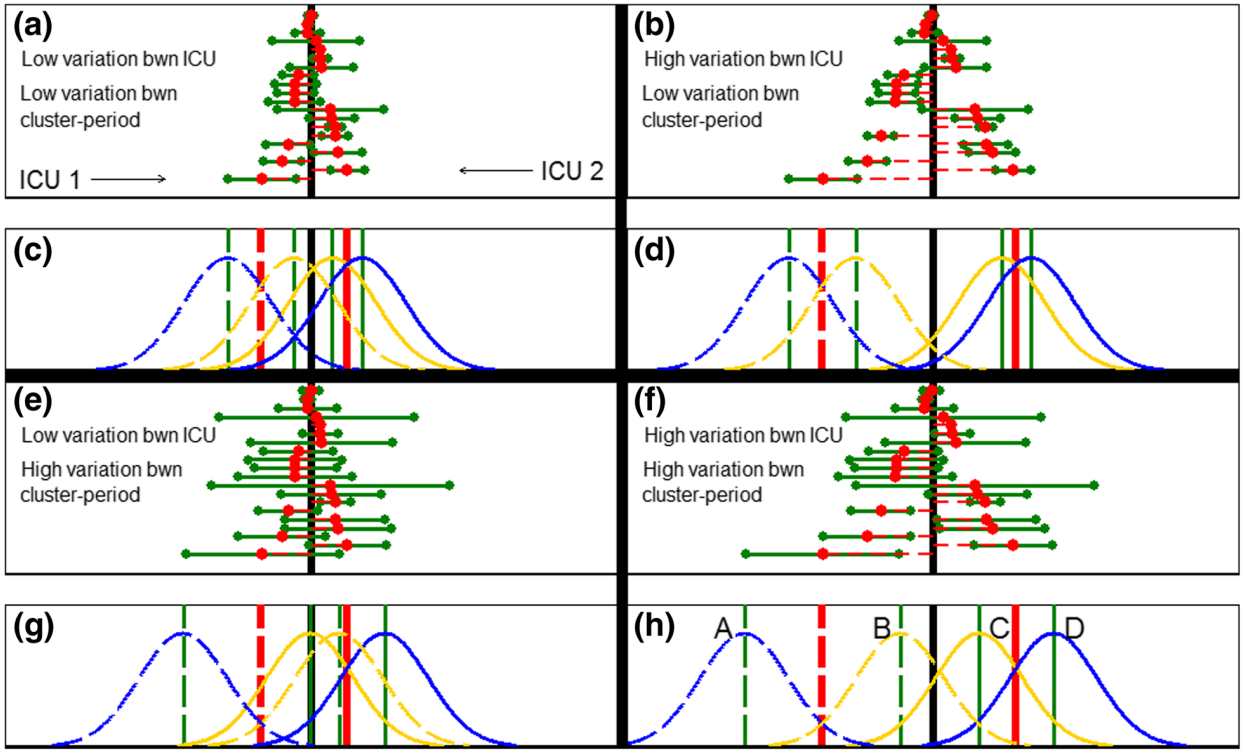


Fig. 2 Variation in true mean length(s) of stay (LOS) between intensive care units (ICUs) and between periods within ICUs. Low variation in the true mean LOS between ICUs is shown in the left column (**a**, **c**, **e**, **g**) and high variation in the right column (**b**, **d**, **f**, **h**). Low variation in the true mean LOS between periods within ICUs is shown in the top row (**a**, **c**, **b**, **d**) and high variation in the bottom row (**e**, **g**, **f**, **h**). **a**, **b**, **e**, **f** the true mean LOS for each of the 20 hypothetical ICUs are marked by a *red circle*, with the difference between the true overall mean LOS and the true mean LOS for each ICU indicated by a *dashed red horizontal line*. The two true cluster-period mean LOS for each ICU are marked with a *green circle* to the left and right of the true ICU mean LOS. The difference between the true ICU mean LOS and the true cluster-period mean LOS is indicated by a *green horizontal line*. The *black vertical line* indicates true overall mean LOS. **c**, **d**, **g**, **h** the *red vertical line* indicates the true ICU mean LOS and the *green vertical line* indicates the true cluster-period mean LOS for each period in each of two ICUs. For (**a**) $WPC = 0.02$, $BPC = 0.01$; for (**b**) $WPC = 0.06$, $BPC = 0.05$; for (**e**) $WPC = 0.06$, $BPC = 0.01$; for (**f**) $WPC = 0.10$, $BPC = 0.05$. ICU 1 is shown with *solid lines* and ICU 2 is shown in *dashed lines* in (**h**). The *yellow (blue) curve* indicates a normal distribution of patient LOS within each cluster-period where the cluster was allocated to intervention S (T). For (**d**) the distribution of patient LOS in each of the four cluster-periods are labelled A to D. WPC: within-cluster within-period correlation (ρ); BPC: within-cluster between-period correlation (η)

Two of the 20 example ICUs are depicted in Figs. 2c, d, g and h. ICU 1 is shown with *solid lines* and ICU 2 is shown in *dashed lines*. As previously, the mean LOS in each ICU is marked by a *red line*, and the mean LOS in each cluster-period is marked by a *green line*. The distribution of the individual patient LOS within each cluster-period follows a normal distribution, and is shown with four *yellow or blue curves*. The distribution of the LOS for patients receiving intervention S are coloured *yellow*, and the distribution of those receiving intervention T are coloured *blue*.

Within each cluster-period, patients have a range of individual LOS centred at the true cluster-period mean LOS (*green line*). Nonetheless, the patients in each cluster-period are from distinct distributions labelled as A, B, C, and D in Fig. 2h (these labels apply also to Fig. 2c, d and g). In each cluster-period, we assume that the variability of the individual patient LOS is the same, and hence the *yellow and blue curves* have the same shape and are only shifted in location between the four cluster-periods.

Summary of the sources of variation in the CRXO design

We have illustrated how the cluster randomisation aspect of the CRXO design leads to the formation of clusters of patients defined by ICU, while the crossover aspect of the design leads further to the formation of *cluster-periods* of patients within each cluster.

We have also illustrated how the cluster randomisation and crossover aspects of the CRXO design can lead to three sources (or components) of variation in the responses of patients in a CRXO trial: variation in the mean LOS between ICUs; variation in the mean LOS

between cluster-periods; and variation between individual patient LOS within a cluster-period.

The within-cluster within-period correlation and the within-cluster between-period correlation

In this section we show how the three sources of variation outlined in the preceding section can be used to quantify the similarity in LOS between the groups of patients defined by ICU (cluster) and cluster-period.

The *within-cluster within-period correlation (WPC)* quantifies the similarity of outcomes from patients in the same cluster-period. The *within-cluster between-period correlation (BPC)* quantifies the similarity of outcomes from patients in the same cluster, but in different periods. Specification of these two correlations are required to perform sample size estimates for a CRXO trial.

In the hypothetical circumstance where the LOS of an infinite number of patients admitted to each ICU is measured, we can determine the true WPC and BPC. In practice, the LOS can only be measured on a sample of patients, and the true WPC and BPC will be estimated from this sample of patients, with some amount of random sampling error.

We first describe the sources of variation underlying the BPC, and then extend the description to the WPC.

Within-cluster between-period correlation (BPC)

The BPC measures how much of the total variability in the LOS is due to variability in the ICU mean LOS or analogously how similar patient responses are within the same cluster, but in different periods. The formula for the BPC, η, is:

$$\eta = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CP}^2 + \sigma_I^2}, \quad (1)$$

where σ_C^2 is the variance in mean LOS between clusters (ICUs), σ_{CP}^2 is the variance in mean LOS between cluster-periods, and σ_I^2 is the variance in individual LOS within a cluster-period.

The BPC measures the similarity between the LOS of two patients from the same ICU with one patient from the first period (cluster-period C) and one patient from the second period (cluster-period D).

The *similarity* between the LOS of patients in an ICU *between* cluster-periods arises from the *variability* in the ICU mean LOS *only*. We now refer to Fig. 2 to describe how this relationship between similarity and variability arises. As the ICU mean LOS (*red lines/red circles*) become more dispersed between ICUs, relative to the dispersion (i.e. distance) between cluster-period mean LOS within an ICU (*green lines/green circles*), the distribution of the patient LOS (*yellow/blue curves*) in the cluster-periods A and B become more similar to each other, as do the distribution of patient LOS in cluster-periods C and D.

For example, in Fig. 2c there is little variation in the ICU mean LOS around the overall mean LOS (*black line*) and the distribution of patient LOS in cluster-periods A, B, C and D almost all coincide. As a result, the similarity between the LOS of patients in different cluster-periods *within the same ICU* (e.g. one patient from cluster-period A and one patient from cluster-period B) is comparable to the similarity between the LOS of patients in different ICUs (e.g. one patient from cluster-period A and one patient from cluster-periods C or D). In contrast, in Fig. 2d, there is more separation between the ICU mean LOS and only the distributions of patient LOS from the same ICUs coincide (i.e. cluster-periods A and B, and cluster-periods C and D, coincide). As a result, the LOS of patients in different cluster-periods *within the same ICU* (e.g. one patient from cluster-period A and one patient from cluster-period B) are more similar to each other than to the patients in other ICUs (e.g. one patient from cluster-period A and one patient from cluster-periods C or D). Hence, the BPC is larger in Fig. 2d than in Fig. 2c. The same comparison can be made between Fig. 2g and h.

The within-cluster within-period correlation (WPC)

The WPC measures how much of the total variability in the LOS is due to variability in the ICU mean LOS and the cluster-period mean LOS or analogously how similar patient responses are within a cluster-period. The formula for the WPC, ρ, is:

$$\rho = \frac{\sigma_C^2 + \sigma_{CP}^2}{\sigma_C^2 + \sigma_{CP}^2 + \sigma_I^2}. \quad (2)$$

The WPC measures the similarity in the LOS from two patients in the same cluster-period, e.g. cluster-period C.

The *similarity* between the LOS of patients *within* a cluster-period arises from the *variability* in the ICU mean LOS *and* cluster-period mean LOS. We now refer to Fig. 2 to describe how this relationship between similarity and variability arises. We describe the relationship in two parts: variability in the ICU mean LOS; and variability in the cluster-period mean LOS.

As the ICU mean LOS (*red circles/red lines*) becomes more disperse, relative to the dispersion (i.e. distance) between the cluster-period mean LOS (*green circles/green lines*), the distribution of the individual patient LOS (*yellow/blue curves*) in the four cluster-periods A, B, C and D become more distinct from each other, and hence patients within a cluster-period appear more similar to each other. For example, in Fig. 2c there is little

variation between the ICU mean LOS around the overall mean LOS (*black line*) and the distribution of patient LOS in cluster-periods A, B, C and D almost all coincide. As a result, the similarity between the LOS of two patients in cluster-period A is comparable to the similarity between the LOS of one patient from cluster-period A and one patient from cluster-period B (or C or D). In contrast, in Fig. 2d, there is more separation between the ICU mean LOS and hence more separation of the patient LOS in ICUs 1 and 2. As a result, the LOS of two patients in cluster-period A are more similar to each other than to one patient from cluster-period A (cluster 1) and another patient from cluster-periods C or D (cluster 2). Hence, the WPC is smaller in Fig. 2c than in Fig. 2d. We note that the same comparison can be made between Fig. 2g and h.

Likewise, as the cluster-period mean LOS (*green circles/green lines*) becomes more disperses, relative to the distance between the ICU mean LOS (*red circles/red lines*), the distribution of the individual patient LOS (*yellow/blue curves*) in the four cluster-periods A, B, C and D also become more distinct from each other, and hence patients within a cluster-period become more similar to each other. For example, in Fig. 2d there is little variation between the cluster-period mean LOS around the ICU mean LOS and thus the distribution of patient LOS in cluster-periods A and B (and equivalently C and D) almost coincide. As a result, the similarity between the LOS of two patients in cluster-period A is comparable to the similarity between the LOS of one patient from cluster-period A and one patient from cluster-period B. In contrast, in Fig. 2h, there is more separation between the cluster-period mean LOS and the distribution of patient LOS. As a result, the LOS of two patients in cluster-period A are more similar to each other than to one patient from cluster-period A and another patient from cluster-period B (and even more similar than one patient from cluster-period A and another patient from cluster-periods C or D). Hence the WPC is again smaller in Fig. 2d than in Fig. 2h. We note that the same comparison can be made between Fig. 2c and g.

Precision of the CRXO design compared to the parallel-group cluster randomised design and parallel-group, individually randomised design

In this section, we discuss how the WPC and BPC affect the precision of the estimate of the difference between interventions, and hence the sample size requirement, in a two-period, two-intervention, cross-sectional CRXO trial. We illustrate the two extremes of the CRXO design: when the precision in the CRXO design is equivalent to an IRCT design; and equivalent to a CRCT design. The allocation of interventions to patients in the IRCT, CRCT, and CRXO design are shown in Fig. 1.

To illustrate the effect of the WPC and BPC on precision (and equivalently the components of variation), we continue to assume that the true difference between interventions is zero. We consider a large sample of patients admitted to *one* cluster in a CRXO design, such that the sampling error in the estimated mean LOS for patients is assumed negligible. Therefore, in the single cluster shown in Fig. 3, the separation between the distribution of LOS from patients receiving intervention S (*yellow curve*) and intervention T (*blue curve*) arises solely from the variation in the mean LOS between cluster-periods (σ^2_{CP}). In this section, we show which partitioning of the total variation in LOS into the components of variation leads to the most precision and to the least precision in the CRXO design.

In the CRXO design, the observed mean LOS of patients receiving each intervention can be compared within each cluster because each intervention is delivered in each cluster. As an illustration, in Fig. 3a, the observed difference in mean LOS between patients receiving each intervention could be due to a difference in true cluster-period mean LOS (*green lines*) but not due to differences in the true ICU mean LOS because this component of variation is removed when the two interventions are compared within an ICU.

As the variation in the true cluster-period mean LOS increases, and hence the separation between the *green lines* in Fig. 3a increases, the separation between the *yellow and blue curves* within an ICU increases. Correspondingly, from Eqs. 1 and 2, the difference between the WPC and BPC increases. In conclusion, increasing variability in the cluster-period means leads to increasing uncertainty in the observed difference in the mean LOS between patients receiving each intervention.

In the CRXO design, precision is maximised when there is no variation in LOS between periods within a cluster. In this scenario the separation between the *green lines* in Fig. 3a shrinks and the *yellow and blue curves* coincide, yielding Fig. 3b. The LOS of two patients in the same cluster-period are as similar as the LOS of two patients from the same ICU but in different cluster-periods. Also, from Eqs. 1 and 2, the WPC equals the BPC. Figure 3b now approximates the diagram that one would expect from an IRCT with two ICUs (with the mean LOS for each centre indicated by the *green lines*) and half the patients within each cluster receiving each intervention. This diagram arises in an IRCT because, for large sample sizes and under the assumption of no true differences between interventions, randomisation ensures that the distributions of LOS in each intervention (*yellow and blue curves*) are identical. The CRXO design will, therefore, have the same precision as an IRCT design.

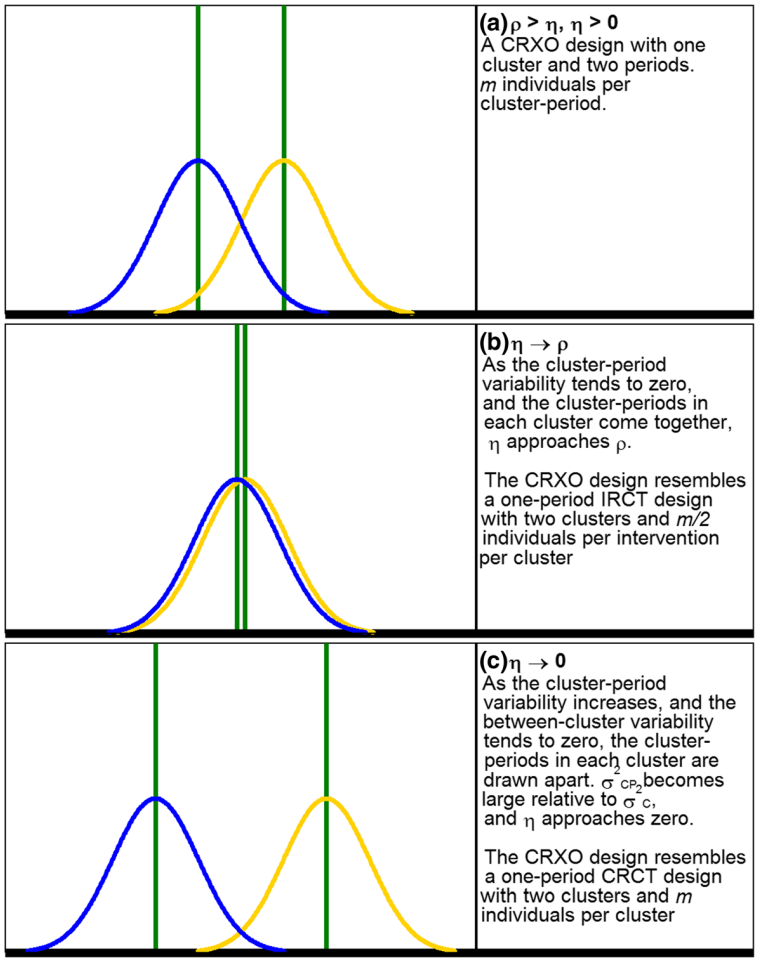


Fig. 3 A single cluster in the cluster randomised crossover (CRXO) design where (**a**) $\rho > \eta$, $\eta > 0$. **b** $\eta \rightarrow \rho$. **c** $\eta \rightarrow 0$. The *green solid vertical lines* indicate difference between true intensive care unit (ICU) mean length of stay (LOS) and true cluster-period mean LOS. The *yellow (blue) curve* indicates a normal distribution of patient LOS within each cluster or cluster-period where the patient or cluster was allocated to intervention S (T). The true difference between intervention S and T is zero. The total variance in LOS remains constant

Conversely, the precision of the CRXO design decreases when the cluster-period variability increases. As the variability between periods within a cluster increases, the separation between the *green lines*, and correspondingly the *yellow and blue curves*, in Fig. 3a increases. The increased separation results in greater variability in the comparison of patient LOS in each intervention within each cluster. For a fixed total variability in ICU LOS, as the variability between periods within a cluster increases, the variability between *different* clusters must reduce. In the limiting case there is no variation at all between clusters ($\sigma_C^2 = 0$), resulting in the BPC equalling zero (Eq. 1). In this case each cluster-period effectively resembles a separate cluster (Fig. 3c). Two patients in different cluster-periods *in the same ICU* are no more similar than two patients *in different ICUs*. Therefore, there is no advantage to the crossover component of the CRXO design and the CRXO will have the same precision as a CRCT design.

In most situations, the BPC will lie between zero and the WPC. In the following section, 'Performing a sample size calculation', we discuss the effect of the BPC and WPC on the sample size required to be able to detect a specified true intervention effect in a CRXO trial with a given level of power, and provide guidance on how to choose values for the BPC and WPC for a sample size calculation.

Performing a sample size calculation

The sample size required to detect a specified true difference between interventions with a given level of power decreases as the precision of the estimate of the intervention effect increases. In the 'Understanding the CRXO design' section, we considered precision in the

CRXO design when the true difference between interventions was assumed to be zero. However, even when the true difference is not zero, the effects of the WPC and BPC on precision described in the previous section continue to hold.

The sample size required for a CRXO trial increases as the cluster-period variability increases, or equivalently as the difference between the WPC and BPC increases. As the value of the BPC increases from zero to the WPC, the sample size required for the CRXO design will decrease from that required for a CRCT design towards the sample size for an IRCT. Therefore, using an appropriate specification of the difference between the WPC and the BPC is essential for performing sample size calculations for the CRXO design.

We now illustrate how to perform a sample size calculation for a two-period, two-intervention CRXO trial with a continuous and binary outcome using ICU LOS and in-ICU mortality data, respectively, from the Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database (APD) [14, 15]. There are 37 tertiary ICUs in Australia and New Zealand, of which 25 to 30 might be expected to participate in a trial.

We compare the sample size requirement for number of individuals and number of clusters (ICUs) from the CRXO design with the requirement from the stratified, multicentre, parallel-group, individually randomised design (IRCT) and the parallel-group cluster randomised design (CRCT) conducted over one period.

Comparisons of the sample size requirements for these different designs can either be made by fixing the total number of clusters across all designs; or by treating the CRXO design as lasting twice as long, i.e. two periods, instead of one period as in the IRCT and CRCT designs. We take the latter approach here so that the WPC is the same in each period.

We include Stata do-files to estimate the required sample size for each trial design, for a chosen set of sample size parameters (see Additional files 1 and 2).

The sample size formulae for a one-period IRCT design, a one-period CRCT design, and a two-period, two-intervention, cross-sectional CRXO design

The sample size formula for the *total number of participants* required for a normally distributed continuous outcome in a two-period, two-intervention CRXO trial, across all clusters and interventions, assuming a constant number of participants recruited to each cluster-period is [8]:

$$N_{CRXO} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 + (m-1)\rho - m\,\eta\right) + 4m,$$

and for a one-period, two-intervention CRCT:

$$N_{CRCT} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 + (m-1)\rho\right) + 2m,$$

and for a one-period, two-intervention, parallel-group IRCT, stratified by cluster, across all clusters and interventions is [16]:

$$N_{IRCT} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 - \rho\right),$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values corresponding to the upper tail probabilities of $\alpha/2$ and β, respectively; α is the two-sided significance level, typically 0.05; $1 - \beta$ is the power to detect the specified difference ($\mu_A - \mu_B$) with probability α; σ^2 is the variance of the outcome; μ_A and μ_B are the outcome means in each arm; m is the number of participants per cluster-period; ρ is the WPC; and η is the BPC.

The formulae presented above include a correction for when the number of clusters small, as suggested in Eldridge and Kerry (p. 149) [2] and Forbes et al. [9]. This leads to an additional $4\,m$ participants in the CRXO design and $2\,m$ participants in the CRCT design. No correction is necessary for the IRCT because the number of individual participants will be large in the example settings.

For a binary outcome we can replace $\frac{2\sigma^2}{(\mu_A - \mu_B)^2}$ with $\frac{p_A(1-p_A)+p_B(1-p_B)}{(p_A - p_B)^2}$ in the above formulae [12], where p_A and p_B are the proportions of the outcomes in each arm.

For the CRXO design, CRCT design and IRCT design, respectively, the formulae to determine the number of clusters (ICUs) needed to achieve the required number of participants are:

$n_{CRXO} = \frac{N_{CRXO}}{2m}$, $n_{CRCT} = \frac{N_{CRCT}}{m}$, and $n_{IRCT} = \frac{N_{IRCT}}{m}$.

Australian and New Zealand Intensive Care Society – Adult Patient Database (ANZICS-APD): estimates of the WPC and BPC

The ANZICS-APD is one of four clinical quality registries run by the ANZICS Centre for Outcome and Resource Evaluation and collects de-identified information on admissions to adult ICUs in Australia and New Zealand. A range of data is collected during patients' admissions, including ICU LOS and in-ICU mortality. In this section we use the ANZICS-APD data from 34 tertiary ICUs to estimate the correlations required to perform sample size calculations for CRXO trials. We estimate the values of the WPC and the BPC from two 12-month periods of data between 2012 and 2013 (Appendix 1).

Continuous outcomes

We follow the methods of Turner et al. to estimate the WPC and BPC (Appendix 1). Using the ICU LOS data,

the estimated WPC was $\hat{\rho} = 0.038$, and the BPC was $\hat{\eta} = 0.032$ (Table 1). The overall mean LOS was 5.3 log-hours, with a standard deviation 1.39 log-hours.

Binary outcomes

We follow the methods of Donner et al. to estimate the WPC and BPC (Appendix 1). Using the in-ICU mortality data, the estimated WPC was $\hat{\rho} = 0.010$, and the BPC was $\hat{\eta} = 0.007$. The overall mortality rate was 8.7%.

Sample size example for ICU LOS

Suppose we wish to design a two-period, two-intervention, CRXO trial to have 80% power to detect a true reduction in ICU LOS of 0.1 log-hours (1.1 h) using a two-sided test with a Type-I error rate of 5%. In practice, the choice of reduction in ICU LOS should be the minimally clinically important reduction, determined in consultation with subject matter experts. A 0.1 log-hours' reduction is equivalent to a 10% reduction, and is a reasonable minimally clinically important reduction in ICU LOS.

The standard deviation is estimated to be 1.2 log-hours (3.3 h). As an illustration, we assume that in a 12-month period, 200 patients in each ICU will meet the inclusion criteria for the trial. The CRXO trial will, therefore, run for 2 years and include 400 patients per ICU, with 200 patients receiving each intervention in each ICU.

For comparison, we consider an IRCT and a CRCT run for a 12-month period, with 100 patients receiving each intervention in each ICU in the IRCT and all 200 patients receiving one intervention in each ICU in the CRCT.

Using the estimates that we calculated from the ANZICS-APD data for the WPC and BPC, the total number of patients and ICUs for each design are summarised in Table 2 (see Appendix 2 for calculations).

T3

The total number of participants required for the CRXO design is $N_{CRXO} = 10{,}564$. To include 10,564 participants, we require $n_{CRXO} = 27$ ICUs, each recruiting 200 participants in each of the *two* 12-month periods. If instead we conducted a CRCT over a *single* 12-month time period, the total number of participants required would be $N_{CRCT} = 39{,}065$. Assuming that 200 patients are eligible in each ICU, we would need $n_{CRCT} = 196$ ICUs. The total number of participants required for an IRCT conducted over a 12-month period is $N_{IRCT} = 4345$. With 200 patients per ICU (100 patients per intervention), the total number of ICUs required is $n_{IRCT} = 22$.

In this example, the CRXO design required five more clusters (ICUs) than the IRCT design; however, the CRXO design is run for twice as long. The CRCT design would require 7.3 times as many clusters as the CRXO design. Given that there are only 37 tertiary ICUs in Australia and New Zealand, a CRCT trial would not be feasible.

We can examine the sensitivity of the CRXO sample size calculation to a different BPC. If the BPC was $\eta = 0.010$ rather than $\eta = 0.032$, then the CRXO design requires $N_{CRXO} = 30{,}433$ participants. The total number of ICUs required to obtain the required number of participants is $n_{CRXO} = 77$. The total number of ICUs required has now increased by 50, and the trial would no longer be feasible in the Australia and New Zealand region within tertiary ICUs only. Note that when the number of patients admitted in

Table 1 Calculation of the within-cluster, within-period correlation (WPC) and within-cluster, between-period correlation (BPC) for intensive care unit (ICU) log-length of stay (LOS) in the Australian and New Zealand Intensive Care Society – Adult Patient Database (ANZICS-APD)

$\hat{\sigma}^2_{ICU} = 0.045$
$\hat{\sigma}^2_{CP} = 0.008$
$\hat{\sigma}^2_{I} = 1.360$
$\hat{\rho} = \frac{\hat{\sigma}^2_{ICU}+\hat{\sigma}^2_{CP}}{\hat{\sigma}^2_{ICU}+\hat{\sigma}^2_{CP}+\hat{\sigma}^2_{I}} = \frac{0.045+0.008}{0.045+0.008+1.360} = 0.038$
$\hat{\eta} = \frac{\hat{\sigma}^2_{ICU}}{\hat{\sigma}^2_{ICU}+\hat{\sigma}^2_{CP}+\hat{\sigma}^2_{I}} = \frac{0.045}{0.045+0.008+1.360} = 0.032$

Table 2 Number of individuals and number of clusters required for a cluster randomised crossover (CRXO), cluster randomised controlled trial (CRCT) and individually randomised controlled trial (IRCT) trial with $\rho = 0.038$ for all designs and specified η for CRXO design

	Number of required individuals	Number of required ICUs
CRXO		
$\rho = 0.038$, $\eta = 0.032$	10,564	27
$\rho = 0.038$, $\eta = 0.010$	30,433	77
CRCT	39,065	196
IRCT	4345	22

ICU intensive care unit

Table 3 Number of individuals and number of clusters required for a cluster randomised crossover (CRXO), cluster randomised controlled trial (CRCT) and individually randomised controlled trial (IRCT) trial with $\rho = 0.010$ for each design and specified η for the CRXO design

	Number of required individuals	Number of required ICUs
CRXO		
$\rho = 0.010$, $\eta = 0.007$ (equal cluster sizes)	51,581	22
$\rho = 0.010$, $\eta = 0.006$ (equal cluster sizes)	63,811	27
$\rho = 0.010$, $\eta = 0.007$ (unequal cluster sizes)	41,208	23
CRCT	13,4792	113
IRCT	10,090	9

ICU intensive care unit

each cluster-period is relatively large, we would observe a similar increase in the sample size if we had underestimated the WPC by 0.023, rather than overestimated the BPC by 0.023.

Sample size example for in-ICU mortality

In a second example, suppose that we wish to design a study to have 80% power to detect a true reduction in in-ICU morality from 8.7% to 7.2% (absolute difference of 1.5%) using a two-sided test with a Type-I error rate of 5%. From the ANZICS-APD admission data, we estimate that in a 12-month period, 1200 patients will be admitted in each ICU and eligible for inclusion in the trial. The total number of patients and ICUs for each design are summarised in Table 3 (see Appendix 2 for calculations).

For a CRXO design, using the estimates for the WPC, the BPC, and the cluster-period size we calculated from the ANZICS-APD, the total number of participants required is $N_{CRXO} = 51{,}581$. Since we expect 1200 patients in each ICU for each of the two 12-month periods, the required number of ICUs is $n_{CRXO} = 22$. If we had used a CRCT, the required number of participants is $N_{CRCT} = 134, 792$. Assuming that 1200 patients admitted over a single 12-month period, we would need $n_{CRCT} = 113$ ICUs. The total number of participants required for the IRCT design is $N_{IRCT} = 10{,}090$. For a trial run over 12 months, with 1200 patients per ICU (600 patients per intervention), the total number of ICUs required is $n_{IRCT} = 9$.

In this example, the CRXO design required 2.4 times as many clusters (ICUs) as the IRCT design, and is run for twice as long. Despite the increase in required clusters, the CRXO is still a feasible design, unlike the CRCT design, which would require 5.1 times as many clusters as the CRXO design.

We can examine the sensitivity of the CRXO sample size calculation to a different BPC. If the BPC was $\eta = 0.006$, rather than $\eta = 0.007$, then the total number of participants required is $N_{CRXO} = 63{,}811$. Since we expect 1200 patients for each cluster-period, we would need to include $n_{CRXO} = 27$ ICUs, i.e. 54 cluster-periods. This demonstrates that a small change in the assumed BPC can have a marked impact on the number of required ICUs and patients.

Unequal cluster-period sizes

We have so far assumed that the cluster-period size is constant. In reality, it is likely that different ICUs will include a differing number of participants [17, 18]. An extension to the sample size formula for this scenario is provided by [9]. When the analysis is based on unweighted cluster-period means, the arithmetic mean in the sample size formula given for the CRXO design can be replaced by the harmonic mean:

$$m_h = n\sum_{i=1}^{n} \frac{1}{m_i}.$$

We assume that the cluster-period size is the same in each period within a cluster. For further extensions, see Forbes et al. [9].

From the ANZICS-APD data, we estimate that the harmonic mean is $m_h = 900$. Therefore then the required number of patients is $N_{CRXO} = 41{,}208$, and the required number of ICUs is:

$$n_{CRXO} = \frac{41208}{2 \times 900} = 23.$$

Allowing for unequal cluster-period sizes has increased the required number of clusters slightly from 22 to 23.

Guidance on how to choose the WPC and the BPC for the sample size calculation

As was seen in the 'Understanding the CRXO design' section, the difference between the WPC and BPC is key in determining the sample size for a CRXO design.

Approaches for choosing the within-cluster intracluster correlation (ICC) in sample size calculations for parallel-group CRCTs have been discussed [19–22]. Similar considerations apply when choosing the WPC in a CRXO design. In particular, because the ICC estimates are subject to large uncertainty [23], reviewing multiple relevant estimates of the ICC is recommended. These ICC estimates may be obtained from trial reports, lists published in journal articles or from routinely collected data.

Identification of the factors which influence the magnitude of the within-cluster ICC can assist trialists in selecting ICC estimates that are relevant to their planned trial. Typically, the trial outcome itself is less predictive of the value of the ICC than factors such as: the type of outcome variable (i.e. process outcomes that measure adherence to protocol and policy or individually measured outcomes) [19], the prevalence of the outcome [20], the size of the natural cluster of individuals that the randomised clusters are formed from [20], and the characteristics of the individuals and clusters [22].

The duration of time over which the outcome variables were measured may also affect the value of the within-cluster ICC. As the measurements of individuals within a cluster become further apart, the similarity between the measurements might be expected to decrease. Using an estimate of the within-cluster ICC that was determined over a different duration of time than the intended period length of the planned trial assumes that there is no variation in the within-cluster ICC over

time, and we are unaware of any research investigating if this is justified.

In contrast, we are aware of only two publications reporting estimates of the BPC [24, 25]. Therefore, until reporting of the BPC becomes more common [26], estimates of the BPC are likely to rely on the analysis of routinely collected data, pilot or feasibility study data, or a reasoned best-guess. As for the within-cluster ICC in cluster randomised trials, estimating the BPC from feasibility or a single routinely collected data source is likely to be subject to considerable uncertainty [27].

In forming a best guess, it is helpful to recognise that the difference between the WPC and BPC is a measure of changes over time within a cluster's environment that affect the outcomes of each individual in that cluster (e.g. a change in policy in one ICU). Over short time periods or in clusters with stable environments and patient characteristics, it might be reasonable to expect little change over time and, therefore, the BPC will be similar to the WPC. However, if this assumption is untrue and the BPC is less that the WPC, a sample size calculation assuming that the two correlations are equal will lead to an underpowered study. It may be prudent to assume that the BPC is less than the WPC. To this end, suggestions have been made to set the BPC to: half the WPC [12]; and to 0.8 of the WPC [11].

In the ANZICS-APD the ratio of the BPC to WPC is 0.7 for ICU mortality and 0.8 for ICU LOS, which is consistent with the suggestion made by Hooper and Bourke [11]. In the absence of multiple estimates or precise estimates of the ICCs, a conservative approach in selecting the BPC is recommended to avoid an underpowered trial. Further, a sensitivity analysis exploring the effect of the choice of ICC on the sample size is recommended.

Common mistakes when performing sample size calculations and analyses

Many trialists have made strong assumptions about the values of the WPC and the BPC in their sample size and analysis methodology [13]. In this section we illustrate the consequences of using incorrect sample size methodology on the estimated sample size and power.

Assume the outcomes are independent

In a review of CRXO trials, 34% of sample size calculations made the assumption that the observations were independent [13]. There are two scenarios where this assumption is reasonably appropriate: when the WPC and the BPC are equal and the sample size calculation was stratified by centre; or when the WPC and the BPC are both zero.

The first scenario arises when the outcomes of two individuals in the same cluster are equally similar if the individuals are in different periods as if the individuals are in the same period (i.e. there is no change in the WPC over time within a cluster). In this fortuitous case the precision gained by crossover aspect of the CRXO design equals the precision lost by cluster randomisation (apart from a factor of 1-WPC, which is usually small [16]). The second scenario arises when there is no similarity between the outcomes of any two individuals, which is unlikely.

The effect on power of assuming that the outcomes are independent will depend on the cluster-period size and the difference between the WPC and the BPC. Loss of power will increase as both the difference between the two ICCs increases and the cluster-period size increases.

We illustrate the potential effect on power and sample size assuming the outcomes are independent using a published sample size calculation. Roisin [28] estimated that the seven wards (clusters) participating in their trial required a minimum of 3328 patients to have 80% power to detect a reduction in proportion of hospital acquisition of methicillin-resistant *Staphylococcus aureus* (MRSA) from 3% to 1.5%. From the ANZICS-APD data, we estimate a WPC of 0.010, and a BPC of 0.007 for in-ICU mortality in the ICU setting. As an example only, we assume that the estimates of the correlations for ICU mortality are similar to the correlations for ICU MRSA acquisition. Given that a total of 2505 patients were eligible for inclusion in the study, we determined the average cluster-period size to be 179. From these estimates, we determine that a sample size of 5385 is required to achieve the specified power, which is a 62% increase from the published sample size requirement of 3328.

Assume a parallel-group cluster randomised design instead of a cluster randomised crossover design

Another common approach when performing sample size calculations for CRXO trials is to use methods designed for parallel-group CRCT trials. Applying CRCT sample size methodology to a CRXO design makes the assumption that: the BPC is zero; and that the WPC calculated over all periods in the trial is the same as the WPC calculated for a single period. Under the assumption that the BPC is zero, the outcomes of individuals within a cluster, but in different periods, are no more similar than outcomes of individuals in different clusters. That is, the individuals in different periods are assumed to be independent. When the BPC is not zero, the CRCT design effect does not account for the gain in precision achieved by the crossover aspect of the CRXO design, leading to a potentially overpowered trial. Trials that use CRCT sample size methods become

progressively more overpowered as the true BPC becomes larger and the cluster-period sizes increase.

We illustrate the potential effect on power and the sample size requirement using CRCT sample size methodology by means of a published sample size calculation. van Duijn [29] estimated that eight ICUs (clusters) participating in their trial would include 135 patient measurements per cluster-period. Using CRCT sample size methodology, each of the 16 cluster-periods (two periods per ICU) were assumed to be separate clusters of 135 patients. van Duijn [29] assumed a within-cluster ICC of 0.01, and hence they estimated that the trial required 1842 patients to have 80% power to detect a reduction in proportion of ICU patients with antibiotic-resistant gram-negative bacteria from 55% to 45%. From the ANZICS-APD data, we estimate a WPC of 0.010, and a BPC of 0.007, as in the example in the previous section. From these estimates, we determine that a sample size of 1623 is required to achieve the specified power, which is 12% less than the sample size required for a CRCT.

Discussion

Sample size calculations for CRXO trials need to account for both the cluster randomisation and crossover aspects of the design to ensure that an appropriate number of participants are recruited to adequately address the trial's hypotheses. There are simple, sample size formulae available for a two-period, two-intervention, cross-sectional CRXO design; however, the implementation of these formulae has been limited [13]. Such limited use of the formula may be due to a lack of recognition that formulae are available, a lack of availability of estimates of the parameters required within the formulae, or a lack of trialists' understanding of those parameters.

We have illustrated how the cluster randomisation and crossover aspects of the CRXO design give rise to similarity in both the responses of individuals within the same cluster and within the same cluster-period; and have described the parameters required to perform sample size calculations for CRXO trials. We have provided guidance on how to choose the parameters required for the sample size calculation and perform sample size calculation using those parameters.

While our focus has been on the two-intervention, two-period, cross-sectional CRXO design, more complex designs with additional periods and interventions are possible. The sample size and analysis methodology is more complex in these designs. For example, in a design with more than two periods, additional assumptions are required about the similarity between individuals in the same cluster in the same time period, and 1, 2, or 3, etc. time periods apart. Careful consideration should always be given to whether cluster randomisation is necessary [30], and whether the risk of the intervention effect from one period carrying over to the next period is minimal [6].

In addition to consideration of the sample size methodology, it is also essential to appropriately account for the cluster and the cluster-period in the analysis. Very few published trials do so [13]. Failure to account for the cluster-period in an individual level analysis leads to inflated Type-I error rates [31]. Methods to analyse CRXO trials have been published by Turner et al. and Forbes et al. [5, 9].

Conclusions

Sample size calculations for CRXO trials must account for both the cluster randomisation and crossover aspects of the design. In this tutorial we described how the CRXO design can be understood in terms of components of variation in the individual outcomes, or equivalently, in terms of correlations between the outcomes of individual patients. We illustrated how to perform sample size calculations for continuous and binary outcomes, and provided guidance on selecting estimates of the parameters required for the sample size calculation.

Appendix 1

Estimates of the WPC and BPC

To illustrate the impact of the WPC and BPC on the sample size calculation, we estimate the values of the WPC and BPC by using previously published methods for continuous and binary outcomes [5, 12].

Continuous outcomes

ICU LOS is right-skewed, so we begin by log-transforming this variable, so that the assumptions of the model used to estimate the correlations are more likely to be met. We use LOS to represent log(LOS) throughout. We estimate the values of the WPC and the BPC from the variances estimated by fitting the following model [5]:

$$Y_{ijk} = \mu + \pi + u_i + v_{ij} + e_{ijk},$$

where there are $i = 1, \ldots, n$ ICUs, $j = 1, 2$ 12-month periods and $k = 1, \ldots, m_{ij}$ patients in the i^{th} ICU (cluster) and j^{th} period; Y_{ijk} is the LOS for the k^{th} patient in the j^{th} cluster-period in the i^{th} ICU (cluster); μ is the overall mean LOS; π is the fixed period effect; $u_i \sim N(0, \sigma_C^2)$ is the difference from the overall mean LOS for each ICU mean LOS; $v_{ij} \sim N(0, \sigma_{CP}^2)$ is the difference from the ICU mean LOS for each cluster-period mean LOS, and $e_{ijk} \sim N(0, \sigma_I^2)$ is the difference from the cluster-period mean LOS for each patient LOS; σ_C^2, σ_{CP}^2, and σ_I^2 are the variances for the ICU (cluster) mean LOS, cluster-period mean LOS and patient LOS within each cluster-period, respectively.

Because we are fitting the model to registry data, rather than clinical trial data of the actual treatments to be considered, we estimate the model parameters under the assumption of a null treatment effect, and hence have not included a fixed treatment effect. A fixed treatment effect should be included when estimating the variance components from data from the actual clinical trial.

The model was fitted in Stata 14 with the mixed command using restricted maximum likelihood estimation: mixed log(LOS) periodeffect || cluster: || cluster_ period:, reml.

Binary outcomes

We estimate the value of the WPC for within-ICU mortality by fitting the analysis of variance (ANOVA) estimator for the intracluster correlation [12]:

$$\hat{\rho} = \frac{MSC - MSW}{MSC + (m_0 - 1)MSW},$$

$$MSC = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n} m_{ij}\left(\hat{P}_{ij} - \hat{P}_j\right)^2}{\sum_{j=1}^{2}(n-1)},$$

$$MSW = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n} m_{ij}\hat{P}_{ij}\left(1 - \hat{P}_{ij}\right)}{\sum_{j=1}^{2}\left(N_j - n\right)},$$

$$m_0 = \frac{N - \sum_{j=1}^{2}\sum_{i=1}^{n} m_{ij}^2/N_j}{\sum_{j=1}^{2}(n-1)},$$

where there are $i = 1, \ldots, n$ ICUs and $j = 1$, 2 12-month periods; m_{ij} is the number of patients in the i^{th} ICU (cluster) and j^{th} period; N_j is the total number of patients in each period and N is the total number of patients overall; $\hat{P}_{ij}$ is the estimated mortality rate in each cluster-period; and $\hat{P}_j$ is the estimated mortality rate in period j.

And by fitting the Pearson pairwise estimator for the BPC [12]:

$$\hat{\eta} = \frac{\sum_{i=1}^{n}\left(Y_{1i} - m_{1i}\hat{P}_1\right)\left(Y_{2i} - m_{2i}\hat{P}_2\right)}{\sqrt{\left(\sum_{i=1}^{n} m_{2i}\left(Y_{1i} - 2Y_{1i}\hat{P}_1 + m_{1i}\hat{P}_1^2\right)\right)\left(\sum_{i=1}^{n} m_{1i}\left(Y_{2i} - 2Y_{2i}\hat{P}_2 + m_{2i}\hat{P}_2^2\right)\right)}},$$

where Y_{1i} and Y_{2i} are the number of deaths in two adjacent time periods on the i^{th} ICU.

Appendix 2

Sample size calculations

In this section we provide the details of the sample size calculations presented in the 'Performing a sample size calculation' section, using the estimates for the WPC and BPC that we calculated from the ANZICS-APD data in Appendix 1.

Sample size calculation for ICU LOS

Total number of participants and ICUs required for the CRXO design

$$N_{CRXO} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 + (m-1)\rho - m\,\eta\right) + 4m,$$

$$N_{CRXO} = 2 \times (1.96 + 0.84)^2 \frac{2 \times 1.2^2}{(5.3-5.2)^2}(1 + (200-1)0.038 - 200 \times 0.032) + 4 \times 200 = 10564$$

Since we expect 200 patients in each ICU for each of the two 12-month periods, the number of ICUs needed to achieve the required number of participants is:

$$n_{CRXO} = \frac{N_{CRXO}}{2m} = \frac{10564}{2 \times 200} = 27.$$

If the BPC was $\eta = 0.010$ rather than $\eta = 0.032$, then:

$$N_{CRXO} = 2 \times (1.96 + 0.84)^2 \frac{2 \times 1.2^2}{(5.3-5.2)^2}(1 + (200-1)0.038 - 200 \times 0.010) + 4 \times 200 = 30433$$

The total number of ICUs required to obtain the required number of participants is:

$$n_{CRXO} = \frac{N_{CRXO}}{2m} = \frac{30433}{2 \times 200} = 77.$$

Total number of participants and ICUs required for the CRCT design

$$N_{CRCT} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 + (m-1)\rho\right) + 2m,$$

$$N_{CRCT} = 2\,(1.96 + 0.84)^2 \frac{2 \times 1.2^2}{(5.3-5.2)^2}(1 + (200-1)0.038) + 2 \times 200 = 39065$$

Assuming that 200 patients are eligible in each ICU over the 12-month trial period, we would need to include:

$$n_{CRCT} = \frac{N_{CRCT}}{m} = \frac{39065}{200} = 196 \text{ ICUs}.$$

Total number of participants and ICUs required for the IRCT design

$$N_{IRCT} = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\left(\mu_A - \mu_B\right)^2}\left(1 - \rho\right),$$

$$N_{IRCT} = 2\,(1.96 + 0.84)^2 \frac{2 \times 1.2^2}{(5.3-5.2)^2}(1 - 0.038) = 4345.$$

For a trial run over 12 months, with 200 patients per ICU (100 patients per intervention), the total number of ICUs required is:

$$n_{IRCT} = \frac{N_{IRCT}}{m} = \frac{4345}{200} = 22.$$

Sample size calculation for in-ICU mortality

Total number of participants and ICUs required for the CRXO design

$$N_{CRXO} = 2 \times (z_{\alpha/2} + z_{\beta})^2 \frac{p_A(1-p_A) + p_B(1-p_B)}{(p_A - p_B)^2} (1 + (m-1)\rho - m\eta) + 4m,$$

$$N_{CRXO} = 2 \times (1.96 + 0.84)^2 \frac{0.087 \times (1-0.087) + 0.072 \times (1-0.072)}{(0.087-0.072)^2} (1 + (1200-1) \times 0.010 - 1200 \times 0.007) + 4 \times 1200 = 51581$$

The number of ICUs needed to achieve the required number of participants is:

$$n_{CRXO} = \frac{N_{CRXO}}{2m} = \frac{51581}{2 \times 1200} = 22.$$

If the BPC was $\eta = 0.006$, rather than $\eta = 0.007$, then the total number of participants required is:

$$N_{CRXO} = 2 \times (1.96 + 0.84)^2 \frac{0.087 \times (1-0.087) + 0.072 \times (1-0.072)}{(0.087-0.072)^2} (1 + (1200-1) \times 0.010 - 1200 \times 0.006) + 4 \times 1200 = 63811$$

We would need to include:

$$n_{CRXO} = \frac{N_{CRXO}}{2m} = \frac{63811}{2 \times 1200} = 27 \text{ ICUs.}$$

Total number of participants and ICUs required for the CRCT design

$$N_{CRCT} = 2 (z_{\alpha/2} + z_{\beta})^2 \frac{p_A(1-p_A) + p_B(1-p_B)}{(p_A - p_B)^2} (1 + (m-1)\rho) + 2m,$$

$$N_{CRCT} = 2 (1.96 + 0.84)^2 \frac{0.087 \times (1-0.087) + 0.072 \times (1-0.072)}{(0.087-0.072)^2} (1 + (1200-1) \times 0.010) + 2 \times 1200 = 134792$$

We would need $n_{CRCT} = \frac{N_{CRCT}}{m} = \frac{134792}{1200} = 113$ ICUs.

Total number of participants and ICUs required for the IRCT design

$$N_{IRCT} = 2 (z_{\alpha/2} + z_{\beta})^2 \frac{p_A(1-p_A) + p_B(1-p_B)}{(p_A - p_B)^2} (1-\rho),$$

$$N_{IRCT} = 2 (1.96 + 0.84)^2 \frac{0.087 \times (1-0.087) + 0.072 \times (1-0.072)}{(0.087-0.072)^2} (1-0.010) = 10090$$

The total number of ICUs required is:

$$n_{IRCT} = \frac{N_{IRCT}}{m} = \frac{10090}{1200} = 9.$$

Additional files

Additional file 1: Continuous outcomes sample size Stata do file. Stata do file to perform sample size calculations for continuous outcomes using formulae presented in the 'Performing a sample size calculation' section, for a given set of sample size parameters. (DO 1 kb)

Additional file 2: Binary outcomes sample size Stata do file. Stata do file to perform sample size calculations for binary outcomes using formulae presented in the 'Performing a sample size calculation' section, for a given set of sample size parameters. (DO 2 kb)

Abbreviations

ANZICS-APD: Australia and New Zealand Intensive Care Society – Adult Patient Database; BPC: Within-cluster between-period correlation; CRCT: Cluster randomised controlled trial; CRXO: Cluster randomised crossover; ICC: Intracluster correlation; ICU: Intensive care unit; IRCT: Individually randomised controlled trial; LOS: Length of stay; WPC: Within-cluster within-period correlation

Acknowledgements

Not Applicable.

Funding

This research was in part supported by a National Health and Medical Research Council (NHMRC) project grant (1108283).
SJA was supported in part by a Monash University Graduate Scholarship and a National Health and Medical Research Council of Australia Centre of Research Excellence grant (1035261) to the Victorian Centre for Biostatistics (ViCBiostat).
JEM was supported by a National Health and Medical Research Council (NHMRC) Australian Public Health Fellowship (1072366).

Availability of data and materials

Not applicable.

Authors' contributions

SJA led the development of all sections and drafted the manuscript. JEM contributed to the development of all sections and provided critical review of the manuscript. KH contributed to the development of the graphical illustrations and corresponding sections, and provided critical review of the manuscript. DP provided guidance on the ANZIC-APD data and contributed to the development of the sample size examples. ABF conceived of the graphical illustrations, contributed to the development of all sections and provided critical review of the manuscript. All authors read and approved the final manuscript.

Authors' information

Not applicable

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Public Health and Preventive Medicine, Monash University, The Alfred Centre, Melbourne, VIC 3004, Australia. [2]Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [3]Australian and New Zealand Intensive Care Society Centre for Outcome and Resource Evaluation, Ievers Terrace, Carlton, VIC 3154, Australia. [4]Department of Intensive Care, The Alfred Hospital, Commercial Road, Melbourne, VIC

3004, Australia. [5]Australian and New Zealand Intensive Care Research Centre, School of Public Health and Preventive Medicine, Monash University, The Alfred Centre, Melbourne, VIC 3004, Australia.

Received: 30 July 2016 Accepted: 19 July 2017
Published online: 15 August 2017

References

1. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57–61.
2. Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester: Wiley; 2012.
3. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess. 1999;3(5):iii–92.
4. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. Am J Epidemiol. 1981;114(6):906–14.
5. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. Stat Med. 2007;26(2):274–89.
6. Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. Contemp Clin Trials. 2007;28(3):316–23.
7. Hills M, Armitage P. The two-period cross-over clinical trial. Br J Clin Pharmacol. 1979;8(1):7–20.
8. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Stat Med. 2008;27(27):5578–85.
9. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. Clin Trials. 2015;12(1):34–44.
10. Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. J Educ Behav Stat. 2011;36(4):472–90.
11. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. BMJ. 2015;350:h2925.
12. Donner A, Klar N, Zou G. Methods for the statistical analysis of binary data in split-cluster designs. Biometrics. 2004;60(4):919–25.
13. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. J Clin Epidemiol. 2016;74:40–50.
14. Stow PJ, Hart GK, Higlett T, George C, Herkes R, McWilliam D, Bellomo R, Committee ADM. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. J Crit Care. 2006;21(2):133–41.
15. Kaukonen KM, Bailey M, Pilcher D, Cooper DJ, Bellomo R. Systemic inflammatory response syndrome criteria in defining severe sepsis. N Engl J Med. 2015;372(17):1629–38.
16. Vierron E, Giraudeau B. Sample size calculation for multicenter randomized trial: taking the center effect into account. Contemp Clin Trials. 2007;28(4):451–8.
17. Konstantopoulos S. Power analysis in two-level unbalanced designs. J Exp Educ. 2010;78(3):291–317.
18. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med. 2001; 20(3):377–90.
19. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials. 2005;2(2):99–107.
20. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. J Clin Epidemiol. 2005;58(3):246–51.
21. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. Am J Epidemiol. 1999;149(9):876–83.
22. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57(8):785–94.
23. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. Stat Med. 2002; 21(24):3757–74.
24. Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials. 2016;17:402.
25. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med. 1994; 13(1):61–78.
26. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016;35(26):4718–28.
27. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? Stat Methods Med Res. 2016;25(3):1039–56.
28. Roisin S, Laurent C, Denis O, Dramaix M, Nonhoff C, Hallin M, Byl B, Struelens MJ. Impact of rapid molecular screening at hospital admission on nosocomial transmission of methicillin-resistant staphylococcus aureus: cluster randomised trial. PLoS One. 2014;9(5):e96310.
29. van Duijn PJ, Bonten MJ. Antibiotic rotation strategies to reduce antimicrobial resistance in Gram-negative bacteria in European intensive care units: study protocol for a cluster-randomized crossover controlled trial. Trials. 2014;15:277.
30. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. CONSORT 2010 Statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
31. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. Stat Med. 2017;36(2):318–33.

Chapter 6

Sample size calculations for cluster randomised crossover trials in Australian and New Zealand intensive care research

This chapter provides detailed considerations on performing a sample size calculation for a CRXO trial in the intensive care setting. This paper is important to inform future CRXO trial design, by providing estimates of the required sample size parameters for trials conducted within the Australia and New Zealand intensive care setting. Provision of these estimates is required, since as identified in Chapter 3, no published CRXO trials have reported both the within-cluster within-period ICC and within-cluster between-period ICC from the outcome data. This chapter also extends the previously published sample size formulae for CRXO trials to stratified trial designs. This extension is important in the context of trials undertaken in the Australia and New Zealand intensive care setting, because the parameters used in the sample size calculation differ by strata (e.g. geographic region or hospital type).

The methods, guidance and examples provided in this paper target health researchers designing CRXO trials in the intensive care setting, but can be applied to both stratified and unstratified CRXO trials in other settings. Estimates of the required sample size parameters are calculated for all-cause in-hospital mortality from the Australian and New Zealand Intensive Care Society Adult Patient Database clinical registry. These estimates may be applicable to other CRXO trials, because the value of the ICC may be predicted by factors such as: the type of outcome variable, the prevalence of the outcome, the size of the natural cluster of individuals that the randomised clusters are formed from, and the characteristics of the individuals and clusters.

Chapter 6 is presented as a manuscript, which was published in *Critical Care and Resuscitation* in June 2018. The pages have been renumbered for the thesis, but the manuscript is otherwise unchanged. In this paper, the terminology used for "Intracluster

correlation" is "correlation".

Supplementary tables and additional files referred to in the manuscript are appended to this thesis as follows:

Location in thesis	**Referred to in manuscript**	**Content of appendix**
Appendix F	Appendix I	Sample size formulae
	Appendix II	Methods for data inclusion and analysis
	Appendix III	Unstratified sample size calculation example: patients requiring mechanical ventilation
	Appendix IV	Stratified sample size calculation example
	Supplementary file	Stata ado file

This is the first time a stratified sample size formula for CRXO trials with risk ratio as the outcome measure has been published, and further details of the derivation are provided in Appendix G.

ORIGINAL ARTICLES

Sample size calculations for cluster randomised crossover trials in Australian and New Zealand intensive care research

Sarah J Arnup, Joanne E McKenzie, David Pilcher, Rinaldo Bellomo and Andrew B Forbes

ABSTRACT

Objective: The cluster randomised crossover (CRXO) design provides an opportunity to conduct randomised controlled trials to evaluate low risk interventions in the intensive care setting. Our aim is to provide a tutorial on how to perform a sample size calculation for a CRXO trial, focusing on the meaning of the elements required for the calculations, with application to intensive care trials.

Data sources: We use all-cause in-hospital mortality from the Australian and New Zealand Intensive Care Society Adult Patient Database clinical registry to illustrate the sample size calculations.

Methods: We show sample size calculations for a two-intervention, two 12-month period, cross-sectional CRXO trial. We provide the formulae, and examples of their use, to determine the number of intensive care units required to detect a risk ratio (RR) with a designated level of power between two interventions for trials in which the elements required for sample size calculations remain constant across all ICUs (unstratified design); and in which there are distinct groups (strata) of ICUs that differ importantly in the elements required for sample size calculations (stratified design).

Results: The CRXO design markedly reduces the sample size requirement compared with the parallel-group, cluster randomised design for the example cases. The stratified design further reduces the sample size requirement compared with the unstratified design.

Conclusions: The CRXO design enables the evaluation of routinely used interventions that can bring about small, but important, improvements in patient care in the intensive care setting.

Crit Care Resusc 2018; 20 (2): 117-123

Designing randomised controlled trials (RCTs) in the intensive care setting can be challenging. A recent review of observed effect sizes in 38 mortality trials calculated an average effect size of 1.4%, in contrast to the average effect size of 10.1% that was hypothesised when the trials were planned.[1] Detection of such small intervention effects using individual randomisation, with reasonable levels of statistical power (eg, > 80%), can lead to designs in which it is not possible to recruit the required number of participants. Further, individual randomisation is often not feasible in intensive care settings for interventions such as infection control and "bundles of care", which involve multiple practice changes simultaneously. This is because individual randomisation in intensive care settings involves a high risk of "contamination" between intervention arms. Individual randomisation is also not feasible for whole-of-intensive care unit processes, such as admission and discharge policies, because varying procedure by individual patient is impractical.

For evaluation of the above interventions, cluster randomised designs are generally preferred. In these designs, all patients in each ICU (cluster) receive the same intervention.[2] Cluster randomisation allows the intervention to be delivered as if it were standard operating practice and may increase the efficiency of data collection.[3] However, cluster randomisation is likely to result in the need for more participants than individual randomisation. The statistical power in a cluster randomisation trial depends on the number of ICUs, the number of patients in each ICU, and the similarity of the outcome responses of patients within each ICU.[2] As the outcome responses of patients within an ICU become more similar to each other than to patients in different ICUs, (within-cluster correlation), cluster randomisation will further reduce the statistical power compared with individual randomisation.

An alternative to the parallel-group, cluster randomised design is the cluster randomised crossover (CRXO) design.[4,5] Bellomo and colleagues[3] have called for an increase in the use of the CRXO design to evaluate routinely used and low-risk interventions such as oxygen therapy, ulcer-prophylaxis therapy and intravenous fluids and nutrition in RCTs.

In a two-intervention, two-period CRXO design, each ICU receives the two interventions, but they receive them sequentially in two separate periods of time called "cluster-periods". The order in which the interventions are delivered to each ICU is randomised to control for changes over time that are independent of the intervention (eg, policy changes), but which might have an impact on the trial outcomes. By comparing the interventions within each ICU, the ICU-specific component of variation is removed

from the estimate of the difference between interventions. Therefore, as for an individually randomised crossover trial, the inclusion of the crossover element has the benefit of reducing the required number of participants compared with a parallel-group design.[6]

The sample size calculation is a critical element in designing RCTs. To date, a tutorial on how to calculate the sample size for a CRXO trial, with a focus on designing intensive care trials, has not been available. Further, in practice, there are often distinct groups (strata) of ICUs that differ in the elements required for sample size calculation, such as mortality rates or within-cluster correlations. These elements need to be accounted for in the sample size calculations. The purpose of this article is to fill these gaps in the literature through the provision of examples of sample size calculations for a set of hypothetical intensive care trials.

Sample size considerations for CRXO trials

A sample size calculation for a CRXO trial needs to account for both the cluster randomisation and crossover aspects of the design. We now briefly discuss sample size and design elements. A detailed discussion by Arnup and colleagues is available elsewhere.[7]

Our focus is a two-intervention, two-period design in which different individuals are included in each period. Each ICU is randomised to receive intervention A for one period and then to cross over to intervention B for the second period, or vice versa. The basic sample size formula for the number of ICUs required to detect a risk ratio (RR) with power 1–β between two interventions, with probability α, is shown in Equation 1 (see Appendix I, online at cicm.org.au/journal.php).

Within-period correlation

When cluster randomisation is used, the responses of patients within the same ICUs are often more similar to each other than to patients in different ICUs.[2] This degree of similarity depends on how variable the response rate is across ICUs. In the CRXO design, the similarity of the responses from two patients within the same ICU, within the same period of time, is quantified by the within-ICU *within-period correlation* (WPC).

Between-period correlation

When a trial design with multiple time periods is used, we also need to consider the similarity of responses of patients within the same ICU, but in different periods of time. We quantify this similarity by the within-ICU *between-period correlation* (BPC). If the ICU environment is relatively stable, we might expect the responses of two patients from the same ICU in different periods to be similar, but less so than if patients were from the same period, because of potential changes in the ICU environment over time. This typically results in the BPC being less than the WPC.

Parallel-group versus CRXO designs

To illustrate the potential benefit of the crossover aspect in the CRXO design, we compare the power to detect an RR of 0.80 (20% relative risk reduction) between two interventions in a parallel-group cluster randomised trial and a crossover cluster randomised trial, assuming a baseline risk of 10% and an ICU admission rate of 1000 patients per year. The sample size formula for the number of ICUs required to detect an RR with power 1–β between two interventions with probability for a parallel-group cluster randomised trial is shown in Equation 2 (see Appendix I).

For a parallel-group, cluster randomised trial conducted over 24 months, the cluster size is 2000. For this example, we assume that the WPC = 0.02. For a CRXO trial conducted over 24 months, with two 12-month periods, the cluster-period size is 1000. We assume that the WPC = 0.02 and the BPC = 0.01. Given these assumptions, the required number of clusters for the parallel-group design is 67, and the required number of clusters for the crossover design is 38. Hence, by including a crossover, the required number of clusters has been reduced by 45%. The required number of patients for the parallel-group design is therefore 67 × 2000 = 134 000; and the required number of patients for the crossover design is 38 × 2 × 1000 = 76 000. The inclusion of a crossover has reduced the required number of patients by 43%.

Effect of within-period correlation

To illustrate the effect of the WPC on the sample size requirements, we begin by discussing the extreme values of the WPC. When the WPC is 1, all patients within the same ICU in the same time period have the same response, so there is no gain in information from sampling more than one patient. When the WPC is 0, the responses of two patients in the same ICU are no more similar than two patients in different ICUs, so the information contained in the ICU is the same as the information that would be obtained from completely independent patients. Therefore, as the WPC increases from 0, the precision of the estimate of the intervention effect in each time period decreases. As a result, the power to detect a specified RR between interventions also decreases.

In the CRXO trial described in the previous section, the WPC was 0.02 (and the BPC was 0.01), and as a result, 38 ICUs (76 000 patients) were required to detect an RR of 0.80 with 80% power (Figure 1, grey solid curve). However, if the WPC is 0.03 (and the BPC remains 0.01), then 38 ICUs

will only achieve 54% power (Figure 1, black solid curve) to detect the same size effect. The effect of the WPC in a CRXO trial acts in the same way as the intracluster correlation coefficient in a parallel-group cluster randomised trial (as shown by Campbell and colleagues[8]).

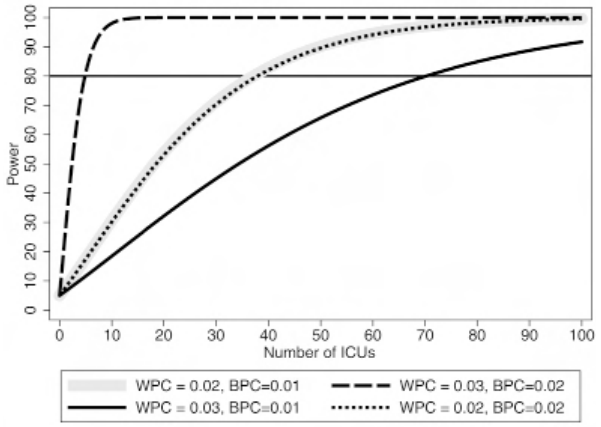
Effect of between-period correlation

To understand the effect of the BPC on the sample size requirement, we first note that the outcome of a specific patient depends on which ICU (cluster) the patient was admitted to, any time effects unique to that ICU, the intervention the patient receives, and factors idiosyncratic to the individual patient. In a CRXO design, we compare interventions A and B by comparing the responses of patients in the first period with the responses of patients in the second period (or equivalently responses in the second period with the first period), *within each ICU*. By performing these comparisons within each ICU, the component of the patient outcomes that is due to the ICU (cluster) that remains constant over time will cancel out from the comparison. The removal of this between-ICU variability enables us to obtain a more precise estimate of the intervention effect.

We now consider the extreme values of the BPC to demonstrate its effect. When the BPC is the same as the WPC, the responses of patients in two different periods are as similar as two responses from patients in the same period; this is the largest value that the BPC can take because the responses of two patients in different periods cannot be more similar than the responses from patients in the same period. In this scenario, all of the ICU-specific effect is removed from the comparison between intervention A and B. In contrast, when the BPC is zero, there is no similarity at all between patients in the same ICU in different periods and there is no value in using crossover design, because no part of the ICU-specific effect is removed from the comparison. Therefore, as the BPC becomes closer to the WPC, more of the ICU-specific effect is removed from the comparison between interventions A and B, and as a result, the power to detect a specified RR between interventions increases. (See Arnup and colleagues, 2017, for a detailed graphical explanation of this phenomenon.[7])

As an example, when the BPC is half the value of the WPC (WPC = 0.02, BPC = 0.01), 38 ICUs are required to detect an RR of 0.80 with 80% power (Figure 1, grey solid curve). The same power can also be achieved if the absolute difference between the WPC and the BPC remains 0.01, but the values of the WPC and the BPC change; that is, WPC = 0.03, BPC = 0.02 (Figure 1, short-dashed curve).[9] However, if the BPC takes the maximum value of equal to WPC (ie, WPC = BPC = 0.02), then 38 ICUs will achieve far greater power, close to 100% (Figure 1, long-dashed curve).

Figure 1. Power to detect a risk ratio of 0.8 (baseline risk, 10%; ICU size, 1000 patients/ cluster-period) between interventions with varying within- and between-period correlations*



ICU = intensive care unit. WPC = within-period correlation. BPC = between-period correlation. * The point at which the horizontal line intersects the power curve indicates the total number of ICUs required to achieve 80% power.

Effect of cluster size

As for parallel-group cluster randomised trials, power depends to a far greater extent on the number of ICUs than on the number patients within each ICU.[2] For example, for a parallel-group cluster randomised trial designed to detect a RR of 0.80 (an absolute risk reduction from 10% to 8%) with an intracluster correlation of 0.01 and with 20 ICUs, doubling the size of the trial from 10 000 to 20 000 patients by increasing the ICU size from 500 to 1000 patients per period increases the power from 51% to 55%. In contrast, doubling the size of the trial from 20 to 40 ICUs with 500 patients per period increases the power to 80%. Therefore, to achieve the same level of power for a larger number of ICUs requires far fewer patients overall, compared with a smaller number of ICUs with more patients per ICU.

Period effects

Changes in the trial environment between the first period and the second period can lead to changes in the patient responses in the second period that are unrelated to the delivered intervention. Such "period effects" alter all patient responses in the second period by the same amount, independent of the intervention given in that period. In a balanced CRXO design, in which the same number of clusters receive intervention A and B in each period, any period effects are removed from the comparison between interventions A and B, and hence do not introduce bias to

the estimate of the intervention effect. However, when the design is not balanced, one needs to explicitly accommodate period effects into the analysis. The sample size formulae can be modified slightly to accommodate period effects, but that detail is beyond the scope of this article. Forbes and colleagues have discussed a relevant approach to this.[9]

Carry-over effects

A requirement of the design is that the effect of the intervention given in the first period does not carry over to affect the responses of patients in the second period. Otherwise, the estimated intervention effect may be biased. Carry-over can occur at the level of the individual patient, or at the level of the ICU environment. Including different patients in each period (ie, a cross-sectional design) removes the risk of carry-over at the level of the individual patient, but if the intervention administered in the first period can lead to changes in the behaviour of the health care team or in the ICU environment that persists into the second period, then the potential for carry-over exists. No statistical method can detect or remove carry-over effects from a two-intervention, two-period CRXO design, and therefore it is essential that the risk of carry-over is minimised by the trial design itself, for example by including an appropriate washout period.[10]

Stratification

ICUs can form strata with distinct characteristics (eg, diagnostic casemix; different ratios of high mortality risk emergency patients to elective surgical cases; different hospital protocols for delivery of care, such as infection control and admission and discharge policies; variation in the availability of therapeutic services within the hospital, such as interventional radiology, cardiac catheterisation and extracorporeal membrane oxygenation; or hospital location) which lead to variations between strata in the outcome measure and other elements required for sample size calculations, for example, mortality rate, ICU admission rate, and the WPC and BPC values, respectively. Failure to account for differences in these elements can lead an inappropriate sample size. An extension to the sample size formula to include multiple strata is given in Equation 3 (see Appendix I). We provide an illustration in Section 3.

Adjustments to sample size formula with small numbers of clusters

When the estimated number of ICUs is small (eg, fewer than 30), it is recommended than one additional ICU is included per intervention (two ICUs in total) in the sample size calculation.[2,9]

Sample size examples

In this section we show how to perform sample size calculations for a hypothetical trial of interventions to reduce all-cause in-hospital mortality taking place over a total duration of 2 years (ie, a cluster period duration of 12 months). We begin with unstratified (ie, "conventional") CRXO trials that include patients from only one stratum. We then extend to cases in which event rates, and hence clustering effects, vary between two strata: tertiary ICUs, and metropolitan and rural ICUs combined.

We obtain estimates of the WPC and BPC for all-cause in-hospital mortality, the annual number of ICU admissions, and baseline mortality rates from the Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database (APD) clinical quality registry. The ANZICS-APD is a clinical quality registry, managed by the ANZICS Centre for Outcome and Resource Evaluation, which collects de-identified information on admissions to adult ICUs in Australia and New Zealand.[11] The methods used for data extraction and calculation of the correlations and annual number of admissions is shown in detail in Appendix II. A file with code to calculate power in the statistical package Stata, version 14 (StataCorp), is also provided.

The estimates calculated from the ANZICS-APD (Table 1) are used to design a two-intervention, two 12-month period CRXO trial, where all-cause in-hospital mortality in all admitted patients is the primary outcome. In Appendix III, we show the analogous calculation for a trial including a restricted patient group — those receiving mechanical ventilation. The formula for the sample size for the required number of ICUs is shown in Equation 1 in Appendix I.

Unstratified sample size calculation example

Suppose a CRXO trial is being planned to compare a buffered crystalloid fluid with saline in all patients requiring

Table 1. Number of ICUs meeting inclusion criteria,* and other ICU characteristics, by ICU stratum and combined across all strata

ICU stratum	Included ICUs (*n*)	Average annual admissions (*n*)	Harmonic mean annual admissions (*n*)	Mortality rate (%)	WPC	BPC
Tertiary	34	1356	1114	9.05%	0.006	0.005
Metropolitan and rural	54	638	553	6.65%	0.008	0.007
Combined	88	911	684	8.03%	0.009	0.006

ICU = intensive care unit. WPC = within-period correlation. BPC = between-period correlation. * See Appendix II, online at cicm.org.au/journal.php.

fluids. The effect to be detected is a 10% reduction in all-cause in-hospital mortality (RR = 0.90). The investigators wish to detect this effect with 90% power, assuming a 5% significance level.

The sample size formula requires values of the "harmonic mean" number of patients per ICU per 12-month period (see Appendix I for definition); the event rate of the outcome (in this example, mortality); and the WPC and BPC. Using the estimates of these values calculated from the APD data (see row labelled "combined" in Table 1) gives a sample size requirement of 105 ICUs (191 310 patients).

A case for stratification

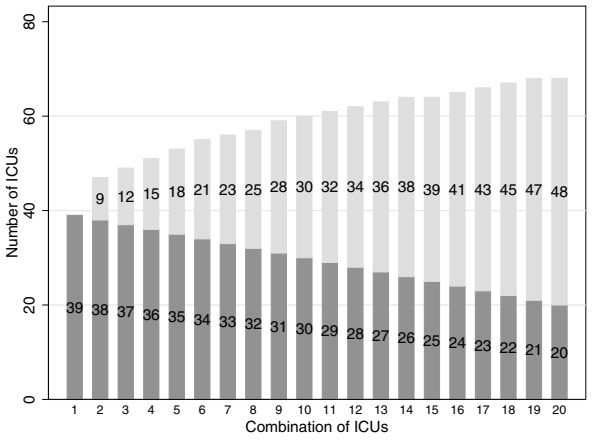
Rather than combining all ICUs together in a single calculation, as in Equation 1 (see Appendix I), we can include the differences between types of ICUs (strata) to reduce the overall sample size requirement. Table 1 shows that the number of annual admissions to tertiary ICUs is more than twice that of the metropolitan and rural ICUs; the mortality rate is highest in tertiary ICUs; and that there are small but important differences in WPC and BPC values across strata. The lower value of the WPC in the tertiary ICUs indicates less variability in mortality rates between tertiary ICUs compared with the metropolitan and rural ICUs. For WPC and BPC values, the differences between the tertiary and the metropolitan and rural ICU values are similar, suggesting that the variability of the ICU environment in these hospitals is similar over time.

Stratified sample size calculation example

The variance (or equivalently precision) of the specified intervention effect is a key element in all sample size calculations. For a stratified sample size calculation, this concept is extended to incorporate the different variances in each stratum, which arise from the different sample size elements in each stratum (eg, different mortality rates, WPC and BPC). Additionally, for a stratified sample size calculation, triallists need to decide how many ICUs (clusters) they will sample from each stratum. Different combinations of numbers across the strata will change the total required number of clusters.

Allowing for stratification can bring about important benefits in terms of reducing the number of required clusters in a CRXO design. We extend our previous example, allowing for the differences in the estimates of annual number of admissions, mortality rates and correlations (WPC and BPC) between the tertiary and the combined metropolitan and rural strata (see Appendix IV). As shown in Figure 2, depending on the combination of numbers of ICUs chosen from each stratum, the reduction in the required number of ICUs (clusters) and patients is *at least* 30%.

Figure 2. Sample size requirements for total number of ICUs, and number of ICUs from each of the strata, required to obtain 90% power to detect a risk ratio of 0.90 between interventions*



ICU = intensive care unit. * Each bar represents a different number (combination) of ICUs from each of the strata. Dark grey = tertiary ICUs. Light grey = metropolitan and rural ICUs.

Effect of stratification on sample size requirement

Stratification reduces the sample size requirement for a CRXO trial when ICUs can be differentiated by a known factor or set of factors (in this case, the type of hospital), and this factor strongly affects the outcome of interest in the trial (in this case mortality: mortality varies by type of hospital).

To understand the effect of stratification, first note that for any sample size calculation, the sample size will decrease when the intervention effect is estimated with more precision. In a CRXO trial, this occurs as the difference the WPC and the BPC decreases (See Section 2). In the ANZICS-APD data, there are substantial differences in mortality rates *across* the two strata, and as a result, the within-stratum variability in mortality rates (and hence the WPC within each stratum), is markedly smaller than for all ICUs combined (Table 1). Because the BPC remains relatively constant across the two strata, the difference between the WPC and the BPC is much smaller in the individual strata than in the combined data, and this serves to reduce the sample size.

Proportion of ICUs to select from each stratum

The next step is to determine the proportion of the total number of ICUs that are tertiary, with the remainder being metropolitan or rural. The tertiary ICUs offer considerable advantages: not only are they larger in size (patient numbers), but they also have a higher mortality rate (and hence a larger absolute intervention effect). As a result,

the smallest total required number of ICUs will be obtained when the trial is restricted to tertiary ICUs. However, if there is an insufficient number of tertiary ICUs to satisfy the sample size requirement, or for practical reasons or generalisability concerns, it may not be appropriate to restrict a trial to only tertiary ICUs. While the overall sample size requirement does increase as the fraction of tertiary ICUs is reduced, this may result in an unfeasible trial becoming feasible, as shown in the example in Appendix IV.

Discussion

We have discussed how to perform a sample size calculation for a two-period, two-intervention, cross-sectional CRXO trial in the intensive care setting. We have also provided estimates of the elements required to perform these calculations using the ANZIC-APD data. A review of the statistical methods used to determine the sample size for CRXO trials found that the methodology was frequently inadequate.[12] Inadequate methods may have been used because of lack of knowledge of the appropriate sample size methodology, limited availability of the elements needed to perform the sample size calculations or because of a lack of practical examples of how to implement a sample size formula. We have addressed these issues in this article.

We have provided a sample size formula to determine the number of ICUs required to detect a constant *relative* risk, rather than a constant *absolute* risk reduction, between two interventions. When considering multiple strata, each with their own baseline mortality (event) rate, evidence suggests that it is more plausible to expect a constant relative reduction rather than absolute reduction.[13]

When the sample size calculation accounts for multiple ICU strata, the same power can be achieved with differing numbers of ICUs from each stratum. For example, a trial can be designed with an equal contribution of ICUs from each stratum, or designed so that one stratum provides most of the required ICUs. In the ANZICS-APD, when the intervention is aimed at reducing all-cause mortality, the absolute minimum number of total ICUs will be obtained by including only tertiary ICUs. However, stratifying the sample size calculation provides flexibility to adjust the required number of ICUs from each stratum when the number of ICUs is limited by the availability of ICUs in the tertiary (or other) stratum.

Additional considerations

We have considered the two-intervention, two-period CRXO design with different patients in each period. Adaptations of this design, such as increasing the number of periods[14] or including the same participants across periods,[15] are possible, but they also increase the complexity of the sample size calculation and analysis.[16]

The sample size calculation is sensitive to the difference between the WPC and the BPC, therefore both correlations should be chosen carefully.[7] Considerations for choosing the WPC are similar to those for choosing an intracluster correlation for a parallel-group, cluster randomised trial.[17-20] Choosing the BPC is likely to rely on routinely collected data, pilot and feasibility data, or a reasoned best guess. However, in the absence of such data, recommendations of half the WPC[21] and 0.8 of the WPC have been made.[22]

The intended analysis should match the sample size methodology. One potentially appropriate analysis method to estimate the RR in a stratified CRXO trial is to use generalised estimating equations (GEE) for a binomially distributed outcome, with a logarithmic link and an exchangeable "working" correlation between individuals within an ICU. The ICU strata can be included as a covariate in the model, as can a term for period effects.

There is an active field of research investigating how well statistical analysis methods perform in cluster randomised trials with small numbers of clusters or low event rates. Our examinations displayed appropriate performance for an RR of 0.9 and two strata with event rates of 7% and 9%. However results from other research has shown that caution may need to be exercised with power formulae with fewer than 12 clusters[23] or when the event rates are 6% or lower.[9]

Conclusion

Sample size determination for CRXO trials requires the use of an appropriate sample size formula together with appropriate estimates of its component elements. We have provided the sample size formulae, estimates of the elements required by the formulae using ANZICS-APD data, and examples of how to determine the required sample size for unstratified and stratified CRXO trials in the intensive care setting.

Author details

Sarah J Arnup[1]
Joanne E McKenzie[1]
David Pilcher[2,3,4]
Rinaldo Bellomo[4]
Andrew B Forbes[1]

1 School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia.
2 Centre for Outcome and Resource Evaluation, Australian and New Zealand Intensive Care Society, Melbourne, VIC, Australia.
3 Department of Intensive Care, The Alfred Hospital, Melbourne, VIC, Australia.

4 Australian and New Zealand Intensive Care Research Centre, School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia.

Correspondence: andrew.forbes@monash.edu

References

1 Aberegg SK, Richards DR, O'Brien JM. Delta inflation: a bias in the design of randomized controlled trials in critical care medicine. *Crit Care* 2010; 14: R77.

2 Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester, United Kingdom: John Wiley and Sons, 2012.

3 Bellomo R, Forbes A, Akram M, Bailey M, Pilcher DV, Cooper DJ. Why we must cluster and cross over. *Crit Care Resusc* 2013; 15: 155-157.

4 Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med* 2007; 26: 274-289.

5 Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. *Contemp Clin Trials* 2007; 28: 316-323.

6 Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. *Trials* 2009; 10: 27.

7 Arnup SJ, McKenzie JE, Hemming K, Pilcher D, Forbes AB. Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. *Trials* 2017; 18: 381.

8 Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract* 2000; 17: 192-196.

9 Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. *Clin Trials* 2015; 12: 34-44.

10 Jones B, Kenwood MG. Design and analysis of cross-over trials. 2nd edition. London: Chapman and Hall/CRC, 2003.

11 Stow PJ, Hart GK, Higlett T, et al. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. *J Crit Care* 2006; 21: 133-141.

12 Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *J Clin Epidemiol* 2016; 74: 40-50.

13 Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002; 21: 1575-1600.

14 Connolly SJ, Philippon F, Longtin Y, et al. Randomized cluster crossover trials for reliable, efficient, comparative effectiveness testing: design of the Prevention of Arrhythmia Device Infection Trial (PADIT). *Can J Cardiol* 2013; 29: 652-658.

15 Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. *J Educ Behav Stat* 2011; 36: 472-490.

16 Kasza J, Hemming K, Hooper R, et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res* 2017; Jan 1: 962280217734981. doi: 10.1177/0962280217734981 [Epub ahead of print].

17 Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* 2005; 2: 99-107.

18 Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol* 2005; 58: 246-251.

19 Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *Am J Epidemiol* 1999; 149: 876-883.

20 Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004; 57: 785-794.

21 Donner A, Klar N, Zou G. Methods for the statistical analysis of binary data in split-cluster designs. *Biometrics* 2004; 60: 919-925.

22 Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015; 350: h2925.

23 Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Stat Med* 2017; 36: 318-333. ❑

Chapter 7

Summary and conclusions

7.1 Introduction

New statistical methods are constantly being developed and existing methods refined to address complexities in data and developments in study design. While there are early examples of use of the CRXO design, dating back to 1974, only in recent years has there been substantial uptake of the design. As such, the methods for the design and analysis have not been fully developed. Without such developments, estimates of interventions obtained from trials using the design may be misleading because they may potentially be biased, expressed with incorrect uncertainty or both. Guidance on the implementation of the methods, and reporting, may also require development. The aim of the research presented in this thesis was to assess the design, reporting quality, and statistical methods used in CRXO trials that evaluate clinical and public health interventions, and subsequently provide tools that may facilitate improvement in areas where shortcomings were identified.

In order to examine the design, reporting and statistical methods used to date, a systematic review was undertaken. To facilitate improvement in the identified shortcomings, the following was developed. First, the content for a CRXO trial reporting guideline was proposed, which included potential modifications to existing guidelines and identification of areas where reporting items may need to be developed. Second, an intuitive graphical approach was developed to demonstrate the effect of the cluster randomisation and multiple period aspect of the design on the required sample size. Development of a graphical representation aimed to provide greater accessibility of the concepts to a more general research audience. Third, guidance was provided for sample size calculation, including: recommendations for choosing the parameters required in the sample size formula; estimates of the parameters required for sample size calculation in intensive care research; and worked examples of sample size calculation tailored to both health researchers working in intensive care research and applied biostatisticians. Finally, the existing sample size formulae were extended to settings where the values of the parameters required by the sample size formulae vary between strata, such as hospital type or geographical location. The results from this research contribute to an improved understanding of the CRXO design, the design and analysis of future CRXO trials, and reporting of CRXO trials; and hence provide a technical

foundation for more reliable evidence for health care decision making, particularly in the intensive care setting.

This chapter summarises the key findings from each of the chapters, discusses overall findings, and concludes with proposals for future research.

7.2 Summary of thesis chapters

7.2.1 Chapters 2 and 3: A protocol and systematic review of CRXO trials

The protocol for a systematic review of published CRXO trials was presented in Chapter 2. The results of the systematic review were presented in Chapters 3 and 4. The aim of the review was to summarise and evaluate the characteristics of the reported trials, the motivations for using the CRXO design, the values of the CRXO design parameters, the justification and methodology used for the sample size calculation and analyses, and the quality of reporting the CRXO design aspects (presented in Chapter 4).

Key findings from this research included the following:

- The most common type of cluster was intensive care units.
- The two-period two-group cross-sectional design was the most commonly used design, but more complex multiple period and cohort designs were also used.
- Designs very commonly used only a small number of clusters (median 9 [IQR: 4–21]), and the size of the clusters were generally small (median 30 [IQR 14–77]). Statistical methods for sample size calculation and analysis for CRXO designs with small numbers of clusters, of small cluster size, have not been proposed or evaluated in the published literature.
- Justification provided by researchers for both the cluster randomisation and multiple period aspects of the CRXO design were commonly not reported, despite justification for design choice being a recommended reporting item for randomised trials [14].
- The methodology used to determine the sample size was commonly not reported. The reported methods were frequently inadequate for the trial design. Cluster randomisation was often accounted for, but most calculations did not account for the multiple period aspect of the design, leading to potentially conservative estimates of the required number of participants.
- Nearly all analyses were performed at the individual participant level. Only 7% of analyses were performed at the level of the cluster.
- The methodology was usually appropriate when the analysis was performed at the cluster level. Analyses undertaken using individual-level data commonly failed to account for both the cluster randomisation aspect and multiple period aspect of the design, or accounted for only the cluster randomisation aspect. In both cases, the

precision of the estimate of the intervention effect is too high, potentially leading to false positive claims regarding the effectiveness of the intervention.

7.2.2 Chapter 4: A proposal for reporting items and an assessment of reporting quality for CRXO trials

The aims of the research presented in Chapter 4 were to examine the reporting quality of published CRXO trials and to stimulate discussion around the content of a reporting guideline. Possible reporting items for CRXO trials were proposed to achieve these aims.

Key findings from this research included the following:

- Incomplete reporting of the design aspects that are unique to the CRXO design was common, including the identification of trial design in the title and abstract; provision of a rationale for both the cluster randomisation and crossover aspects of the design; discussion of the potential for and minimisation of carryover; parameters used in determining the sample size; and the values of within-cluster within-period ICC and within-cluster between-period ICC observed in the outcome data.

- Approaches to reporting baseline data, the flow of participants through the trial, and other trial characteristics is challenging and requires further discussion. The presentation of the baseline data and flow diagram will depend on whether the trial uses cohort or cross-sectional design.

- There is a need for a CONSORT statement extension to provide specific guidance on reporting for CRXO trials.

7.2.3 Chapter 5: A graphical illustration of the components of variation from the CRXO design and a sample size tutorial

The infrequent use of appropriate sample size methodology in published CXRO trials was the motivation for the remaining chapters in this thesis. The aims of the research presented in Chapter 5 were to facilitate researchers' understanding of the design by providing graphical illustrations of the effect of the CRXO design aspects on sample size calculation, and to provide guidance for determining the required sample size using previously published formulae.

Key findings from this research included the following:

- A graphical illustration that showed the intuition behind the key parameters in the CRXO design was developed. Illustrations were used to show how the cluster randomisation and crossover aspects of the design give rise to similarity in the responses of individual participants both within the same cluster, and within the same cluster period; and the relationship between the components of variation in the responses of individual participants and the within-cluster ICCs.

- The illustrations showed how the precision of the estimate of the intervention effect in a CRXO trial depends on the difference between the within-cluster within-period ICC and within-cluster between-period ICC.

- The relationship between the difference in the within-cluster within-period ICC and within-cluster between-period ICC, and the required sample size for a CRXO trial, was explained. It was shown that when the within-cluster within-period ICC and within-cluster between-period ICC were equal, the precision gained by crossover aspect of the CRXO design equalled the precision lost by cluster randomisation and the CRXO design had similar precision to an individually randomised design. But when the within-cluster between-period ICC was zero there was no advantage in a CRXO trial over a parallel-group cluster randomised trial.

- The examples of sample size calculation demonstrated that the sample size was sensitive to small changes in values of the within-cluster ICCs, except when cluster sizes were large and the difference between the within-cluster within-period and within-cluster between-period ICCs was constant.

7.2.4 Chapter 6: Sample size calculations for CRXO trials in Australian and New Zealand intensive care research

The aim of the research presented in Chapter 6 was to extend the sample size formulae to stratified trial designs and provide guidance specifically for health researchers designing trials in the intensive care setting.

Key findings from this research included the following:

- The parameters required for sample size calculations in the Australian and New Zealand intensive care setting varied according to hospital type (tertiary compared to metropolitan and rural).

- When clusters varied according to factors that affected the trial outcome, stratifying the sample size formula by those factors led to a reduction in the number of clusters required to detect a constant risk ratio, as compared with the unstratified sample size formula.

- Using data from the Australian and New Zealand Intensive Care Society Adult Patient Database clinical registry, it was shown that the stratified sample size formula resulted in trial designs that had reasonable (>80%) power to detect small effect sizes with the available number of ICUs. Without stratification, it was not possible to design a trial to detect the same effect size with the finite number of available ICUs.

- The stratification in the sample size formula provided flexibility to adjust the required number of ICUs from each stratum when the number of ICUs was limited in one stratum.

7.3 Overall discussion

The central aim of this thesis is to facilitate improvement in the understanding, design, reporting and statistical methods used in CRXO trials. The key findings from this thesis are that the statistical methods used for the sample size calculation and analysis of CRXO trials rarely accounted appropriately for the CRXO design, and that reporting of the design and analysis aspects unique to CRXO trials was frequently incomplete. In this section, issues relating to facilitating improvement in these areas are discussed.

The first shortcoming identified in this thesis is that the statistical methods employed in CRXO trials rarely accounted appropriately for the design. Four possible explanations for the use of inappropriate methods are: 1) limited understanding amongst health researchers of the need to use specialist methods for the sample size calculation and analysis of CRXO trials; 2) lack of recognition of the availability of existing sample size and analysis methodology, and knowledge of how to implement these methods; 3) limited availability of the typical magnitude of the parameters required to perform a sample size calculation, and uncertainty in the meaning of the parameters; and 4) lack of appropriate sample size calculation and analysis methodology. Each of these explanations is now considered.

Limited understanding of the components of variation induced by the CRXO design may lead to the use of inappropriate methodology. The tutorials presented in this thesis aim to assist health researchers in employing appropriate statistical methodology by increasing their understanding of the design. The development of a CONSORT statement extension for CRXO trials may increase researchers' recognition of the methodological requirements for this design. While reporting guidelines aim to improve the reporting of trial results, they can provide an awareness of the unique methodological aspects of a particular design, which health researchers can consider in the planning process. The work presented in this thesis underpins the development of a reporting guideline for CRXO trials (Chapter 4).

The use of inappropriate methods may also occur because health researchers are unaware that methodology exists for CRXO trials. Statistical methods for performing sample size calculations were first published in 2008 [26] and for analysis in 2007 [5][6], but the research in this thesis indicates that these methods are rarely used in practice (Chapter 3). These methods were published in journals whose target audience is biostatisticians, and may therefore not reach the readership of health researchers. Also, health researchers may find the literature inaccessible and be uncertain how to implement the published methods. The publications presented in this thesis attempt to bridge the gap between statisticians and health researchers by providing worked examples of the sample size formulae. Chapters 5 and 6 contain published Stata do-files that calculate the required sample size and power for a CRXO trial. In addition, an online application has recently become available to estimate the design effect, power or variance for CRXO trial [33]. Additional publications illustrating appropriate analysis methods, and the development of software packages that perform the calculations, may facilitate increased use of appropriate methodology.

Limited understanding of the meaning of the parameters required by the sample size formulae, and limited availability of estimates of the parameters, may lead to the use of

inappropriate sample size methodology. In these scenarios, health researchers may choose inappropriate estimates of the required parameters, or employ formulae appropriate for different trial designs, such as formulae for parallel-group cluster randomised designs. These issues are addressed in multiple ways in this thesis. Parameter estimates are provided for trials evaluating the effect of routine interventions on all-cause mortality in the Australian and New Zealand intensive care setting. Proposed items to be included in the CONSORT CRXO extension aim to achieve reporting of the values of parameters required for sample size calculation. Finally, an understanding of the parameters in the sample size formulae is essential, because the parameters are dependent on the characteristics of the individual trial. This thesis provides guidance and worked examples of the steps required to obtain appropriate estimates. In the absence of data or published estimates to inform trial design, it is hoped that these measures will enable health researchers to determine a conservative, but appropriate, sample size estimate.

A final explanation for the use of inappropriate methodology is that appropriate methodology had not been established for the trial design. Publications were identified for inclusion in the systematic review during 2014, and at that time, analysis methodology had been published only for two-period two-group cross-sectional designs, for continuous and binary outcomes [5][6][38]. For sample size calculations, only continuous outcomes had been considered [26]. These methods were appropriate for many of the trials included in the systematic review, yet almost none of these trials employed the published methods. This suggests that even when appropriate methodology exists, other factors, such as those discussed in the previous paragraphs, inhibit the use of these methods.

Many of the trials included in the systematic review used more complex designs than those investigated in the published methodology. The design of these trials can provide motivating examples for future methodological work. Indeed, two papers have recently addressed sample size calculation and analysis of designs with more than two periods [33][40]. The discussion of future methodological development is continued in the future work section.

The second shortcoming identified by this thesis is that the reporting of CRXO trials is frequently incomplete. This is a common finding across randomised controlled trials [2][43][50][51][52][53]. The CONSORT statement and its extensions were developed to improve the quality of reporting for randomised controlled trials. The work presented in this thesis showed that there is a need for an extension to CRXO trials. Because no published reporting guidelines exist for CRXO trials, the research presented in this thesis simultaneously i) proposed possible reporting items, and indicated areas where items may need to be developed; and ii) used these proposed items to determine the need for a separate guideline. The recommended next steps include setting up a consensus process, including participants with relevant expertise, to decide upon the specific items and their wording [42]. The CONSORT statement extension for CRXO trials is under development with the first Delphi and consensus meeting having taken place in August 2018. The reporting items proposed in this thesis informed the item development that took place in this meeting.

Despite the introduction of the CONSORT statement for parallel-group randomised

trials and extensions, there has only been modest improvement in reporting quality and reporting quality remains below an acceptable level [2][51][54]. Journal endorsement of the CONSORT statement may improve reporting [55][56][57]. However, such improvement may be minimal without the addition of editorial policies that require authors to comply with the guidelines as a condition of publication [58]. In a review the online 'Instruction for Authors' in high impact medical journals, as few as 14% required authors to comply with guidelines as a condition of publication [59]. To bring about improvement in reporting quality, there remains an urgent need to evaluate strategies to encourage compliance with reporting guidelines at both the level of the journal editors and the publication authors.

7.4 Future research

In addition to the recommendations for future research presented in each of the chapters, the following suggestions are presented:

- The identified characteristics of trials using the CRXO design can be used to inform future methodological development. The work presented in this thesis found that the following trial design aspects were common: three interventions, more than two periods, periods of variable time length, and cohort designs. There has been limited development and evaluation of analysis and sample size methods for such designs [33][39][40].

- The performance of published methods for CRXO trials with small numbers of clusters has not been extensively studied. The median number of clusters in CRXO trials was 9 [IQR: 4–21]. Based on recommendations for parallel-group cluster randomised trials, trials using either generalised estimating equation or generalised linear model methodology with robust standard errors might be appropriate with at least 30 clusters [30]. However, recent research in CRXO designs does not provide such clear-cut recommendations. A random effects model may require more than 30 clusters [32], but with generalised estimating equations a lesser number may suffice [53]. The recommendations for both analysis methods were based on results from numerical simulations, and therefore further work is required to clarify the distinctions between the simulations and results.

- There has been limited evaluation of methods used to determine the sample size for CRXO trials [24][36][39]. Prior research suggests that caution needs to be exercised when the endpoint event rates are low (less than 6%) [24]. This work was for an unstratified CRXO design however, further work is required to determine recommendations for the stratified design.

- The tutorial publications presented in this thesis focussed on improving understanding of the components of variation in the CRXO design and performing appropriate sample size calculations. Tutorial publications that address the use of inappropriate analysis methodology may also improve the use of appropriate statistical analyses for

this design. As future methodological research establishes best practise for analysis, tutorials may help translate this research into practice.

- Identification of biases that may be unique to the CRXO design is warranted. Knowledge of these biases can then allows health researchers to minimise the impact of the biases when planning trials. As an example of bias introduced by the trial design, consider a cross-sectional design where ICUs form the clusters and the intervention is administered for the duration of participants' stay in ICU. A participant may stay in the ICU across multiple periods, especially if admitted near the time of a pre-planned crossover. The participant may initially receive one intervention, but still be staying in the ICU when the intervention assigned to the cluster is switched. In this situation, the participant could either continue to receive the initial intervention, or cross over to the subsequent intervention. Both options have the potential to introduce bias, and require further consideration. One strategy to minimise the possibility of this situation occurring is to use an appropriate washout period between periods or to exclude patients within a certain time period prior to the crossover, for example, the final week of a one month period.

- It is important to evaluate the effectiveness of publications that aim to improve the use of a trial design. Such evaluation provides opportunity to refine strategies that did not result in improvement and increased use of strategies that did lead to improvement. The research conducted for this thesis led to five publications that aimed to improve the use of the CRXO design, and informed item development for a CONSORT statement extension for CRXO trials. Once the CONSORT extension is complete, the impact of this on the use of the CRXO design could be assessed by performing a subsequent review. The method used to assess the quality of reporting in stepped-wedge cluster randomised trials following the recently published CONSORT statement extension for stepped-wedge cluster randomised trials could be employed [60].

- The published papers from this thesis motivated the inclusion of the CRXO design in the Cochrane handbook for systematic reviews of interventions. Further work is needed to develop an extension for the Cochrane risk of bias tool to include the CRXO design.

- This thesis focussed on the use of the CRXO design in trials that evaluate health interventions. However the CRXO design could feasibly be used in other disciplines, including agricultural and environmental sciences, social, behavioural and educational sciences. While each discipline will face specific challenges when implementing the design, accounting for the cluster randomisation and cross over aspects of the design in the design and analysis stages will always be required. Therefore, investigation into the use of the CRXO design in other disciplines may suggest strategies to overcome the shortcomings found in health research. In particular, novel methods for sample size calculation and analysis may already be in use in other disciplines, and the appropriateness of methods currently used in health research may have been evaluated

in other contexts. A broad cross discipline systematic review of methodology and trials may therefore be of value.

References

[1] Grimes DA, Schulz KF: **An overview of clinical research: the lay of the land.** *Lancet* 2002, **359**(9300):57-61.

[2] Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG: **The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed.** *BMJ* 2010, *340*:c723.

[3] Eldridge S, Kerry S: *A Practical Guide to Cluster Randomised Trials in Health Services Research.* UK: John Wiley & Sons, Ltd; 2012.

[4] Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG: **Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.** *Health Technol Assess* 1999, **3**(5):iii-92.

[5] Parienti JJ, Kuss O: **Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies.** *Contemp Clin Trials* 2007, **28**(3):316-23.

[6] Turner RM, White IR, Croudace T, Group PIPS: **Analysis of cluster randomized cross-over trial data: a comparison of methods.** *Stat Med* 2007, **26**(2):274-89.

[7] Hutton, JL: **Are distinctive ethical principles required for cluster randomized controlled trials?** *Stat Med* 2001, **20**(3):473-88.

[8] Eldridge S, Ukoumunne OC, Carlin JB: *The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. Int Stat Rev* 2009, **77**(3): 378-94.

[9] Donner A, Birkett N, Buck C: **Randomization by cluster. Sample size requirements and analysis.** *Am J Epidemiol* 1981, **114**(6):906-14.

[10] Higgins JPT, Green S, editors: *Cochrane handbook for systematic reviews of interventions.* Version 5.1.0 [updated March 2011]. Oxford: The Cochrane Collaboration, 2011. Available from http://handbook.cochrane.org/.

[11] Giraudeau B, Caille A, Le Gouge A, Ravaud P: **Participant Informed Consent in Cluster Randomized Trials: Review.** *PLoS One* 2012, **7**(7), e40436.

[12] Giraudeau B, Ravaud P: **Preventing bias in cluster randomised trials.** *PLoS Med* 2009, **6**(5), e1000065.

[13] Puffer S, Torgerson D, Watson J: **Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals.** *BMJ* 2003, **327**(7418): 785-9.

[14] Schulz KF, Altman DG, Moher D, Group C: **CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials.** *J Clin Epidemiol* 2010, **63**(8):834-40.

[15] Edwards SJ, Braunholtz DA, Lilford RJ, Stevens AJ: **Ethical issues in the design and conduct of cluster randomised controlled trials.** *BMJ* 1999, **318**(7195):1407-9.

[16] Eldridge SM, Ashby D, Feder GS: **Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care.** *Clin Trials* 2005, **2**(2):91-8.

[17] Lignou S: **Informed consent in cluster randomised trials: new and common ethical challenges.** *J Med Ethics* 2018, **44**(2):114-20.

[18] Weijer C, Grimshaw JM, Eccles MP, McRae AD, White A, Brehaut JC, Taljaard M: **The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials.** *PLoS Med* 2012, **9**(11): e1001346.

[19] Hemming K, Taljaard M, Forbes G, Eldridge SM, Weijer C: **Ethical implications of excessive cluster sizes in cluster randomised trials.** *BMJ Qual Saf* 2018, **27**:664-70.

[20] Hills M, Armitage P: **The two-period cross-over clinical trial.** *Br J Clin Pharmacol* 1979, **8**(1):7-20.

[21] Jones B, Kenwood MG: *Design and analysis of cross-over trials.* 2nd ed. London: Chapman & Hall/CRC; 2003.

[22] Rietbergen C, Moerbeek M: **The design of cluster randomized crossover trials.** *J Educ Behav Stat* 2011, **36**(4):472-90.

[23] Bellomo R, Forbes A, Akram M, Bailey M, Pilcher DV, Cooper DJ: **Why we must cluster and cross over.** *Crit Care Resuscitation* 2013, **15**(3):155-7.

[24] Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R: **Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care.** *Clin Trials* 2015, **12**(1):34-44.

[25] Reich N, Milstone A: **Improving efficiency in cluster-randomized study design and implementation: taking advantage of a crossover.** *Open Access J Clin Trials* 2014, **6**:11-5.

[26] Giraudeau B, Ravaud P, Donner A: **Sample size calculation for cluster randomized cross-over trials.** *Stat Med* 2008, **27**(27):5578-85.

[27] Louis TA, Lavori PW, Bailar JC 3rd, Polansky M: **Crossover and self-controlled designs in clinical research.** *N Engl J Med* 1984, **310**(1):24-31.

[28] Aberegg SK, Richards DR, O'Brien JM: **Delta inflation: a bias in the design of randomized controlled trials in critical care medicine.** *Crit Care* 2010, **14**(2):R77.

[29] Donner A, Klar N: *Design and Analysis of Cluster Randomized Trial in Health Research.* UK: John Wiley & Sons, Ltd; 2000.

[30] Hayes RJ, Moulton LH: *Cluster randomised trials.* 2nd Ed. Boca Raton, FL, USA: Chapman and Hall/CRC; 2017.

[31] Campbell MJ, Walters SJ: *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research.* UK: John Wiley & Sons, Ltd; 2014.

[32] Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC: **Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome.** *Stat Med* 2017, **36**(2):318-333.

[33] Kasza, J, Forbes, A: **Information content of cluster-period cells in stepped wedge trials.** *Biometrics* 2019, DOI: 10.1111/biom.12959.

[34] Liang K-Y, Zeger SL: **Longitudinal data analysis using generalized linear models.** *Biometrika* 1986, **73**(1):13-22.

[35] Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, Wolfson M: **A comparison of two bias-corrected covariance estimators for generalized estimating equations.** *Biometrics* 2007, **63**(3):935-41.

[36] Li F, Forbes AB, Turner EL, Preisser JS: **Power and sample size requirements for GEE analyses of cluster randomized crossover trials.** *Stat Med* 2019, **38**(1):636-49.

[37] Rutterford C, Copas A, Eldridge S: **Methods for sample size determination in cluster randomized trials.** *Int J Epidemiol* 2015, **44**(3):1051-67.

[38] Donner A, Klar N, Zou G: **Methods for the statistical analysis of binary data in split-cluster designs.** *Biometrics* 2004, **60**(4):919-25.

[39] Hooper R, Teerenstra S, de Hoop E, Eldridge S: **Sample size calculation for stepped wedge and other longitudinal cluster randomised trials.** *Stat Med* 2016, **35**(26):4718-28.

[40] Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB, ANZICS Centre for Outcomes Resource Evaluation Committee: **Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials.** *Stat Methods Med Res* 2019, **28**(3):703-16.

[41] Taljaard M, Teerenstra S, Ivers NM, Fergusson DA: **Substantial risks associated with few clusters in cluster randomized and stepped wedge designs.** *Clin Trials* 2016, **13**(4):459-63.

[42] Moher D, Schulz KF, Simera I, Altman DG: **Guidance for developers of health research reporting guidelines.** *PLoS Med* 2010, **7**(2):e1000217.

[43] Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S: **Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review.** *J Clin Epidemiol* 2015, **68**(6):716-23.

[44] Campbell MK, Fayers PM, Grimshaw JM: **Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research.** *Clin Trials* 2005, **2**(2):99-107.

[45] Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ: **Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data.** *J Clin Epidemiol* 2005, **58**(3):246-51.

[46] Gulliford MC, Ukoumunne OC, Chinn S: **Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994.** *Am J Epidemiol* 1999, **149**(9):876-83.

[47] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ: **Patterns of intra-cluster correlation from primary care research to inform study design and analysis.** *J Clin Epidemiol* 2004, **57**(8):785-94.

[48] Hooper R, Bourke L: **Cluster randomised trials with repeated cross sections: alternatives to parallel group designs.** *BMJ* 2015, *350*:h2925.

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

[50] Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG: **Design, analysis, and presentation of crossover trials.** *Trials* 2009, **10**:27.

[51] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, Grimshaw JM, Weijer C, Zwarenstein M, Donner A:

Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ* 2011, **343**:d5886.

[52] Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ: **Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation.** *J Clin Epidemiol* 2011, **64**(9):936-48.

[53] Li T, Yu T, Hawkins BS, Dickersin K: **Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis.** *PLoS One* 2015, **10**(8): e0133023.

[54] Richardson M, Garner P, Donegan S:**Cluster Randomised Trials in Cochrane Reviews: Evaluation of Methodological and Reporting Practice.** *PLoS One* 2016, **11**(3): e0151818.

[55] Moher D, Jones A, Lepage L: **Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation.** *JAMA* 2001, **285**(15):1992-1995.

[56] Devereaux PJ, Manns BJ, Ghali WA, Quan H, Guyatt GH: **The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the consolidated standards of reporting trials (CONSORT) checklist.** *Control Clin Trials* 2002, **23**(4):380-8.

[57] Turner L, Shamseer L, Altman DG, Schulz KF, Moher D: **Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review.** *Syst Rev* 2012, **1**:60.

[58] Hopewell S, Ravaud P, Baron G, Boutron I: **Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis.** *BMJ* 2012, **344**:e4178

[59] Hopewell S, Altman DG, Moher D, Schulz KF: **Endorsement of the CONSORT statement by high impact factor medical journals: a survey of journal editors and journal "instructions to authors".** *Trials* 2008, **9**:20.

[60] Hemming K, Carroll K, Thompson J, Forbes A, Taljaard M: **Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review** *J Clin Epidemiol* 2019, **107**:77-88.

Appendix A

Additional file 1 accompanying Chapter 2 - Systematic review data extraction form

CRXO systematic review data extraction form

Date: 15 May 2014

The unit of analysis for the review is ***study or trial***, not article. In many cases a study will be split into multiple articles, i.e. a protocol or design article, an article reporting the primary outcome(s), and many other articles reporting secondary outcomes.

The primary outcome for the study will be defined from the following hierarchy:

- The first primary outcome in the protocol document or first published paper for the study if there is no protocol document.
- The outcome used for the sample size calculation.
- The first outcome listed in the methods section of the abstract.

Section 1: Study Identifiers

Study ID	(Autocompleted)
Is the article a protocol paper?	0=No, 1=Yes
First Author Surname	String
Publication Year	Integer, 1946 to 2014
Journal	Categorical
Reviewer's initials	Categorical
Date of review	Date
Notes	Not to be analysed

Section 2: Full Text Screening

Does the article report a research trial that used or planned to use a CRXO design which incorporated the following design elements (*The article will only be marked for inclusion if "yes" is answered to all):*

Outcomes were measured on humans in a study or trial at either cluster or individual level.	(0=No, 1=Yes, 2=Unclear)
Allocation of the intervention was at cluster level (The allocation does not have to be at random).	(0=No, 1=Yes, 2=Unclear)
Each cluster received each intervention, or at least some clusters crossed over from one intervention to another (e.g. two-intervention-four-sequence designs AA, AB, BA, BB).	(0=No, 1=Yes, 2=Unclear)
Each cluster received each intervention in a sequence over time, rather than concurrently in time.	(0=No, 1=Yes, 2=Unclear)

Section 3: Title and Abstract

Rationale: To assess how CRXO trials are identified in the title and abstract.

Title	
Is the trial identified as a **cluster** randomised **crossover** trial in title? (*note words 'cluster' and 'crossover' must be used, placement of hyphens is unimportant*)	(0=No, 1=Yes)
Abstract	
Is the trial identified as a **cluster** randomised **crossover** trial in abstract? (*note words 'cluster' and 'crossover' must be used, placement of hyphens is unimportant*)	(0=No, 1=Yes)
If no, copy verbatim from abstract how the unit of randomisation was described in the abstract	TEXT
If no, copy verbatim from abstract how the cross over of interventions was described in the abstract	TEXT

Section 4: Justification for the CRXO design

Rationale: To understand why researchers are using the CRXO design and how they justify that decision.

Why was the CRXO design chosen? *For each of the following points enter (0=Not Discussed, 1=Yes, 2=Unclear). Select as many points as apply.*

*Justification given by authors for **cluster** randomisation*	
Intervention can **only** act at the cluster level, and therefore impossible to randomise individually *(e.g. if the intervention is an educational program for health care practitioners, or a program implemented publicly via radio or newspaper, the intervention will reach a group of people)*.	(0=Not Discussed, 1=Yes, 2=Unclear)
Practical/ethical/cost/administrative difficulties with randomising at an individual level.	(0=Not Discussed, 1=Yes, 2=Unclear)
Contamination likely between participants at the level of person/people delivering the intervention *(e.g. an educational intervention may be delivered to health care practitioners, and it may impossible for them to only apply the intervention to some individuals in their care and not others. Therefore contamination would occur in an individually randomised trial).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Contamination likely between participants in a cluster *(e.g. a behavioural intervention may be*	(0=Not Discussed, 1=Yes, 2=Unclear)

delivered to schools, and it may be impossible to prevent primary caregivers from exchanging experiences, thereby contaminating each arm of the trial in an individually randomised trial).	
To ensure intervention is fully delivered *(if it is expected that compliance with the trial protocol will be reduced if members of a cluster were individually randomised)*	(0=Not Discussed, 1=Yes, 2=Unclear)
Outcome data only available at cluster level	(0=Not Discussed, 1=Yes, 2=Unclear)
Other, specify	TEXT
*Justification given by authors for **crossover** design*	
Increased efficiency to overcome loss of power through randomising in clusters (*i.e. the authors specifically cite a reduction in precision/power due to cluster randomisation or the 'design effect' as the reason for the crossover element).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Reduced efficiency in estimating the intervention effect due to a limited number of clusters available for inclusion in the trial *(i.e. the authors cite the limited number of clusters as the reason for the crossover element).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Clusters are expected to have very different characteristics from each other (*i.e. the authors cite that they expect or wish to allow for clusters being very different in characteristics which might affect the outcome, and wish to crossover so that each cluster 'acts as own control')*	(0=Not Discussed, 1=Yes, 2=Unclear)
Other, specify	TEXT

Section 5: Trial Objectives

Rationale: What are the levels of the primary objective being addressed with the CRXO trial?

Copy **verbatim** objective or hypothesis from Introduction	TEXT
Is the primary objective at the cluster level, individual level, or both	1. Cluster level 2. Individual level 3. Both 4. Not stated 5. Not clear, explain

Section 6: Population Details

Rationale: What settings are CRXO trial being used in?

Disease or domain under study (can select multiple)	1. Cancer 2. Cardiovascular 3. Central nervous system/musculoskeletal 4. Digestive/endocrine 5. Nutritional and metabolic 6. Gynaecology 7. Pregnancy and birth 8. Infectious diseases 9. Mental health and behavioural conditions 10. Pathological conditions 11. Symptoms and signs 12. Respiratory disease 13. Urogenital 14. Blood and immune system 15. Ear and nose 16. Eye 17. General health 18. Genetic disorders 19. Injuries 20. Accidents and wounds 21. Mouth and dental 22. Skin 23. Other
Country of trial (List all if 5 or less, otherwise state multinational)	TEXT
Setting (select one)	1. Primary care practices/health care clinics 2. Communities/geographical areas 3. Households/families 4. Aged care facility 5. Hospital 6. Schools 7. Workplaces 8. Other,specify
Do the methods define the cluster unit?	(0=No, 1=Yes, 2=Description unclear)
Clusters receiving intervention (select one)	1. Primary care practices (practice includes multiple health care professionals) 2. Individual health professional 3. Communities/Residential areas 4. Households/families 5. Hospital, specify unit/ward type 6. Nursing home/aged care 7. Schools 8. Worksites 9. Other, specify
Additional comments about clusters receiving intervention	TEXT

Section 7: Study Design

Rationale: This section is intended to capture both the key design features of the published CRXO trial and how the design features are reported.

An answer of "yes" means that the design aspect could be reconstructed from the information provided in the article.

The number of participating clusters	Integer, 99 = Not reported
The number of periods	Integer, 99 = Not reported
The number of interventions	
Intervention treatments (*active interventions*)	Integer, 99 = Not reported
Control treatments (*e.g. no treatment, usual care. Enter 0 if all interventions are active*)	Integer, 99 = Not reported
List the different unique intervention sequences, (*i.e. AB, BA; or AA, BB, AB, BA*. *Copy* ***verbatim*** *from text*)	TEXT
Do the authors ***discuss*** how many interventions each ***participant*** will receive, i.e. if the participant can remain in cluster for longer than one period?	(0=Not Discussed, 1=Yes, 2=Unclear)
Is each period ***designed*** to include the same or different participants? (*i.e. are measurements repeated or not repeated on participants?)*	(0=Same participants, 1=Different participants, 2=Not Stated)
Are there any other relevant design features that may lead to additional correlation within the outcomes? E.g. hierarchical designs where there is clustering at different levels; wards within hospitals, GPs within general practices. (*copy* ***verbatim*** *from text*)	TEXT

Section 8: Carry over

Rationale: To describe whether the risk of carry over being is acknowledged and managed.

Do the authors ***discuss*** the possibility of carry over of intervention effects between periods?	(0=Not Discussed, 1=Yes, 2=Unclear*),* Page and paragraph number
Do the methods ***detail*** how the risk of carry over effects will be minimised by the study design *(e.g. washout period, different subjects in each period)*	(0=No, 1=Yes-Sufficient to replicate, 2=Yes-Insufficient to replicate description, 3=NA-Carry over not possible) Copy in text verbatim on the details.

Section 9: Blinding, bias and consent

Rationale: To understand how randomisation or allocation of interventions was performed, the risk of bias in CRXO trials, and the adequacy of reporting.

An answer of "yes" means that the design aspect could be reconstructed from the information provided in the article.

Allocation sequence	
Was the allocation sequence randomly generated? *(Where random is taken to mean: random number table, computer random number generator, coin tossing, shuffling cards or envelopes, throwing a dice, drawing of lots, minimisation)*	(0=No, 1=Yes-Sufficient to replicate, 2=Yes-Insufficient to replicate, 3=Unclear)
Selection bias	
Research team	
Do the people allocating the intervention sequence to the ***clusters*** know what the intervention sequence is? *(Allocation concealment)*	(0=No, 1=Yes, 2=Unclear)
Do the people recruiting/identifying ***participants*** know which intervention sequence has been assigned to the cluster?	(0=No, 1=Yes, 2=Unclear, 3=All participants recruited/identified before cluster randomisation)
Can the people recruiting/identifying ***participants*** influence which people are recruited/identified for inclusion in the study?	(0=No, 1=Yes, 2=Unclear) Provide text to justify judgement – *e.g. participants are identified systematically from administrative data so identifier cannot influence inclusion.*
Individual participants	
Who provides consent for the individual ***participant*** to ***receive intervention***?	1. Individual. Consent is given by individual prior to intervention. 2. Cluster level. Individual participant does not give consent for intervention and cannot opt out of intervention. Consent is given by cluster spokesperson. 3. Opt out. Individual participant does not give consent for intervention. Intervention will be given unless participant opts out of intervention. Consent is given by cluster spokesperson. 4. Delayed consent. Consent is obtained from individual or their next of kin to continue intervention, but intervention is initiated without individual consent. Initial consent is given by cluster spokesperson.

	5. Other 6. Unclear
If the individual participant provides consent does the ***participant*** have knowledge of the intervention they will receive ***prior*** to consenting?	(0=No, 1=Yes, 2=Unclear, 3=NA (if option 2))
Is the intervention concealed to ***participants during*** the study? (*I.e. is the intervention blinded?*)	(0=No, 1=Yes, 2=Unclear)
Consent for data collection	
Does consent for ***data collection*** occur at the cluster level, individual level, or is not required?	(0=Not stated, 1=cluster, 2=individual, 3=not required, 4=Unclear)
Performance bias	
Was the intervention concealed at cluster level (*i.e. were health care professionals delivering intervention blind to the intervention*)?	(0=No, 1=Yes, 2=Unclear)
Detection bias	
Were any patient, or individual level, reported outcomes collected (*e.g. pain, depression)?*	(0=No, 1=Yes, 2=Unclear)
If yes, list patient or individual level reported outcomes	TEXT
Were any subjective outcomes *(e.g. clinician rated depression, condition specific mortality)* collected by study personnel, clinicians, or outcome assessors *(i.e. not reported by the individual level participant)?*	(0=No, 1=Yes, 2=Unclear)
If yes, list which outcomes you considered subjective	TEXT
If yes, were the study personnel, clinicians, or outcome assessors who were assessing the subjective outcomes blind to the intervention assignment?	(0=No, 1=Yes, 2=Unclear)

Section 10: Intervention

Rationale: To describe the type of interventions being used in CRXO

Type of experimental intervention (select all that apply. 0=No, 1=Yes, 2 =Unclear)

Educational/quality improvement interventions targeted at health care professionals (e.g., distribution of educational materials, outreach visits, audit and feedback)	(0=No, 1=Yes, 2 =Unclear)
Quality improvement interventions targeted at the organisation of health care or health delivery service (e.g., financial, shifting of professional roles, multi-disciplinary teams, integration of services, changes in setting or equipment, home visits by nurses)	(0=No, 1=Yes, 2 =Unclear)
Participant health promotion or educational intervention (e.g., promotion of breastfeeding, smoking cessation intervention, decision aid, disease screening promotion)	(0=No, 1=Yes, 2 =Unclear)
Direct participant therapeutic intervention (e.g., **experimental** intervention includes drug/vaccine/vitamin supplement, insecticide spraying, surgery, testing of new clinical pathway – distinguish from **indirect** changes to patient therapies as a result of guideline adherence)	(0=No, 1=Yes, 2 =Unclear)
Other, specify	TEXT
Details of experimental intervention	TEXT

Control intervention (select one)	1. Not reported 2. No active intervention, i.e. usual care 3. Minimal application for experimental intervention 4. Placebo intervention 5. Other active intervention 6. Other, specify
Details of control intervention	TEXT

Section 11: Sample size

Rationale: To assess how sample size calculations are being performed and justified

Correlation terminology:

Indiviudual i, Cluster j, Period k

Within-cluster within-period correlation: Corr(y_ijk, y_i'jk)

Within-cluster between-period correlation: Corr(y_ijk, y_i'jk')

For the questions which ask for a justification, these are yes/no questions, either a justification was provided or it was not. However the "unclear" option remains because circumstances may arise where it isn't clear if the question applies.

Was a sample size/power calculation presented?	(0=No, 1=Yes-Sufficient to replicate, 2=Yes-Insufficient to be reproduced, 3=Unclear)
Which outcome was the sample size calculation based on?	TEXT
What was the scale of the outcome?	1. Continuous 2. Binary 3. Categorical 4. Count 5. Time to event 6. Other, specify
Was there a justification for number of periods?	(0=No, 1=Yes, 2=Unclear) Page and paragraph number
Was there a justification for number of clusters?	(0=No, 1=Yes, 2=Unclear) Page and paragraph number
Was there a justification for number of participants per cluster?	(0=No, 1=Yes, 2=Unclear) Page and paragraph number
Were equal (as opposed to unequal) cluster sizes assumed in the calculation?	(0=Unequal, 1=Equal, 2=Unclear) Page and paragraph number
Was the ***within-cluster within-period*** clustering taken into consideration in the calculation?	(0=No, 1=Yes, 2=Unclear) Page and paragraph number
Please provide additional details about the cluster level clustering if "within-cluster within-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the value?	1=correlation 2=variance components 3=design effect 4=other, specify
If yes, what was the value?	Float
If yes, what was the reference or source for the value for the ICC? (*e.g. pilot study, previous published research, best guess, unpublished research)*	TEXT
Was the within-cluster between-period clustering taken into consideration in the calculation?	(0=No, 1=Yes, 2=Unclear) TEXT

Please provide additional details about the period level clustering if "within-cluster between-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the value?	1=correlation 2=variance 3=other, specify
If yes, what was the value?	Float
If yes, what was the reference/source of the value for the ICC? (*e.g. pilot study, previous published research, best guess, unpublished research)*	TEXT
If a reference or method was provided for the sample size calculation, provide the reference or details of the method *(copy **verbatim** from article)*	TEXT
Any additional comments?	TEXT

Section 12: Outcomes and Results

Rationale: To describe the outcome measures being assessed with the CRXO design, and how they were being assessed.

For the primary outcome from the study, and the first secondary outcome that is reported in the abstract that is of a different data type to the primary outcome, answer the following:

Specify outcome (*copy **verbatim** from text*)	TEXT
Classify how the outcome was identified from the study:	Primary outcome: 1. First primary outcome in the protocol document or published article 2. The outcome used for the sample size calculation 3. The first outcome listed in the article abstract Secondary outcomes: 4. First outcome reported in abstract that is of a different data type to the primary outcome
What type of data is the outcome? (select one)	1=Continuous 2=Binary 3=Categorical 4=Count 5=Time to event 6=Other, specify
How was the statistical analysis concerning the intervention effect performed for the outcome? (*copy **verbatim** from text*)	TEXT
Was a justification given for the choice of analysis? *(E.g. Was a justification given for why they chose a multilevel individual level approach rather than a cluster level approach)*	(0=No, 1=Yes, 2=Unclear)

What references were provided for the statistical analysis and/or justification of the analysis?	TEXT
Was the within-cluster within-period clustering accounted for in the analysis of the outcome?	(0=No, 1=Yes, 2=Unclear)
Please provide additional details about the cluster level clustering if "within-cluster within-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the clustering measure?	1=ICC 2=variance components 3=coefficient of variation 4=not reported 5=other, specify
If yes, what was the value of the clustering measure?	99=Not reported
Was the within-cluster between-period clustering accounted for in the analysis of the outcome?	(0=No, 1=Yes, 2=Unclear)
Please provide additional details about the period level clustering if "within-cluster between-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the clustering measure?	1=ICC 2=variance components 3=coefficient of variation 4=not reported 5=other, specify
If yes, what was the value of the clustering measure?	99=Not Reported
Were any other levels of clustering accounted for in the analysis? (*copy **verbatim** from text*)	TEXT
Was the intervention effect adjusted for any covariates?	(0=No, 1=Yes, 2=Unclear)
If yes, were the covariates individual or cluster level?	(1=Individual, 2=Cluster, 3=Both, 4=Unclear)
If yes, was the adjustment performed at individual or cluster level?	(1=Individual, 2=Cluster, 3=Both, 4=Unclear)
Provide details covariate adjustment *(copy **verbatim** from article)*	TEXT

Section 13: Baseline Characteristics (Table 1)

Rationale: To describe how the baseline characteristics were summarised (Table 1).

In the following table, select all statements that apply.

By intervention (i.e. separate summaries for the control and the intervention groups) *(NA if same participants are included in both interventions)*	(0=No, 1=Yes, 2=NA)
By period (i.e. separate summaries for each period for each intervention) *(NA if same participants are included in both periods)*	(0=No, 1=Yes, 2=NA)
By cluster (I.e. a separate summary for each cluster)	(0=No, 1=Yes)
Other, specify	TEXT

What were the sizes of the analysed clusters? If the values are not reported directly, but can be calculated from the supplied data, then perform the calculation and enter that value. I.e. 834 participants from 18 clusters gives a mean cluster size of 46.

What was the mean cluster size ***overall***? (99 = Not determinable)	Float
Was an indication provided for the variation in cluster size between clusters?	(0=No, 1=Yes)
*If yes – copy text **verbatim** from article*	
What was the mean cluster size in each ***intervention***? (99 = Not determinable)	
Intervention 1 (name:)	Float,
Insert rows for each intervention	
What was the mean cluster size in each ***period***? (99 = Not determinable)	
Period 1 (name:)	Float,
Insert rows for each intervention	
What was the mean cluster size in each ***intervention and period*** (ie in each cluster period)? (99 = Not determinable)	
Period 1, Intervention 1	Float
Insert rows for each period and intervention	
Were any other summary statistics of the cluster sizes provided in the article? *(Copy text **verbatim** from article) (e.g. coefficient of variation, harmonic mean)*	TEXT
Was an indication provided for the variation in cluster size over time? *(e.g. between periods)*	(0=No, 1=Yes)
*If yes – copy text **verbatim** from article*	TEXT

Section 14: Missing data

Rationale: To summarise whether missing data is being reported in the CRXO trials, and how the data is being accounted for in analyses

Was missing data discussed in the article?	(0=No, 1=Yes)
How was missing data reported? (*copy **verbatim** from text*)	TEXT
How did the authors account for missing data in the analysis? (*copy **verbatim** from text*)	TEXT

Appendix B

Supplementary data accompanying Chapter 3 - Supplementary tables S1 to S4, Systematic review data extraction form, Reference list of included trials

Supplementary Tables

Table S1: Summary of changes in methods from protocol methods

Protocol method	Deviation from protocol method	Justification
Inclusion criteria		
The trial was undertaken in humans; the allocation of the intervention was to clusters of individuals rather than individuals themselves; each cluster received each intervention in a sequence over time (conventional crossover design), or at least some clusters crossed over from one intervention to another (such as two- treatment-four-sequence designs AA, AB, BA, and BB)	At least some clusters crossed each way between at least two interventions. *Type of deviation*: Addition	The intervention sequence in a CRXO design is randomly ordered to control for period effects. Several extracted trials applied a pre-post design to all clusters or applied a pre-post design to some clusters and a control intervention to the remaining clusters in all periods.
	The intervention given in the one period was not deliberately intended by design to affect individuals in subsequent periods. *Type of deviation*: Addition	Several extracted trials used designs where it was intended that the interventions in each intervention sequence would be compared separately, rather than pooled, to evaluate any ordering effects.
Full text review		
Two reviewers will assess the full text articles.	All eligible articles were double screened along with 20% of articles that were initially determined to be ineligible by SA *Type of deviation*: Amendment	The number of located CRXO trials was much greater than we had anticipated and we did not have the resources available for all full text articles to be double screened. We therefore amended our process to reduce the number full text articles that were double screened, but placed greater emphasis in this process on making the correct decision regarding non-inclusion of ineligible trials.
Data extraction form	Changes to data extraction were made after further piloting of the original data extraction form. See additional file 1. *Type of deviation:* Amendment	After piloting the original data extraction form, changes were required to improve the clarity of the questions and to ensure that all data was extracted as intended in the original data extraction form.
Data extraction		
Two reviewers will independently extract data using an electronic data extraction form developed for this review	One author (SA) extracted data from all trials, and data from 20% of the trials was independently double extracted by the co-authors. Three of the five authors (SA, JM, AF) reviewed the discrepancies arising from the double data	The number of located CRXO trials was much greater than we had anticipated and we did not have the resources available for all articles to have double data extraction.

	extraction, and discussed processes for further reviewing items where there was inconsistency.	

Table S2: Country where the trial was conducted

Country	**N = 91 n (%)**
USA	30 (33%)
UK	10 (11%)
The Netherlands	9 (10%)
More than 1 country	5 (5%)
Canada	5 (5%)
Australia	4 (4%)
France	4 (4%)
China	2 (2%)
Denmark	2 (2%)
Germany	2 (2%)
Sweden	2 (2%)
Thailand	2 (2%)
Austria	1 (1%)
Belgium	1 (1%)
Estonia	1 (1%)
Finland	1 (1%)
Greece	1 (1%)
Kenya	1 (1%)
New Zealand	1 (1%)
Pakistan	1 (1%)
South Korea	1 (1%)
South Africa	1 (1%)
Switzerland	1 (1%)
Taiwan	1 (1%)
Tanzania	1 (1%)
Zambia	1 (1%)

Table S3: Type of randomised cluster

Randomised cluster type	**N = 91 n (%)**
Hospital or ward	45 (49%)
ICU	19
Other wards	26
Individual health care provider	13 (14%)
School or class	13 (14%)
Class or classroom	7
School	5
Group of students	1
Emergency medical team	6 (7%)
Primary care practice	4 (4%)
Individual (mouth, muscles)	2 (2%)
Dementia unit or facility	2 (2%)
Aged care facility	2 (2%)
Community or geographical area	1 (1%)
Household or family group	1 (1%)
Workplace	1 (1%)
Outpatient clinic	1 (1%)

ICU: Intensive care unit

Table S4: Justifications for cluster randomised crossover designs by type of cluster randomised

	All cluster types	**Type of cluster randomised**			
	N total	**Hospital**	**School**	**Health care provider**	**Other[1]**
	N = 91	**N = 45**	**N = 13**	**N=13**	**N=20**
Justifications					
Justification for cluster randomisation	36 (40%)	22 (49%)	4 (31%)	3 (23%)	7 (35%)
Justification for cross-over	38 (42%)	22 (49%)	5 (39%)	7 (54%)	4 (20%)
Cluster randomisation	**N = 36**	**N = 22**	**N = 4**	**N = 3**	**N= 7**
Contamination likely between participants at the level of person/people delivering the intervention	11 (31%)	4 (18%)	1 (25%)	3 (100%)	3 (43%)
Contamination likely between participants in a cluster	11 (31%)	8 (36%)	2 (50%)	0	1 (14%)
Practical/ethical/cost/administrative difficulties with randomising at an individual level	11 (31%)	6 (27%)	2 (50%)	1 (33%)	2 (29%)
Ensure the intervention is fully delivered	5 (14%)	3 (14%)	0	0	2 (29%)
Reflective of how the intervention will be applied in practice	4 (11%)	3 (14%)	0	0	1 (14%)
Control for cluster level variation or achieve a balance of cluster covariates across interventions	3 (8%)	1 (5%)	1 (25%)	0	1 (14%)
Intervention only acts at the cluster level, impossible to randomise individually	2 (6%)	2 (9%)	0	0	0
Maximise the number of participants	1 (3%)	1 (5%)	0	0	0
Ensure blinding of participants is possible	1 (3%)	1 (5%)	0	0	0
Accepted trial design for school based interventions	1 (3%)	0	1 (25%)	0	0
Cross over	**N = 38**	**N = 22**	**N = 5**	**N = 7**	**N = 4**

Clusters expected to have different characteristics or to account for cluster level confounding	27 (71%)	17 (77%)	4 (80%)	4 (57%)	2 (50%)
Reduced efficiency in estimating the intervention effect due to a limited number of clusters available for inclusion in the trial	5 (13%)	2 (9%)	1 (20%)	1 (14%)	1 (25%)
Increased efficiency to overcome loss of power through randomising in clusters	4 (11%)	2 (9%)	2 (40%)	0	0
Control period to "act as own control" (with no further justification)	5 (13%)	4 (18%)	0	1 (14%)	0
Allow for within cluster comparison (with no further justification)	1 (3%)	1 (5%)	0	0	0
Increase power or precision (no further justification)	4 (11%)	1 (5%)	0	2 (29%)	1 (25%)
Increase participation	3 (8%)	2 (10%)	1 (20%)	0	0
Reduce bias as each cluster contributes the same number of participants in each period	2 (5%)	1 (5%)	0	0	1 (25%)
Allay ethical concerns by ensuring the intervention is received by all clusters	1 (3%)	1 (5%)	0	0	0
Allow for blinding of the allocation sequence. Participants would not know when they were in the control period	1 (3%)	0	0	1 (14%)	0
Permit historical and concurrent controls for each cluster	1 (3%)	1 (5%)	0	0	0

[1] Other cluster types include: Aged care facilities, dementia facilities, primary care practices, and outpatient facilities; households and geographic regions; and worksite departments, emergency responder teams, and individual patients (units receiving treatment were individual teeth or muscles).

Additional files

Additional file 1: Data extraction form

CRXO systematic review data extraction form

Date: 16 December 2014

Section 1: Study Identifiers

Study ID	(Autocompleted)
1.1 Is the article a protocol paper?	0=No, 1=Yes
1.1b Are there any other papers associated with this study? List references	String
1.2 First Author Surname	String
1.3 Publication Year	Integer, 1946 to 2014
1.5 Reviewer's initials	Categorical
1.6 Date of review	Date
1.7 Notes	Not to be analysed

The unit of analysis for the review is ***study or trial***, not article. In many cases a study will be split into multiple articles, i.e. a protocol or design article, an article reporting the primary outcome(s), and many other articles reporting secondary outcomes.

The primary outcome for the study will be defined from the following hierarchy:

- The first primary outcome in the protocol document or first published paper for the study if there is no protocol document.
- The outcome used for the sample size calculation.
- The first outcome listed in the methods section of the abstract.

Section 2: Full Text Screening

Does the article report a research trial that used or planned to use a CRXO design which incorporated the following design elements (*The article will only be marked for inclusion if "yes" is answered to all):*

2.1 Outcomes were measured on humans in a study or trial at either cluster or individual level.	(0=No, 1=Yes, 2=Unclear)
2.2 Allocation of the intervention was at cluster level (The allocation does not have to be at random).	(0=No, 1=Yes, 2=Unclear)
2.3 Each cluster received each intervention, or at least some clusters crossed over from one intervention to another (e.g. two-intervention-four-sequence designs AA, AB, BA, BB).	(0=No, 1=Yes, 2=Unclear)
2.4 Each cluster received each intervention in a sequence over time, rather than concurrently in time.	(0=No, 1=Yes, 2=Unclear)
2.5 At least some clusters crossed each way between at least two interventions (e.g. one cluster received AB and one cluster received BA)	(0=No, 1=Yes, 2=Unclear)
2.6 The intervention given in the one period was not deliberately intended by design to affect individuals in subsequent periods (e.g. interventions intended to change the prescribing behaviour of health care provider, where patients form the cluster)	(0=No, 1=Yes, 2=Unclear)

Section 3: Title and Abstract

Rationale: To assess how CRXO trials are identified in the title and abstract.

Title	
Is the trial identified as a **cluster** randomised **crossover** trial in title? (*note words 'cluster' and 'crossover' must be used, placement of hyphens is unimportant*)	(0=No, 1=Yes)
Abstract	
Is the trial identified as a **cluster** randomised trial in abstract? (*note words 'cluster' must be used, placement of hyphens is unimportant*)	(0=No, 1=Yes)
If no, copy verbatim from abstract how the unit of randomisation was described in the abstract	TEXT
Is the trial identified as a **crossover** trial in abstract? (*note words 'crossover' must be used, placement of hyphens is unimportant*)	(0=No, 1=Yes)
If no, copy verbatim from abstract how the cross over of interventions was described in the abstract	TEXT

Section 4: Justification for the CRXO design

Rationale: To understand why researchers are using the CRXO design and how they justify that decision.

Why was the CRXO design chosen? *For **each** of the following points enter (0=Not Discussed, 1=Yes, 2=Unclear). Select as many points as apply.*

*Justification given by authors for **cluster** randomisation*	
Intervention can **only** act at the cluster level, and therefore impossible to randomise individually *(e.g. if the intervention is an educational program for health care practitioners, or a program implemented publicly via radio or newspaper, the intervention will reach a group of people)*.	(0=Not Discussed, 1=Yes, 2=Unclear)
Practical/ethical/cost/administrative difficulties with randomising at an individual level.	(0=Not Discussed, 1=Yes, 2=Unclear)
Contamination likely between participants at the level of person/people delivering the intervention *(e.g. an educational intervention may be delivered to health care practitioners, and it may impossible for them to only apply the intervention to some individuals in their care and not others. Therefore contamination would occur in an individually randomised trial).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Contamination likely between participants in a cluster *(e.g. a behavioural intervention may be delivered to schools, and it may be impossible to prevent primary caregivers from exchanging experiences, thereby contaminating each arm of the trial in an individually randomised trial).*	(0=Not Discussed, 1=Yes, 2=Unclear)
To ensure intervention is fully delivered *(if it is expected that compliance with the trial protocol will be reduced if members of a cluster were individually randomised)*	(0=Not Discussed, 1=Yes, 2=Unclear)
Outcome data only available at cluster level	(0=Not Discussed, 1=Yes, 2=Unclear)
Other, specify	TEXT

*Justification given by authors for **crossover** design*	
Increased efficiency to overcome loss of power through randomising in clusters (*i.e. the authors specifically cite a reduction in precision/power due to cluster randomisation or the 'design effect' as the reason for the crossover element).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Reduced efficiency in estimating the intervention effect due to a limited number of clusters available for inclusion in the trial *(i.e. the authors cite the limited number of clusters as the reason for the crossover element).*	(0=Not Discussed, 1=Yes, 2=Unclear)
Clusters are expected to have different characteristics from each other or to account for cluster level confounding (*i.e. the authors cite that they expect or wish to allow for clusters being very different in characteristics which might affect the outcome, and wish to crossover so that each cluster 'acts as own control')*	(0=Not Discussed, 1=Yes, 2=Unclear)
Other, specify	TEXT

Section 5: Trial Objectives

Rationale: What are the levels of the primary objective being addressed with the CRXO trial?

Copy **verbatim** objective or hypothesis from Introduction	TEXT

Section 6: Population Details

Rationale: What settings/conditions are CRXO trials being used in?

Disease or domain under study (can select multiple)	1. Cancer 2. Cardiovascular 3. Central nervous system/musculoskeletal 4. Digestive/endocrine 5. Nutritional and metabolic 6. Gynaecology 7. Pregnancy and birth and paediatrics 8. Infectious diseases 9. Mental health and behavioural conditions 10. Pathological conditions 11. Symptoms and signs 12. Respiratory disease 13. Urogenital 14. Blood and immune system 15. Ear and nose 16. Eye 17. General health / public health 18. Genetic disorders 19. Injuries 20. Accidents and wounds 21. Mouth and dental 22. Skin 23. Other
Country of trial (List all if 5 or less, otherwise state multinational)	TEXT
Setting (select one)	1. Primary care practices/health care clinics 2. Communities/geographical areas 3. Households/families 4. Aged care facility 5. Hospital 6. Schools 7. Workplaces 8. Other,specify
List if setting is other	TEXT
Do the methods define the cluster unit?	(0=No, 1=Yes, 2=Description unclear)
Clusters receiving intervention (select one)	1. Primary care practices (practice includes multiple health care professionals)

	2. Individual health professional 3. Communities/Residential areas 4. Households/families 5. Hospital, specify unit/ward type 6. Nursing home/aged care 7. Schools 8. Worksites 9. Other, specify
If hospital, specify ward or unit	TEXT
If other, specify	TEXT
Additional comments about clusters receiving intervention	TEXT

Section 7: Study Design

Rationale: This section is intended to capture both the key design features of the published CRXO trial and how the design features are reported.

An answer of "yes" means that the design aspect could be reconstructed from the information provided in the article.

The number of participating clusters	Integer, 99 = Not reported
The number of periods	Integer, 99 = Not reported
The number of interventions	
Intervention treatments (*active interventions*)	Integer, 99 = Not reported
Control treatments (*e.g. no treatment, usual care. Enter 0 if all interventions are active*)	Integer, 99 = Not reported
List the different unique intervention sequences, (*i.e. AB, BA; or AA, BB, AB, BA*.)	TEXT
Is a diagram included to describe design?	(0=No, 1=Yes-complete, 2=Yes-incomplete/unclear)
Is each period ***designed*** to include the same or different participants in each period? (*i.e. are measurements repeated or not repeated on participants?)*	(0=Same participants, 1=Different participants, 2=Mix of same and different participants, 3=Unclear)
Do the authors ***discuss*** how many interventions each ***participant*** can receive, i.e. if the participant can remain in cluster for longer than one period or could be included again in a later period?	(0=Not Discussed, 1=Yes, 2=Unclear, 3=NA)
If the study is designed to include different participants in each period, does the study design make it possible for participants to be included in more than one period?	(0=No, 1=Yes, 2=Unclear)
Describe why you think it is possible that participants are in more than one period	TEXT
Are there any other relevant design features that may lead to additional correlation within the outcomes? E.g. hierarchical designs where there is clustering at different levels; wards within hospitals, GPs within general practices. (*copy **verbatim** from text*)	TEXT

Section 8: Carry over

Rationale: To describe whether the risk of carry over being is acknowledged and managed.

Do the authors ***discuss*** the possibility of carry over of intervention effects between periods?	(0=Not Discussed, 1=Yes, 2=Unclear*),* Page and paragraph number
Is a washout period included?	(0=No and absence not explained, 1=Yes or absence explained, 2 = Not clear)
Was carry over managed in any other way?	(0=None listed, 1=Yes, text copied below)
Copy verbatim other ways in which carry over was managed	TEXT
If carry over possible, assess the risk of carry over effects (select one)	1=Unlikely 2=Possible 3=Likely 4=Unclear
Describe your rationale for the assessment of the risk of carry over in above question	TEXT

Section 9: Blinding, selection bias and consent

Rationale: To understand how randomisation or allocation of interventions was performed, the risk of bias in CRXO trials, and the adequacy of reporting.

An answer of "yes" means that the design aspect could be reconstructed from the information provided in the article.

Allocation sequence	
Was the allocation sequence randomly generated? *(Where random is taken to mean: random number table, computer random number generator, coin tossing, shuffling cards or envelopes, throwing a dice, drawing of lots, minimisation)*	(0=No, 1=Yes-Sufficient to replicate, 2=Yes-Insufficient to replicate, 3=Unclear)
Were any covariates used in the randomisation scheme? *(ie stratification, minimisation, matching based on one or more covariates?)*	(0=None listed, 1=Yes, 2=Unclear)
If yes, complete for each covariate used in sample size calculation	
Covariate name: (seccova_covname)	*Covariate level (1=individual, 2=cluster, 3=both, 4=unclear): (seccova_level)*
Selection bias	
Research team	
Is the allocation sequence known to the people allocating the intervention sequence to the clusters? I.e. can particular intervention sequences be deliberately matched to the clusters? *(Allocation concealment)*	(0=No, 1=Yes, 2=Unclear)
Do the people recruiting/identifying ***participants*** know which intervention sequence has been assigned to the cluster?	(0=No, 1=Yes, 2=Unclear, 3=NA-All participants recruited/identified before cluster randomisation, 4=NA-no recruitment/identification takes place)
Can the people recruiting/identifying ***participants*** influence which people are recruited/identified for inclusion in	(0=No, 1=Yes, 2=Unclear) Provide text to justify judgement – *e.g. participants are identified systematically from administrative*

the study?	*data so identifier cannot influence inclusion.*
Provide text to justify judgement	TEXT
Individual participants	
Who provides consent for the individual ***participant*** to ***receive intervention***?	1. Individual. Consent is given by individual prior to intervention. 2. Cluster level. Individual participant does not give consent for intervention and cannot opt out of intervention. Consent is given by cluster spokesperson. 3. Opt out. Individual participant does not give consent for intervention. Intervention will be given unless participant opts out of intervention. Consent is given by cluster spokesperson. 4. Delayed consent. Consent is obtained from individual or their next of kin to continue intervention, but intervention is initiated without individual consent. Initial consent is given by cluster spokesperson. 5. Other 6. Unclear
Provide details if 'other' or 'unclear' is selected	TEXT
If the individual participant (or other person on their behalf) provides consent, does the ***participant*** have knowledge of the intervention to be receive first, ***prior*** to consenting? I.e. can the participant choose to take part because of the intervention they will receive first?	(0=No/unlikely, 1=Yes/possible, 2=Unclear, 3=NA (if option 2))
Is the intervention concealed to ***participants during*** the study? (*I.e. is the intervention blinded?*)	(0=No, 1=Yes, 2=Unclear)

Consent for data collection	
If the individual (or person on their behalf) does not provide consent, does consent for data collection occur at the cluster level, individual level, or is not required?	(0=Not reported, 1=cluster, 2=individual, 3=not required, 4=Reported but unclear)
Performance bias	
Was the intervention concealed at cluster level (*i.e. were health care professionals delivering intervention blind to the intervention*)?	(0=No, 1=Yes, 2=Unclear)
Detection bias	
Were any outcomes reported by the patient or individual collected (e.g. pain, depression)? (*e.g. pain, depression)?*	(0=No, 1=Yes, 2=Unclear)
If yes, list patient or individual level reported outcomes	TEXT
Were any subjective outcomes (e.g. clinician rated depression, condition specific mortality) collected by study personnel, clinicians, or outcome assessors (i.e. not reported by the individual level participant)?	(0=No, 1=Yes, 2=Unclear)
If yes, list which outcomes you considered subjective	TEXT
If yes, were the study personnel, clinicians, or outcome assessors who were assessing the subjective outcomes blind to the intervention assignment?	(0=No, 1=Yes, 2=Unclear)

Section 10: Intervention

Rationale: To describe the type of interventions being used in CRXO

Type of experimental intervention (select all that apply. 0=No, 1=Yes, 2 =Unclear)

Educational interventions that are **targeted at health care professionals** (e.g., distribution of educational materials, outreach visits, audit and feedback)	(0=No, 1=Yes, 2 =Unclear)
Quality improvement interventions **targeted at the organisation of health care or health delivery service.** The intervention is a new method for delivering or organising an existing health care service (e.g. 2 week vs 4 week attending physician rotation with trainees (change in delivery of medical training), daily vs on demand radiographs for mechanically ventilated patients (change in delivery of routine procedure), rapid detection test vs culture test of screening for MRSA carriage on admission (change in method of performing routine screening), financial, shifting of professional roles, multi-disciplinary teams, integration of services, changes in setting or equipment.) Distinguish from interventions to assess the effectiveness of a method or service that is performed or delivered at the level of the health care provider.	(0=No, 1=Yes, 2 =Unclear)
Intervention is **targeted at the health care professional** to indirectly alter patient outcomes. The intervention involves a change in the practise or behaviour of the health care professional (e.g. gloving procedure during venipuncture, hand and forearm cleaning for surgery.) Distinguish from quality improvement interventions where the intervention is intended to change the process of delivery of an existing health care service.	
Intervention is **targeted at the cluster environment** rather than individuals within the cluster. The intervention indirectly affects the individuals within the cluster through changes in the environment. E.g. ward cleaning regime, air quality maintenance.	
Participant health promotion or educational intervention. **Intervention is delivered directly to individual**. (e.g., promotion of breastfeeding, smoking cessation intervention, decision aid, disease screening promotion)	(0=No, 1=Yes, 2 =Unclear)
Intervention is delivered directly to participants in the cluster (e.g. change in drug or drug regime within ward for a given health condition, music therapy, exercise program, vitamin supplementation, insecticide spraying) Direct participant therapeutic intervention) Distinguish from **indirect** changes to patient therapies as a result of guideline adherence or changes at level of those delivering the intervention.	(0=No, 1=Yes, 2 =Unclear)
Other, specify	TEXT
Details of experimental intervention	TEXT
Control intervention (select one)	1. Not reported 2. No active intervention, i.e. usual care 3. Minimal application for experimental intervention 4. Placebo intervention

	5. Other active intervention 6. Other, specify
Specify if other:	TEXT
Details of control intervention	TEXT

Section 11: Sample size

Rationale: To assess how sample size calculations are being performed and justified

Correlation terminology:

Indiviudual i, Cluster j, Period k

Within-cluster within-period correlation: Corr(y_ijk, y_i'jk)

Within-cluster between-period correlation: Corr(y_ijk, y_i'jk')

For the questions which ask for a justification, these are yes/no questions, either a justification was provided or it was not. However the "unclear" option remains because circumstances may arise where it isn't clear if the question applies.

Was a sample size/power calculation presented?	(0=No, 1=Yes-Sufficient to replicate, 2=Yes-Insufficient to be reproduced, 3=Unclear 4=Reason given for no sample size calculation)
If a reference or method was provided for the sample size calculation, provide the reference or details of the method *(copy **verbatim** from article)*	TEXT
Was there a justification for number of periods? *Ie did the authors state why they used the number of periods.*	(0=No, 1=Yes, 2=Unclear, 3=Not justified but set by number available)
Was there a justification for number of clusters? *Ie did the authors state why they used the number of clusters, e.g. only 10 clusters available in region*	(0=No, 1=Yes, 2=Unclear, 3=Not justified but set by number available)
Was there a justification for number of participants per cluster? *Ie did the authors state why they used the number of participants, e.g. sample size calculation, all that were available in class.*	(0=No, 1=Yes, 2=Unclear, 3=Not justified but set by number available)
Which outcome was the sample size calculation based on?	TEXT
What was the scale of the outcome?	1. Continuous 2. Binary 3. Categorical 4. Count 5. Time to event

	6. Other, specify
If "other" provide details	TEXT
Were equal (as opposed to unequal) cluster sizes assumed in the calculation?	(0=Unequal, 1=Equal, 2=Unclear)
Was the ***within-cluster within-period*** clustering taken into consideration in the calculation?	(0=No, 1=Yes, 2=Unclear)
Please provide additional details about the cluster level clustering if you selected unclear or if you feel that clustering was adequately accounted for but "within-cluster within-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the value?	1=correlation 2=variance components 3=design effect 4=other, specify
If other, specify	TEXT
If yes, what was the value?	Float
If yes, what was the reference or source for the value for the ICC? (*e.g. pilot study, previous published research, best guess, unpublished research)*	TEXT
Was the within-cluster between-period clustering taken into consideration in the calculation?	(0=No, 1=Yes, 2=Unclear) TEXT
Please provide additional details about the period level clustering if "within-cluster between-period" does not capture the type of clustering	TEXT
If yes, what was the scale of the value?	1=correlation 2=variance 3=other, specify
If "other", please specify	TEXT
If yes, what was the value?	Float
If yes, what was the reference/source of the value for the ICC? (*e.g. pilot study, previous published research, best guess, unpublished research)*	TEXT
Were any covariates included in the sample size calculation?	(0=No, 1=Yes, 2=Unclear)
If yes, complete for each covariate used in sample size calculation	
Covariate name: *(seccovc_covname)*	*Covariate level (1=individual, 2=cluster, 3=both, 4=unclear):* *(seccovc_level)*
Any additional comments?	TEXT

Section 12: Outcomes and Results

Rationale: To describe the outcome measures being assessed with the CRXO design, and how they were being assessed.

Complete this table for each of the following:

- The primary outcome from the study (or outcome used in sample size or first outcome listed in article abstract methods or otherwise elsewhere in the abstract)
- The first secondary outcome that is reported in the abstract that is of a different data type to the primary outcome
- The first secondary outcome that is reported in the abstract that is of the same data type but analysed by a different method
- The first secondary outcome that is reported in the article that is of a different data type or analysed by a different method

Specify outcome (*copy **verbatim** from text*)	TEXT
Classify how the outcome was identified from the study:	Primary outcome: 1. First primary outcome in the protocol document or published article 2. The outcome used for the sample size calculation 3. The first outcome listed in the article abstract Secondary outcomes: 4. First outcome reported in abstract/protocol that is of a different data type to the primary outcome 5. First outcome reported in article that is of different data type or analysis
What type of data is the outcome? (select one)	1=Continuous 2=Binary 3=Categorical 4=Count 5=Time to event 6=Other, specify
If “other” describe	TEXT
How was the statistical analysis concerning the intervention effect performed for the outcome? (*copy **verbatim** from text*)	TEXT
Was a justification given for the choice of analysis? *(E.g. Was a justification given for why they chose a multilevel individual level approach rather than a cluster level approach)*	(0=No, 1=Yes, 2=Unclear)
What justification was given for the choice of analysis? Copy verbatim:	TEXT
What references were provided for the statistical analysis and/or justification of the	TEXT

analysis?	
Was the within-cluster within-period clustering accounted for in the analysis of the outcome?	(0=No, 1=Yes, 2=Unclear)
Please provide additional details about the cluster level clustering if "within-cluster within-period" does not capture the type of clustering	TEXT
If reported, what was the scale of the clustering measure?	1=ICC 2=variance components 3=coefficient of variation 4=not reported 5=other, specify
If "other" please specify	TEXT
If yes, what was the value of the clustering measure?	99=Not reported
Was the within-cluster between-period clustering accounted for in the analysis of the outcome?	(0=No, 1=Yes, 2=Unclear)
Please provide additional details about the period level clustering if "within-cluster between-period" does not capture the type of clustering	TEXT
If reported, what was the scale of the clustering measure?	1=ICC 2=variance components 3=coefficient of variation 4=not reported 5=other, specify
If "other" please specify	
If yes, what was the value of the clustering measure?	99=Not Reported
Were any other levels of clustering accounted for in the analysis? (*copy **verbatim** from text, e.g. repeated measurements on the same participant in a period*)	TEXT
Was the intervention effect adjusted for any covariates?	(0=No, 1=Yes, 2=Unclear)
Complete for each covariate used in analysis:	
Covariate name: *(seccovb_covname)*	*Which level is the covariate measured at:* *(1=individual, 2=cluster, 3=both, 4=unclear):* *(seccova_measlevel)* *Which level was the adjustment performed at:(1=individual, 2=cluster, 3=both, 4=unclear):*

	(seccova_analysislevel)
Provide details covariate adjustment *(copy **verbatim** from article)*	TEXT
Were the covariates used in randomisation included in the analysis?	(0=No, 1=Yes, 2=NA-none used in randomisation, 3=Unclear)
Any additional comments?	TEXT

Section 13: Baseline Characteristics (Table 1)

Rationale: To describe how the baseline characteristics were summarised (Table 1).

In the following table, select all statements that apply.

By intervention (i.e. separate summaries for the control and the intervention groups) *(NA if same participants are included in both interventions)*	(0=No, 1=Yes, 2=NA)
By period (i.e. separate summaries for each period for each intervention) *(NA if same participants are included in both periods)*	(0=No, 1=Yes, 2=NA)
By cluster (I.e. a separate summary for each cluster)	(0=No, 1=Yes)
By intervention sequence (I.e. a separate summary for each unique sequence of interventions)	(0=No, 1=Yes)
By total (I.e. a total summary for all participants)	(0=No, 1=Yes)
Other, specify	TEXT

Were the covariates used for randomisation reported in Table 1 (or a separate table)?	0=No, 1=Yes-some, 2=Yes-all, 3=NA, 4=Unclear, 5=No table 1
Were the covariates used for the sample size reported in Table 1 (or a separate table)?	0=No, 1=Yes-some, 2=Yes-all, 3=NA, 4=Unclear, 5=No table 1
Were the covariates used for analysis reported in Table 1 (or a separate table)?	0=No, 1=Yes-some, 2=Yes-all, 3=NA, 4=Unclear, 5=No table 1

What were the sizes of the **analysed** clusters? If the values are not reported directly, but can be calculated from the supplied data, then perform the calculation and enter that value. I.e. 834 participants from 18 clusters gives a mean cluster size of 46.

What was the mean number of participants in the cluster period? If multiple measurements are taken on each participant, report mean number of participants, not measurments. (99 = Not determinable)	Float
Was an indication provided for the variation in cluster size between clusters?	(0=No, 1=Yes, 2= No variation)
*If yes – copy text **verbatim** from article*	TEXT
Were any other summary statistics of the cluster sizes provided in the article? *(Copy text **verbatim** from article) (e.g. coefficient of variation, harmonic mean)*	TEXT
Was an indication provided for the variation in cluster size over time? *(e.g. between periods)*	(0=No, 1=Yes, 2 = No variation)
*If yes – copy text **verbatim** from article*	TEXT

Section: Missing data

Rationale: To summarise whether missing data is being reported in the CRXO trials, and how the data is being accounted for in analyses

Was missing data discussed in the article?	(0=No, 1=Yes, 2= Unlikely to be missing data)
How was missing data reported? (summarize, eg in text, diagram, poorly described, unclear): (*copy **verbatim** from text*)	TEXT
How did the authors account for missing data in the analysis? (*copy **verbatim** from text*)	TEXT

Additional file 2: Reference list of trials included in systematic review

Included Trials

1. Lee N, Hui DSC, Zuo Z, Ngai KLK, Lui GCY, Wo SK, Tam WWS, Chan MCW, Wong BCK, Wong RYK, et al: **A prospective intervention study on higher-dose oseltamivir treatment in adults hospitalized with influenza a and B infections.** *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2013, **57:**1511-1519.
2. Adcock KG, Hogan SM, Elci OU, Mills KL: **Do Illustrations Improve Children's Comprehension of Assent Documents?** *The journal of pediatric pharmacology and therapeutics : JPPT : the official journal of PPAG* 2012, **17:**228-235.
3. Connolly SJ, Philippon F, Longtin Y, Casanova A, Birnie DH, Exner DV, Dorian P, Prakash R, Alings M, Krahn AD: **Randomized cluster crossover trials for reliable, efficient, comparative effectiveness testing: design of the Prevention of Arrhythmia Device Infection Trial (PADIT).** *The Canadian journal of cardiology* 2013, **29:**652-658.
4. Milstone AM, Elward A, Song X, Zerr DM, Orscheln R, Speck K, Obeng D, Reich NG, Coffin SE, Perl TM, Pediatric STSG: **Daily chlorhexidine bathing to reduce bacteraemia in critically ill children: a multicentre, cluster-randomised, crossover trial.** *Lancet* 2013, **381:**1099-1106.
5. Climo MW, Yokoe DS, Warren DK, Perl TM, Bolon M, Herwaldt LA, Weinstein RA, Sepkowitz KA, Jernigan JA, Sanogo K, Wong ES: **Effect of daily chlorhexidine bathing on hospital-acquired infection.** *The New England journal of medicine* 2013, **368:**533-542.
6. Washer LL, Chenoweth C, Kim H-W, Rogers MAM, Malani AN, Riddell Jt, Kuhn L, Noeyack B, Jr., Neusius H, Newton DW, et al: **Blood culture contamination: a randomized trial evaluating the comparative effectiveness of 3 skin antiseptic interventions.** *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* 2013, **34:**15-21.
7. Wilder-Smith A, Byass P, Olanratmanee P, Maskhao P, Sringernyuang L, Logan JG, Lindsay SW, Banks S, Gubler D, Louis VR, et al: **The impact of insecticide-treated school uniforms on dengue infections in school-aged children: study protocol for a randomised controlled trial in Thailand.** *Trials* 2012, **13:**212.
8. Lucas BP, Trick WE, Evans AT, Mba B, Smith J, Das K, Clarke P, Varkey A, Mathew S, Weinstein RA: **Effects of 2- vs 4-week attending physician inpatient rotations on unplanned patient revisits, evaluations by trainees, and attending physician burnout: a randomized trial.** *JAMA : the journal of the American Medical Association* 2012, **308:**2199-2207.
9. Twardella D, Matzen W, Lahrz T, Burghardt R, Spegel H, Hendrowarsito L, Frenzel AC, Fromme H: **Effect of classroom air quality on students' concentration: results of a cluster-randomized cross-over experimental study.** *Indoor air* 2012, **22:**378-387.

10. Chant C, Mustard M, Thorpe KE, Friedrich JO: **Nurse- vs nomogram-directed glucose control in a cardiovascular intensive care unit.** *American journal of critical care : an official publication, American Association of Critical-Care Nurses* 2012, **21:**270-278.

11. Lin L-C, Huang Y-J, Watson R, Wu S-C, Lee Y-C: **Using a Montessori method to increase eating ability for institutionalised residents with dementia: a crossover design.** *Journal of clinical nursing* 2011, **20:**3092-3101.

12. Mubi M, Janson A, Warsame M, Martensson A, Kallander K, Petzold MG, Ngasala B, Maganga G, Gustafsson LL, Massele A, et al: **Malaria rapid testing by community health workers is effective and safe for targeting malaria treatment: randomised cross-over trial in Tanzania.** *PloS one* 2011, **6:**e19753.

13. Hill LJB, Williams JHG, Aucott L, Thomson J, Mon-Williams M: **How does exercise benefit performance on cognitive tests in primary-school pupils?** *Developmental medicine and child neurology* 2011, **53:**630-635.

14. Jongerden IP, Buiting AG, Leverstein-van Hall MA, Speelberg B, Zeidler S, Kesecioglu J, Bonten MJ: **Effect of open and closed endotracheal suctioning on cross-transmission with Gram-negative bacteria: a prospective crossover study.** *Critical care medicine* 2011, **39:**1313-1321.

15. Wilson APR, Smyth D, Moore G, Singleton J, Jackson R, Gant V, Jeanes A, Shaw S, James E, Cooper B, et al: **The impact of enhanced cleaning within the intensive care unit on contamination of the near-patient environment with hospital pathogens: a randomized crossover study in critical care units in two hospitals.** *Critical care medicine* 2011, **39:**651-658.

16. Kim N-H, Kim M, Lee S, Yun NR, Kim K-H, Park SW, Kim HB, Kim N-J, Kim E-C, Park WB, Oh M-D: **Effect of routine sterile gloving on contamination rates in blood culture: a cluster randomized trial.** *Annals of internal medicine* 2011, **154:**145-151.

17. Hill L, Williams JHG, Aucott L, Milne J, Thomson J, Greig J, Munro V, Mon-Williams M: **Exercising attention within the classroom.** *Developmental medicine and child neurology* 2010, **52:**929-934.

18. Nthumba PM, Stepita-Poenaru E, Poenaru D, Bird P, Allegranzi B, Pittet D, Harbarth S: **Cluster-randomized, crossover trial of the efficacy of plain soap and water versus alcohol-based rub for surgical hand preparation in a rural hospital in Kenya.** *The British journal of surgery* 2010, **97:**1621-1628.

19. Marks GB, Ezz W, Aust N, Toelle BG, Xuan W, Belousova E, Cosgrove C, Jalaludin B, Smith WT: **Respiratory health effects of exposure to low-NOx unflued gas heaters in the classroom: a double-blind, cluster-randomized, crossover study.** *Environmental health perspectives* 2010, **118:**1476-1482.

20. Parm U, Metsvaht T, Sepp E, Ilmoja ML, Pisarev H, Pauskar M, Lutsar I: **Impact of empiric antibiotic regimen on bowel colonization in neonates with suspected early onset sepsis.** *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 2010, **29:**807-816.

21. Nijssen S, Fluit A, van de Vijver D, Top J, Willems R, Bonten MJM: **Effects of reducing beta-lactam antibiotic pressure on intestinal colonization of antibiotic-resistant gram-negative bacteria.** *Intensive care medicine* 2010, **36:**512-519.

22. Gerald LB, Gerald JK, Zhang B, McClure LA, Bailey WC, Harrington KF: **Can a school-based hand hygiene program reduce asthma exacerbations among elementary school children?** *The Journal of allergy and clinical immunology* 2012, **130:**1317-1324.

23. Hejblum G, Chalumeau-Lemoine L, Ioos V, Boelle P-Y, Salomon L, Simon T, Vibert J-F, Guidet B: **Comparison of routine and on-demand prescription of chest radiographs in mechanically ventilated adults: a multicentre, cluster-randomised, two-period crossover study.** *Lancet* 2009, **374:**1687-1693.

24. Rozendaal FW, Spronk PE, Snellen FF, Schoen A, van Zanten ARH, Foudraine NA, Mulder PGH, Bakker J, Ulti Si: **Remifentanil-propofol analgo-sedation shortens duration of ventilation and length of ICU stay compared to a conventional regimen: a centre randomised, cross-over, open-label study in the Netherlands.** *Intensive care medicine* 2009, **35:**291-298.

25. de Smet AMGA, Kluytmans JAJW, Cooper BS, Mascini EM, Benus RFJ, van der Werf TS, van der Hoeven JG, Pickkers P, Bogaers-Hofman D, van der Meer NJM, et al: **Decontamination of the digestive tract and oropharynx in ICU patients.** *The New England journal of medicine* 2009, **360:**20-31.

26. Kelly P, Katubulushi M, Todd J, Banda R, Yambayamba V, Fwoloshi M, Zulu I, Kafwembe E, Yavwa F, Sanderson IR, Tomkins A: **Micronutrient supplementation has limited effects on intestinal infectious disease and mortality in a Zambian population of mixed HIV status: a cluster randomized trial.** *The American journal of clinical nutrition* 2008, **88:**1010-1017.

27. Jeyaratnam D, Whitty CJM, Phillips K, Liu D, Orezzi C, Ajoku U, French GL: **Impact of rapid screening tests on acquisition of meticillin resistant Staphylococcus aureus: cluster randomised crossover trial.** *BMJ (Clinical research ed)* 2008, **336:**927-930.

28. Harbarth S, Fankhauser C, Schrenzel J, Christenson J, Gervaz P, Bandiera-Clerc C, Renzi G, Vernaz N, Sax H, Pittet D: **Universal screening for methicillin-resistant Staphylococcus aureus at hospital admission and nosocomial infection in surgical patients.** *JAMA : the journal of the American Medical Association* 2008, **299:**1149-1157.

29. Tortolero SR, Markham CM, Addy RC, Baumler ER, Escobar-Chaves SL, Basen-Engquist KM, McKirahan NK, Parcel GS: **Safer choices 2: rationale, design issues, and baseline results in evaluating school-based health promotion for alternative school students.** *Contemporary clinical trials* 2008, **29:**70-82.

30. Bleasdale SC, Trick WE, Gonzalez IM, Lyles RD, Hayden MK, Weinstein RA: **Effectiveness of chlorhexidine bathing to reduce catheter-associated bloodstream infections in medical intensive care unit patients.** *Archives of internal medicine* 2007, **167:**2073-2079.

31. Glazebrook C, Marlow N, Israel C, Croudace T, Johnson S, White IR, Whitelaw A: **Randomised trial of a parenting intervention during neonatal intensive care.** *Archives of disease in childhood Fetal and neonatal edition* 2007, **92:**F438-443.

32. Sloane PD, Williams CS, Mitchell CM, Preisser JS, Wood W, Barrick AL, Hickman SE, Gill KS, Connell BR, Edinger J, Zimmerman S: **High-intensity environmental light in dementia: effect on sleep and activity.** *Journal of the American Geriatrics Society* 2007, **55:**1524-1533.

33. Vessey JA, Sherwood JJ, Warner D, Clark D: **Comparing hand washing to hand sanitizers in reducing elementary school students' absenteeism.** *Pediatric nursing* 2007, **33:**368-372.

34. Lucas BP, Asbury JK, Wang Y, Lee K, Kumapley R, Mba B, Borkowsky S, Asmar A: **Impact of a bedside procedure service on general medicine inpatients: A firm-based trial.** *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* 2007, **2:**143-149.

35. Fowell A, Johnstone R, Finlay IG, Russell D, Russell IT: **Design of trials with dying patients: a feasibility study of cluster randomisation versus randomised consent.** *Palliative medicine* 2006, **20:**799-804.

36. Sjogren T, Nissinen KJ, Jarvenpaa SK, Ojanen MT, Vanharanta H, Malkia EA: **Effects of a physical exercise intervention on subjective physical well-being, psychosocial functioning and general well-being among office workers: a cluster randomized-controlled cross-over design.** *Scandinavian journal of medicine & science in sports* 2006, **16:**381-390.

37. Thomas B, Jones M, Johns P, Trauer T: **P.r.n. medication use in a psychiatric high-dependency unit following the introduction of a nurse-led activity programme.** *International journal of mental health nursing* 2006, **15:**266-271.

38. Hoeffer B, Talerico KA, Rasin J, Mitchell CM, Stewart BJ, McKenzie D, Barrick AL, Rader J, Sloane PD: **Assisting cognitively impaired nursing home residents with bathing: effects of two bathing interventions on caregiving.** *The Gerontologist* 2006, **46:**524-532.

39. Lang E, Afilalo M, Vandal AC, Boivin J-F, Xue X, Colacone A, Leger R, Shrier I, Rosenthal S: **Impact of an electronic link between the emergency department and family physicians: a randomized controlled trial.** *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2006, **174:**313-318.

40. Larson EL, Cimiotti J, Haas J, Parides M, Nesin M, Della-Latta P, Saiman L: **Effect of antiseptic handwashing vs alcohol sanitizer on health care-associated infections in neonatal intensive care units.** *Archives of pediatrics & adolescent medicine* 2005, **159:**377-383.

41. Cook DA, Dupras DM, Thompson WG, Pankratz VS: **Web-based learning in residents' continuity clinics: a randomized, controlled trial.** *Academic medicine : journal of the Association of American Medical Colleges* 2005, **80:**90-97.

42. Vinson DC, Galliher JM, Reidinger C, Kappus JA: **Comfortably engaging: which approach to alcohol screening should we use?** *Annals of family medicine* 2004, **2:**398-404.

43. Lam WWT, Fielding R, Johnston JM, Tin KYK, Leung GM: **Identifying barriers to the adoption of evidence-based medicine practice in clinical clerks: a longitudinal focus group study.** *Medical education* 2004, **38:**987-997.

44. Morton JL, Schultz AA: **Healthy Hands: Use of alcohol gel as an adjunct to handwashing in elementary school children.** *The Journal of school nursing : the official publication of the National Association of School Nurses* 2004, **20:**161-167.

45. Parienti J-J, du Cheyron D, Ramakers M, Malbruny B, Leclercq R, Le Coutour X, Charbonneau P, Members of the NSG: **Alcoholic povidone-iodine to prevent central venous catheter colonization: A randomized unit-crossover study.** *Critical care medicine* 2004, **32:**708-713.

46. Vrentzos GE, Papadakis JA, Vardakis KE, Maliaraki N, Stilianou K, Arvanitis A, Sratigis S, Alivanis P, Margioris AN, Ganotakis ES: **Intravenous administration of vitamin B12 in the treatment of hyperhomocysteinemia associated with end-stage renal disease.** *Journal of nephrology* 2003, **16:**535-539.

47. van Alem AP, Vrenken RH, de Vos R, Tijssen JGP, Koster RW: **Use of automated external defibrillator by first responders in out of hospital cardiac arrest: prospective controlled trial.** *BMJ (Clinical research ed)* 2003, **327:**1312.

48. Detmar SB, Muller MJ, Schornagel JH, Wever LDV, Aaronson NK: **Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial.** *JAMA : the journal of the American Medical Association* 2002, **288:**3027-3034.

49. Parienti JJ, Thibon P, Heller R, Le Roux Y, von Theobald P, Bensadoun H, Bouvet A, Lemarchand F, Le Coutour X, Antisepsie Chirurgicale des mains Study G: **Hand-rubbing with an aqueous alcoholic solution vs traditional surgical hand-scrubbing and 30-day surgical site infection rates: a randomized equivalence study.** *JAMA : the journal of the American Medical Association* 2002, **288:**722-727.

50. Calfee DP, Farr BM: **Comparison of four antiseptic preparations for skin in the prevention of contamination of percutaneously drawn blood cultures: a randomized trial.** *Journal of clinical microbiology* 2002, **40:**1660-1665.

51. Herkner H, Eisenburger P, Havel C, Laggner AN: **Dashing with scooters to in-hospital emergencies: a randomised cross-over experiment.** *Resuscitation* 2002, **52:**293-296.

52. Chan L, MacLehose RF, Houck PM: **Impact of physician reminders on the use of influenza vaccinations: a randomized trial.** *Archives of physical medicine and rehabilitation* 2002, **83:**371-375.

53. Cibere J, Sibley JT, Haga M: **Rheumatologists' adherence to guidelines for misoprostol use in patients at high risk for nonsteroidal antiinflammatory drug gastropathy.** *The Journal of rheumatology* 2002, **29:**339-346.

54. Karchmer TB, Giannetta ET, Muto CA, Strain BA, Farr BM: **A randomized crossover study of silver-coated urinary catheters in hospitalized patients.** *Archives of internal medicine* 2000, **160:**3294-3298.

55. Steele DJ, Medder JD, Turner P: **A comparison of learning outcomes and attitudes in student- versus faculty-led problem-based learning: an experimental study.** *Medical education* 2000, **34:**23-29.

56. Dahler-Eriksen BS, Lauritzen T, Lassen JF, Lund ED, Brandslund I: **Near-patient test for C-reactive protein in general practice: assessment of clinical, organizational, and economic outcomes.** *Clinical chemistry* 1999, **45:**478-485.

57. Glazener CM, Ramsay CR, Campbell MK, Booth P, Duffty P, Lloyd DJ, McDonald A, Reid JA: **Neonatal examination and screening trial (NEST): a randomised, controlled, switchback trial of alternative policies for low risk infants.** *BMJ (Clinical research ed)* 1999, **318:**627-631.

58. Chavasse DC, Shier RP, Murphy OA, Huttly SR, Cousens SN, Akhtar T: **Impact of fly control on childhood diarrhoea in Pakistan: community-randomised trial.** *Lancet* 1999, **353:**22-25.

59. Bader HI, Boyd RL: **Comparative efficacy of a rotary and a sonic powered toothbrush on improving gingival health in treated adult periodontitis patients.** *American journal of dentistry* 1999, **12:**143-147.

60. Flanagan JR, Doebbeling BN, Dawson J, Beekmann S: **Randomized study of online vaccine reminders in adult primary care.** *Proceedings / AMIA Annual Symposium AMIA Symposium* 1999**:**755-759.

61. Desai J, Sharief M, Swash M: **Riluzole has no acute effect on motor unit parameters in ALS.** *Journal of the neurological sciences* 1998, **160 Suppl 1:**S69-72.

62. Jernigan JA, Siegman-Igra Y, Guerrant RC, Farr BM: **A randomized crossover study of disposable thermometers for prevention of Clostridium difficile and other nosocomial infections.** *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* 1998, **19:**494-499.

63. Mendelson MH, Short LJ, Schechter CB, Meyers BR, Rodriguez M, Cohen S, Lozada J, DeCambre M, Hirschman SZ: **Study of a needleless intermittent intravenous-access system for peripheral infusions: analysis of staff, patient, and institutional outcomes.** *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* 1998, **19:**401-406.

64. Borutta A: **Plaque removal efficacy of a newly developed powered toothbrush in the primary dentition of pre-school children.** *The Journal of clinical dentistry* 1997, **8:**151-155.

65. Barrett BJ, Parfrey PS, McDonald J, Haire R, Peachey G: **Concurrent utilization review and inappropriate hospital stay: evaluation of a program.** *Clinical and investigative medicine Medecine clinique et experimentale* 1996, **19:**28-35.

66. Skargren E, Oberg B: **Effects of an exercise program on musculoskeletal symptoms and physical capacity among nursing staff.** *Scandinavian journal of medicine & science in sports* 1996, **6:**122-130.

67. Schwab TM, Callaham ML, Madsen CD, Utecht TA: **A randomized clinical trial of active compression-decompression CPR vs standard CPR in out-of-hospital cardiac arrest in two cities.** *JAMA : the journal of the American Medical Association* 1995, **273:**1261-1268.

68. Nydahl M, Gustafsson IB, Ohrvall M, Vessby B: **Similar serum lipoprotein cholesterol concentrations in healthy subjects on diets enriched with rapeseed and with sunflower oil.** *European journal of clinical nutrition* 1994, **48:**128-137.

69. Evans RS, Classen DC, Pestotnik SL, Lundsgaarde HP, Burke JP: **Improving empiric antibiotic selection using computer decision support.** *Archives of internal medicine* 1994, **154:**878-884.

70. Kellermann AL, Hackman BB, Somes G, Kreth TK, Nail L, Dobyns P: **Impact of first-responder defibrillation in an urban emergency medical services system.** *JAMA : the journal of the American Medical Association* 1993, **270:**1708-1713.

71. Thamlikitkul V, Chokloikaew S, Tangtrakul T, Siripoonkiat P, Wongpreedee N, Danchaivijitr S: **Blood culture: comparison of outcomes between switch-needle and no-switch techniques.** *American journal of infection control* 1992, **20:**122-125.

72. Larson E, McGeer A, Quraishi ZA, Krenzischek D, Parsons BJ, Holdford J, Hierholzer WJ: **Effect of an automated sink on handwashing practices and attitudes in high-risk units.** *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* 1991, **12:**422-428.

73. Anonymous: **Dry chemistry pathology in general practice. An assessment of utilisation, perceptions and impact on patient care. Non-Laboratory Pathology Working Party of the National Health Technology Advisory Panel.** *The Medical journal of Australia* 1991, **154:**511-518.

74. Epperson EL: **Efficacy of 0.9% sodium chloride injection with and without heparin for maintaining indwelling intermittent injection sites.** *Clinical pharmacy* 1984, **3:**626-629.

75. Barker KN, Pearson RE, Hepler CD, Smith WE, Pappas CA: **Effect of an automated bedside dispensing machine on medication errors.** *American journal of hospital pharmacy* 1984, **41:**1352-1358.

76. Hayes MJ, Morris GK, Hampton JR: **Comparison of mobilization after two and nine days in uncomplicated myocardial infarction.** *British medical journal* 1974, **3:**10-13.

77. Szyld E, Aguilar A, Musante GA, Vain N, Prudent L, Fabres J, Carlo WA: **Comparison of Devices for Newborn Ventilation in the Delivery Room.** *The Journal of pediatrics* 2014.

78. Guy RJ, Natoli L, Ward J, Causer L, Hengel B, Whiley D, Tabrizi SN, Donovan B, Fairley CK, Badman SB, et al: **A randomised trial of point-of-care tests for chlamydia**

and gonorrhoea infections in remote Aboriginal communities: Test, Treat ANd GO-the "TTANGO" trial protocol. *BMC infectious diseases* 2013, **13:**485.

79. van Werkhoven CH, Postma DF, Oosterheert JJ, Bonten MJ: **Antibiotic treatment of moderate-severe community-acquired pneumonia: design and rationale of a multicentre cluster-randomised cross-over trial.** *The Netherlands journal of medicine* 2014, **72:**170-178.

80. Hardy K, Price C, Szczepura A, Gossain S, Davies R, Stallard N, Shabir S, McMurray C, Bradbury A, Hawkey PM: **Reduction in the rate of methicillin-resistant Staphylococcus aureus acquisition in surgical wards by rapid screening for colonization: a prospective, cross-over study.** *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2010, **16:**333-339.

81. Hostler D, Everson-Stewart S, Rea TD, Stiell IG, Callaway CW, Kudenchuk PJ, Sears GK, Emerson SS, Nichol G, Resuscitation Outcomes Consortium I: **Effect of real-time feedback during cardiopulmonary resuscitation outside hospital: prospective, cluster-randomised trial.** *Bmj* 2011, **342:**d512.

82. Scales DC, Dainty K, Hales B, Pinto R, Fowler RA, Adhikari NK, Zwarenstein M: **A multifaceted intervention for quality improvement in a network of intensive care units: a cluster randomized trial.** *Jama* 2011, **305:**363-372.

83. Stiell IG, Nichol G, Leroux BG, Rea TD, Ornato JP, Powell J, Christenson J, Callaway CW, Kudenchuk PJ, Aufderheide TP, et al: **Early versus later rhythm analysis in patients with out-of-hospital cardiac arrest.** *The New England journal of medicine* 2011, **365:**787-797.

84. Damsgaard CT, Dalskov SM, Laursen RP, Ritz C, Hjorth MF, Lauritzen L, Sorensen LB, Petersen RA, Andersen MR, Stender S, et al: **Provision of healthy school meals does not affect the metabolic syndrome score in 8-11-year-old children, but reduces cardiometabolic risk markers despite increasing waist circumference.** *The British journal of nutrition* 2014, **112:**1826-1836.

85. Berndt N, Bolman C, Froelicher ES, Mudde A, Candel M, de Vries H, Lechner L: **Effectiveness of a telephone delivered and a face-to-face delivered counseling intervention for smoking cessation in patients with coronary heart disease: a 6-month follow-up.** *Journal of behavioral medicine* 2014, **37:**709-724.

86. Oostdijk EAN, Kesecioglu J, Schultz MJ, Visser CE, de Jonge E, van Essen EHR, Bernards AT, Purmer I, Brimicombe R, Bergmans D, et al: **Effects of decontamination of the oropharynx and intestinal tract on antibiotic resistance in ICUs: a randomized clinical trial.** *Jama* 2014, **312:**1429-1437.

87. Reddy SK, Bailey MJ, Beasley RW, Bellomo R, Henderson SJ, Mackle DM, McArthur CJ, Mehrtens JE, Myburgh JA, McGuinness SP, et al: **A protocol for the 0.9% saline versus Plasma-Lyte 148 for intensive care fluid therapy (SPLIT) study.** *Critical care and resuscitation : journal of the Australasian Academy of Critical Care Medicine* 2014, **16:**274-279.

88. van Duijn PJ, Bonten MJ: **Antibiotic rotation strategies to reduce antimicrobial resistance in Gram-negative bacteria in European intensive care units: study protocol for a cluster-randomized crossover controlled trial.** *Trials* 2014, **15:**277.

89. Pourrat X, Roux C, Bouzige B, Garnier V, Develay A, Allenet B, Fraysse M, Halimi JM, Grassin J, Giraudeau B: **Impact of drug reconciliation at discharge and communication between hospital and community pharmacists on drug-related problems: study protocol for a randomized controlled trial.** *Trials* 2014, **15:**260.

90. Roisin S, Laurent C, Denis O, Dramaix M, Nonhoff C, Hallin M, Byl B, Struelens MJ: **Impact of rapid molecular screening at hospital admission on nosocomial transmission of methicillin-resistant staphylococcus aureus: cluster randomised trial.** *PloS one* 2014, **9:**e96310.

91. Surka S, Edirippulige S, Steyn K, Gaziano T, Puoane T, Levitt N: **Evaluating the use of mobile phone technology to enhance cardiovascular disease screening by community health workers.** *International Journal of Medical Informatics* 2014, **83:**648-654.

Appendix C

Additional files accompanying Chapter 4 - Systematic review search strategies, Supplementary tables S1 and S2

Search strategies

Ovid MEDLINE search

CROSS OVER TERMS

1. (cross-over or cross?over or "cross* over").tw.

2. (switch-over or switch?over or "switch* over" or switch-back or switch?back or "switch* back" or switched).tw.

3. ((change-over or change?over or "change* over") not ((change-over or change?over or "change* over") adj1 time)).tw.

4. (ab*ba* adj3 design*).tw.

5. exp Cross-Over Studies/

6. 1 or 2 or 3 or 4 or 5

CLUSTER ALLOCATION TERMS

7. ((unit$1 or school$1 or hospital$1 or cluster* or region$1 or ward* or practice* or communit* or population* or facility or facilities or practitioner*) adj15 random*).tw.

8. ((unit$1 or school$1 or hospital$1 or cluster* or region$1 or ward* or practice* or communit* or population* or facility or facilities or practitioner*) adj15 interven*).tw.

9. ((group* adj random*) or (group* adj interven*)).tw.

10. 7 or 8 or 9

HUMANS ONLY

11. Humans/

12. Animals/

13. 12 not 11

COMBINE CONCEPTS

14. 6 and 10

15. 14 not 13

PubMed search

CROSS OVER TERMS

1. "cross-over"[tiab] OR crossover[tiab] OR "cross over" [tiab] OR "crossed over"[tiab]

2. "switch-over"[tiab] OR switchover[tiab] OR "switch over"[tiab] OR "switch-back"[tiab] OR switchback[tiab] OR "switch back"[tiab] OR switched[tiab]

3. (change-over[tiab] OR changeover[tiab] OR "change over"[tiab] OR "changed over"[tiab] OR "changes over"[tiab]) not ("change-over time"[tiab] OR "changeover time"[tiab] OR "change over time"[tiab] OR "changed over time"[tiab] OR "changes over time"[tiab])

4. ab*ba[tiab]

5. Cross-Over Studies[mh]

6. #1 OR #2 OR #3 OR #4 OR #5

CLUSTER ALLOCATION TERMS

7. (cluster-randomi*[tiab] OR “cluster randomised”[tiab] OR “cluster randomized”[tiab] OR “cluster randomization”[tiab] OR “cluster randomisation”[tiab])

HUMANS ONLY

8. (Animals[mh] NOT Humans[mh])

COMBINE CONCEPTS

9. #6 AND #7

10.#9 NOT 8

11.#10 NOT MEDLINE[sb]

EMBASE search via embase.com

CROSS OVER TERMS

1. (cross-over or crossover or "cross over" or "crosses over" or "crossed over" or "crossing over"):ti:ab

2. (switch-over or switchover or "switch over" or "switches over" or "switched over" or switch-back or "switchback" or "switch back" or "switches back" or "switched back" or switched):ti:ab

3. ((change-over or changeover or "change over" or "changes over" or "changed over") not ((change-over or changeover or "change over" or "changes over" or "changed over") near/1 time)):ti:ab

4. (abba near/3 design):ti:ab or (abba near/3 designs):ti:ab

5. “crossover procedure”/exp

6. #1 or #2 or #3 or #4 or #5

CLUSTER ALLOCATION TERMS

7. ((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) near/15 (random or randomly or randomise or randomize or randomised or randomized or randomises or randomizes or randomisation or randomization)):ti:ab

8. ((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) near/15 (intervene or intervention or interventions)):ti:ab

9. ((group or groups or grouped) near/1 (random or randomly or randomise or randomize or randomised or randomized or randomises or randomizes or randomisation or randomization)):ti:ab or ((group or groups or grouped) near/1 (intervene or intervention or interventions)):ti:ab

10. #7 or #8 or #9

HUMANS ONLY

11. ‘animal’ not ‘human’

COMBINE CONCEPTS

12. #6 and #10

13. #12 not #11

14. #13 not ‘medline’

CINAHL Plus search

CROSS OVER TERMS

1. TI (("cross-over" or "cross?over" or "cross* over")) OR AB (("cross-over" or "cross?over" or "cross* over"))

2. TI (("switch-over" or "switch?over" or "switch* over" or "switch-back" or "switch?back" or "switch* back" or switched)) OR AB (("switch-over" or "switch?over" or "switch* over" or "switch-back" or "switch?back" or "switch* back" or switched))

3. TI ((("change-over" or "change?over" or "change* over") not (("change-over" or "change?over" or "change* over") n1 time))) OR AB ((("change-over" or "change?over" or "change* over") not (("change-over" or "change?over" or "change* over") n1 time)))

4. TI (ab*ba* n3 design*) OR AB (ab*ba* n3 design*)

5. (MH "Crossover Design")

6. S1 or S2 or S3 or S4 or S5

CLUSTER ALLOCATION TERMS

7. TI (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 random*)) OR AB (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 random*))

8. TI (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 interven*)) OR AB (((unit or units or school or schools or hospital or hospitals or cluster or clusters or region or regions or ward or wards or practice or practices or community or communities or population or populations or facility or facilities or practitioner or practitioners) n15 interven*))

9. TI (((group* n1 random*) or (group* n1 interven*))) OR AB (((group* n1 random*) or (group* n1 interven*)))

10. S7 or S8 or S9

HUMANS ONLY

11. (MH "Human")

12. (MH "Animals")

13. S12 not S11

COMBINE CONCEPTS

14. S6 and S10

15. S14 not S13

16. Exclude MEDLINE

Table S1: Country where the trial was conducted

Country	**n (%)** **(N = 83)**
USA	29 (35%)
UK	10 (12%)
The Netherlands	8 (10%)
Canada	5 (6%)
More than 1 country	3 (4%)
Australia	3 (4%)
France	3 (4%)
China	2 (2%)
Denmark	2 (2%)
Germany	2 (2%)
Sweden	2 (2%)
Austria	1 (1%)
Belgium	1 (1%)
Estonia	1 (1%)
Finland	1 (1%)
Greece	1 (1%)
Kenya*	1 (1%)
Pakistan	1 (1%)
South Korea	1 (1%)
South Africa	1 (1%)
Switzerland	1 (1%)
Taiwan	1 (1%)
Tanzania*	1 (1%)
Thailand*	1 (1%)
Zambia*	1 (1%)

* Developing countries as classified by the International Monetary Fund, 2015

Table S2: Type of randomised cluster

Randomised cluster type	**n (%) (N = 83)**
Hospital or ward	40 (48%)
ICU	17
Other wards	23
Individual health care provider	13 (16%)
School or class	11 (13%)
Class or classroom	7
School	3
Group of students	1
Emergency medical team	6 (7%)
Primary care practice	3 (4%)
Individual (mouth, muscles)	2 (2%)
Dementia unit or facility	2 (2%)
Aged care facility	2 (2%)
Community or geographical area	1 (1%)
Household or family group	1 (1%)
Workplace	1 (1%)
Outpatient clinic	1 (1%)

ICU: Intensive care unit

Appendix D

Blog article accompanying Chapter 4 - A call for cluster randomized cross-over trial reporting guidelines

A call for cluster randomized cross-over trial reporting guidelines - On Medicine

Sarah Arnup & Joanne McKenzie 7 Dec 2016

When asking a research question, you need to ensure the trial design you choose is the most appropriate for that particular question. All have particular benefits and pitfalls that need to be considered and addressed when it comes to reporting. The cluster randomized cross-over trial is one of these potential choices and here, Sarah Arnup and Joanne McKenzie discuss their research into the reporting of this trial design, published today in *Trials*, and highlight the need for a CONSORT extension.

Depending on your research question, a cluster randomised cross-over trial might be the best choice but what should you ensure to cover during reporting?

Geralt

To be able to determine whether a healthcare intervention is effective, we need randomized trials that are appropriately designed and conducted. But, appropriate design and conduct of a trial in and of itself is not enough.

The trial's methods, conduct and results need to be comprehensively reported to convince others of the validity of the results, allow replication of the trial methodology, and the incorporation of the results in synthesis products such as meta-analysis.

Benefits and pitfalls of the cluster randomized cross-over trial design

> The cluster randomized cross-over (CRXO) design provides an exciting opportunity to evaluate interventions in settings where cluster randomization is required, but a parallel group cluster randomized design is not feasible…

The cluster randomized cross-over (CRXO) design (see here and here) provides an exciting opportunity to evaluate interventions in settings where cluster randomization is required, but a parallel group cluster randomized design is not feasible because the number of required clusters is prohibitively large.

In a CRXO trial, schools, hospitals or other groups of individuals ("clusters") are assigned to a sequence of interventions. This design differs to a parallel group cluster randomized design where each cluster is assigned to just one intervention. The CRXO design incorporates a cross-over of interventions at the level of the cluster.

The cross-over aspect means that the interventions are compared within cluster, and hence the between cluster variation is removed from the estimate of the difference between the interventions. As a result, the design generally requires fewer participants than the parallel group cluster randomised design.

While the CRXO design has the potential benefit of requiring fewer participants, the design is not always appropriate for evaluating interventions (e.g. where it is not possible to 'remove' an intervention, such as an educational intervention targeted at clinicians within a hospital); has greater potential for bias than other designs (e.g. through carry-over effects, identification/recruitment of participants into the trial); and, has more complexity in the sample size calculations and statistical methods.

But there's huge room for improvement when it comes to reporting

Our previous research found that in a cohort of CRXO trials, appropriate statistical methods for sample size calculations and statistical analyses were rarely used.

In our paper published today in *Trials*, we examined the completeness of reporting of CRXO trials, and found that they are not well reported. Specifically:

> The justification for using the design was infrequently stated.

- Inconsistent language was used in the title and abstract to describe the trial design. As a result, many CRXO trials may not be indexed appropriately in electronic databases.
- The justification for using the design was infrequently stated. This information is essential to determine if the CRXO design is best suited to address the research question.
- The potential for carryover was infrequently discussed. Clear reporting of the method used to reduce the risk of carryover is required to allow readers to assess the risk of bias to the intervention effect arising from the potential

carryover.
- A justification for the sample size was presented in only half of the trials, and incomplete reporting of the sample size calculation within those trials was common. Clear sample size reporting is important for replicability, transparency and ethical reasons.
- Reporting of which researchers and participants were aware of the allocated interventions was incomplete. The trial is at risk of selection bias (a systematic difference in the characteristics of the groups of individuals that are compared between interventions) if: the person assigning the intervention sequences to clusters is aware of the intervention sequence; the person responsible for recruiting participants knew which intervention sequence had been assigned to the cluster; or individuals within a cluster are aware of the cluster allocation before giving consent.

A justification for the sample size was presented in only half of the trials, and incomplete reporting of the sample size calculation within those trials was common.

Our findings are not unsurprising; they are aligned with a large body of evidence that has consistently demonstrated sub-optimal reporting of randomized trials (see here, here and here).

However, there is evidence to suggest that reporting guidelines improve the completeness of reporting.

What next for cluster randomized cross-over trial reporting?

A potential explanation for the incomplete reporting in the case of CRXO trials is that there are no reporting guidelines that address the unique features of this design.

As part of our research, we therefore proposed reporting items for the CRXO trial design. We based our proposed items on the 2012 cluster randomized trials extension to the CONSORT statement and relevant items that have been proposed for reporting a related design, the stepped wedge design.

A potential explanation for the incomplete reporting in the case of CRXO trials is that there are no reporting guidelines that address the unique features of this design.

The results of our study highlight the need for a CONSORT extension for CRXO trials, with the items proposed in our paper providing the starting point for such an extension. However, even with such an extension, there is no guarantee that these guidelines will be used.

This raises the broader question, how do we get researchers to use reporting guidelines? While there is some evidence to suggest that journal endorsement of the CONSORT Statement improves adherence to reporting guidelines, reporting is still below acceptable levels. Research on the barriers and enablers to appropriate reporting would be valuable, as would evaluations of interventions to improve design, conduct and reporting of randomized trials.

View the latest posts on the On Medicine homepage

Appendix E

Additional files accompanying Chapter 5 - Sample size Stata do files for continous and binary outcomes

Stata do file for continous outcomes sample size calculation

```
version 13
clear all
set more off

*** INPUT PARAMETERS ***
*************************************************************************
local rho = 0.032               // <--- INPUT Estimate of the within-cluster
                                            within-period correlation

local eta = 0.028               // <--- INPUT Estimate of the within-cluster
                                            between-period correlation

local sigma_t = 1.2             // <--- INPUT Estimate of standard deviation
                                            of outcome data

local m = 200                   // <--- INPUT number of patients
                                            PER CLUSTER-PERIOD

local diff = 0.1                // <--- INPUT treatment difference

local za = 1.96                 // Type I error rate 5%
                                  (To replicate paper values, set to 1.96)
```

```
local zb = 0.84                 // Power 80%
                                   (To replicate paper values, set to 0.84)

/*
local za = invnormal(0.975)     // Type I error rate 5%
local zb = invnormal(0.8)       // Power 80%
*/
*******************************************************************************

* CALCULATIONS

* Individually Randomised Controlled Trial
local irct = 2 * (`za'+`zb')^2 * ///
    ( (2*(`sigma_t')^2) / (`diff')^2 ) * (1-`rho')

local irct_clusters = `irct' / `m'


* Cluster Randomised Controlled Trial
local crct = 2 * (`za'+`zb')^2 * ///
    ( (2*(`sigma_t')^2) / (`diff')^2 ) * (1+(`m'-1)*`rho') + 2*`m'

local crct_clusters = `crct' / `m'


* Cluster Randomised Crossover Trial
local crxo = 2 *(`za'+`zb')^2 * ///
    ( (2*(`sigma_t')^2) / (`diff')^2 ) ///
    * (1+(`m'-1)*`rho'-`m'*`eta') + 4*`m'

local crxo_clusters = `crxo' / (2 * `m')


*******************************************************************************
* OUTPUT

* Individually Randomised Controlled Trial
* Number of required individuals:
di ceil(`irct')
* Number of required clusters:
di ceil(`irct_clusters')
```

```
* Cluster Randomised Controlled Trial
* Number of required individuals:
di ceil(‘crct’)
* Number of required clusters:
di ceil(‘crct_clusters’)

* Cluster Randomised Crossover Trial
* Number of required individuals:
di ceil(‘crxo’)
* Number of required clusters:
di ceil(‘crxo_clusters’)

exit
```

Stata do file for binary outcomes sample size calculation

```
version 13
clear all
set more off

*************************************************************************
*** INPUT PARAMETERS ***
*************************************************************************
local rho = 0.011               // <--- INPUT Estimate of the within-cluster
                                           within-period correlation

local eta = 0.008               // <--- INPUT Estimate of the within-cluster
                                           between-period correlation

local m = 1200                  // <--- INPUT number of patients
                                           PER CLUSTER-PERIOD

local p1 = 0.089                // <--- INPUT event rate for treatment 1
local p2 = 0.074                // <--- INPUT event rate for treatment 2

local za = 1.96                 // Type I error rate 5%
                                   (To replicate paper values, set to 1.96)

local zb = 0.84                 // Power 80%
                                  (To replicate paper values, set to 0.84)

/*
local za = invnormal(0.975)  // Type I error rate 5%
local zb = invnormal(0.8)  // Power 80%
*/
*************************************************************************
* CALCULATIONS

* Individually Randomised Controlled Trial
local irct = 2 * (`za'+`zb')^2 * ///
    ( (`p1'*(1-`p1') + `p2'*(1-`p2')) / (`p1'-`p2')^2 ) * (1-`rho')

local irct_clusters = `irct' / `m'

* Cluster Randomised Controlled Trial
local crct = 2 * (`za'+`zb')^2 * ///
```

```
    ( (‘p1’*(1-‘p1’) + ‘p2’*(1-‘p2’)) / (‘p1’-‘p2’)^2 ) ///
    * (1+(‘m’-1)*‘rho’) + 2*‘m’

local crct_clusters = ‘crct’ / ‘m’

* Cluster Randomised Crossover Trial
local crxo = 2 *(‘za’+‘zb’)^2 * ///
    ( (‘p1’*(1-‘p1’) + ‘p2’*(1-‘p2’)) / (‘p1’-‘p2’)^2 ) ///
    * (1+(‘m’-1)*‘rho’-‘m’*‘eta’) + 4*‘m’

local crxo_clusters = ‘crxo’ / (2 * ‘m’)

*******************************************************************************
* OUTPUT
* Individually Randomised Controlled Trial
* Number of required individuals:
di ceil(‘irct’)
* Number of required clusters:
di ceil(‘irct_clusters’)

* Cluster Randomised Controlled Trial
* Number of required individuals:
di ceil(‘crct’)
* Number of required clusters:
di ceil(‘crct_clusters’)

* Cluster Randomised Crossover Trial
* Number of required individuals:
di ceil(‘crxo’)
* Number of required clusters:
di ceil(‘crxo_clusters’)


exit
```

Appendix F

Supplementary material accompanying Chapter 6 - Appendices I to IV, Stata ado file

Appendix I: This appendix was part of the submitted manuscript and has been peer reviewed. It is posted as supplied by the authors.

Unstratified sample size formula for risk ratio in cluster randomised crossover trials

The number of ICUs required to detect a risk ratio (RR) with power 1-β between two interventions with significance level α, is:

$$N_{ICU} = \frac{\left(z_{\alpha/2}+z_{\beta}\right)^2}{\log(RR)^2}\frac{\bar{p}(1-\bar{p})}{p_A^2}\frac{(1+RR^2)}{RR^2\, n_h}\left(1+(n_h-1)\rho-\frac{(2\,RR\, n_h)}{(1+RR^2)}\eta\right) \qquad \textbf{Equation 1}$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values corresponding to the upper tail probabilities of α/2 and β, respectively; RR = $\frac{p_B}{p_A}$ is the specified risk ratio, where p_A and p_B are the outcome rates for arms *A* and *B* respectively; $\bar{p}$ = 0.5 ($p_A + p_B$); ρ is the within-cluster within-period correlation; η is the within-cluster between-period correlation; and $n_h = m/\sum_{i=1}^{m}\frac{1}{n_i}$ is the harmonic mean ICU size, where m is the number of ICUs used in the calculation and n_i is the number of admissions in ICU *i*.

Unstratified sample size formula for risk ratio in parallel-group cluster randomised trials

The number of ICUs required to detect a risk ratio (RR) with power 1-β between two interventions with significance level α, is:

$$N_{ICU} = \frac{\left(z_{\alpha/2}+z_{\beta}\right)^2}{\log(RR)^2}\frac{\bar{p}(1-\bar{p})}{p_A^2}\frac{(1+RR^2)}{RR^2\, n_h}(1+(n_h-1)\rho) \qquad \textbf{Equation 2}$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values corresponding to the upper tail probabilities of α/2 and β, respectively; RR = $\frac{p_B}{p_A}$ is the specified risk ratio, where p_A and p_B are the outcome rates for arms *A* and *B* respectively; $\bar{p}$ = 0.5 ($p_A + p_B$); ρ is the within-cluster correlation; and $n_h = m/\sum_{i=1}^{m}\frac{1}{n_i}$ is the harmonic mean ICU size, where m is the number of ICUs used in the calculation and n_i is the number of admissions in ICU *i*.

Stratified sample size formula for risk ratio in cluster randomised crossover trials

The number of ICUs required to achieve power (1-β) to detect a risk ratio (RR) between two interventions with significance level α, is:

$$N_{ICU} = \frac{(z_{\alpha/2}+z_{\beta})^2}{\log(RR)^2}\sum_s f_s V_s \qquad \textbf{Equation 3}$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values corresponding to the upper tail probabilities of α/2 and β, respectively; RR = $\frac{p_{B_s}}{p_{A_s}}$ is the specified risk ratio, where p_{A_s} and p_{B_s} are the outcome rates for arms *A* and *B* respectively in stratum *s*; and f_s is the fraction of total ICUs recruited from stratum *s*.

V_s is the variance of the outcome in strata *s*, and is given by*:*

$$V_s = \frac{\bar{p_s}(1-\bar{p_s})}{p_{A_s}^2}\,\frac{(1+RR^2)}{RR^2\,n_{h_s}}\left(1+(n_{h_s}-1)\rho_s - \frac{(2\,RR\,n_{h_s})}{(1+RR^2)}\eta_s\right)$$

where $\bar{p_s}$ = 0.5 ($p_{A_s} + p_{B_s}$); ρ_s is the within-cluster within-period correlation (WPC) in stratum *s*; η_s is the within-cluster between-period correlation (BPC) in stratum *s*; and $n_{h_s} = m_s/\sum_{i=1}^{m_s}\frac{1}{n_{si}}$ is the harmonic mean ICU size for strata *s*, m_s is the number of ICUs used in the calculation from strata *s*, and n_{si} is the number of admissions in ICU in stratum *s*.

Appendix II: This appendix was part of the submitted manuscript and has been peer reviewed. It is posted as supplied by the authors.

In this section we describe the methods used to obtain the data required to perform sample size calculations for a CRXO trial.

Data inclusion

We extracted data from ICUs with a minimum of 200 admissions during each year from 2010 to 2015 inclusive. ICUs with fewer admissions would not typically be invited to participate in a clinical trial. Admissions with diagnoses of coronary artery bypass grafts and cardiac valve surgery were excluded also, consistent with the exclusion criteria for many clinical trials conducted within intensive care. In addition, to remove potentially unreliable data, we excluded the following data: sites with large fluctuations in admission totals where the reason for the fluctuations were not known; and individual patient records with inconsistencies in admission and discharge times; or individual patient records with inconsistencies between ICU and hospital mortality outcomes.

Patient records from 115 of the 172 ICUs that submitted data during 2010 to 2015 were analysed. All except four excluded ICUs had less than 200 annual admissions.

Analysis

We present the within-period correlation (WPC) and between-period correlation (BPC) for all cause in-hospital mortality, the annual number of ICU admissions, and baseline mortality rates for a 12-month period length. These data are presented both for all patient records meeting the inclusion requirements; and, additionally, for patients receiving ventilation.

When there is variability in number of admissions between ICUs, the appropriate statistic for summarising ICU size to perform a sample size calculation for the number of ICU required is the *harmonic* mean rather than the arithmetic mean. We report both the arithmetic and harmonic mean in Tables 1 and 2.

We assume that the number of patients per ICU is the same in both periods of the design. Although this is likely to be approximately correct, when the number of patients differ across the periods, and hence the harmonic mean cluster-period size differs across the periods, a pragmatic approximation approach is to use the average of the harmonic means in the sample size formula.

We estimate the values of the WPC and BPC by applying previously published methods for binary outcomes to the mortality data [2,17]: the WPC is estimated by fitting the analysis of variance (ANOVA) estimator for the intracluster correlation; and the BPC by fitting the Pearson pairwise estimator. The WPC is a weighted mean of 2010 to 2015 period, while the BPC is taken as the arithmetic mean of the three adjacent pairs of 12 month periods in 2010 to 2015.

Although we stated earlier (Section 2) that the maximum value of the BPC is equal to the WPC, the WPC and the BPC are computed independently from the data. As a result, the *estimate* of the BPC can be greater than the WPC. This scenario is likely to be due to sampling variation in the estimation of the WPC and the BPC, rather than the BPC truly being greater than the WPC. In cases where the estimated BPC is greater than the WPC, we recommend setting the BPC as equal to the WPC.

All calculations were performed in Stata 14.2 (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.).

Appendix III: This appendix was part of the submitted manuscript and has been peer reviewed. It is posted as supplied by the authors.

Suppose a CRXO trial is being planned to compare usual infection control with selective digestive decontamination in all patients receiving mechanical ventilation. The effect to be detected is a 10% reduction in all cause in-hospital mortality (RR = 0.90). The investigators wish to detect this effect with 90% power, assuming a 5% significance level. The trial is restricted to patients admitted to tertiary ICUs.

The sample size formula requires values of: the harmonic mean number of patients per ICU per 12-month period (see Appendix I for definition); the event rate of the outcome, in this example, mortality; and the WPC and the BPC. Using the estimates of these values calculated from the APD data (see the combined row in Table 2), gives a sample size requirement of 61 ICUs.

ICU Strata	Ventilated (% of total admissions)	Average mean annual number of admissions	Harmonic mean annual number of admissions	Mortality Rate	WPC	BPC
Tertiary	45.3%	614	455	15.3%	0.008	0.007
Metro/Rural	25.4%	165	65	15.4%	0.008	0.007
Combined	36.9%	227	97	15.3%	0.008	0.007

Table 2: The percentage of total admissions where the patient required mechanical ventilation, annual number of admissions, harmonic mean annual number of admissions, mortality rate, and within-period correlation (WPC) and between-period correlation (BPC) for 12-month period length for all included patients, by ICU strata and combined across all strata.

Appendix IV: This appendix was part of the submitted manuscript and has been peer reviewed. It is posted as supplied by the authors.

In this section, we expand the sample size calculation presented in Section 3 to allow for the differences in the estimates of annual number of admissions, mortality rates, and correlations (WPC, BPC) between the tertiary and metropolitan/rural ICU strata. Again, the effect to be detected is a 10% reduction in all cause in-hospital mortality (RR = 0.90). The investigators wish to detect this effect with 90% power, assuming a 5% significance level. The sample size formula for the required number of ICUs is provided by Equation 3 (Appendix I).

The sample size formula requires values, *in each stratum*, of: the harmonic mean number of patients per ICU per 12-month period (see Appendix I for definition); the event rate of the outcome, in this example, mortality; and the WPC and the BPC for both the tertiary and the metropolitan/rural ICU strata. Using the estimates of these values calculated from the APD data (see Table 1), Figure 2 gives the sample size requirement for the total number of ICUs, and the number of ICUs from each strata, for different combinations of numbers of ICUs from each strata. For example, if only tertiary ICUs are included in the trial, the total required number of ICUs is 39 (Combination 1). Alternatively, 32 tertiary ICUs and 25 metropolitan/rural ICUs, giving a total of 57 ICUs, could be included in the trial (Combination 8).

Stata ado file

```
/*
Program to estimate the power achieved to detect a specified Relative Risk
when the sample size parameters vary across multiple strata.

Example call:
crxopower, strata(2) rr(0.85) rho(0.01, 0.01) eta(0.005, 0.005) ///
  pa(0.12, 0.12) hmean(840, 840) numclus(14, 14)

Options:

strata:     The number of strata.
rr:         The Relative Risk of between the event rate between two
            interventions.

rho:        A list of the within-cluster within-period ICC for each strata.
eta:        A list of the within-cluster between-period ICC for each strata.
pa:         A list of the event rate in the control intervention for each
            strata.

hmean:      A list of the harmonic mean cluster size for each strata.
numclus:    A list of the number of clusters to be included from each
            strata.

Optional options:
alpha:     Two-sided Type I error rate, default value is 0.05


History:
12 Dec 2017, Sarah Arnup:  First version

*/

capture program drop crxopower
program define crxopower
version 14.2

syntax, strata(integer) rr(real) ///
 rho(numlist) eta(numlist) pa(numlist) hmean(numlist) numclus(numlist) ///
 [alpha(real 0.05)] [beta(real 0.9)]
```

```
*************************************************************************
* error checking
*************************************************************************

if `strata' < 1 {
 di _newline in red ///
  "Strata must be greater than 0, current value is `strata' "
 exit
}

if `rr' < 0 {
 di _newline in red ///
  "Relative Risk must be greater than 0, current value is `rr' "
 exit
}

foreach v in rho eta pa hmean numclus {
 local temp: word count ``v''
 if `temp' != `strata' {
  di _newline in red "Invalid number of entries in `v' "
  di _newline in red "There are `strata' strata, but `temp' entries in `v' "
  exit
}
}

foreach v in rho eta pa {
 forval s=1/`strata' {
  local b: word `s' of ``v''
  if (`b' < 0) | (`b' > 1) {
   di _newline in red "`v' must be between 0 and 1. Strata `s' `v' is `b' "
   exit
}
}
}

foreach v in hmean numclus {
 forval s=1/`strata' {
  local b: word `s' of ``v''
   if (`b' < 1) {
    di _newline in red "`v' must be greater than 1. Strata `s' `v' is `b' "
    exit
}
```

```
}
}

forval s=1/‘strata’ {
 local b1: word ‘s’ of ‘rho’
 local b2: word ‘s’ of ‘eta’
 if (‘b1’ < ‘b2’) {
  di _newline in red ///
   "eta must not be greater than rho. Strata ‘s’ rho is ‘b1’, eta is ‘b2’ "
  exit
}
}

if (‘alpha’ < 0) | (‘alpha’ > 1) {
 di _newline in red "Alpha must be between 0 and 1, current value ‘alpha’ "
 exit
}

if (‘beta’ < 0) | (‘beta’ > 1) {
 di _newline in red "Beta must be between 0 and 1, current value ‘beta’ "
 exit
}

*******************************************************************************
* define parameters
*******************************************************************************

* power, type 1 error, relative risk
local z_alpha = ‘alpha’/2
local z_beta = ‘beta’
local betarr = log(‘rr’)

* rho, eta, event rate, cluster size, number of clusters
forval s=1/‘strata’ {
 local rho_‘s’: word ‘s’ of ‘rho’
 local eta_‘s’: word ‘s’ of ‘eta’
 local pa_‘s’: word ‘s’ of ‘pa’
 local pb_‘s’= ‘rr’ * ‘pa_‘s’’
 local pbar_‘s’ = (‘pa_‘s’’ + ‘pb_‘s’’)/2
 local n_‘s’: word ‘s’ of ‘hmean’
 local c_‘s’: word ‘s’ of ‘numclus’
}
```

```
* total number of clusters
local c_sum = 0 // initial value
forval s=1/‘strata’ {
 local c_sum = ‘c_sum’ + ‘c_‘s’’
}

* strata weighting
forval s=1/‘strata’ {
 local f_‘s’ = ‘c_‘s’’ / ‘c_sum’
}

*************************************************************************
* power formula
*************************************************************************

* variance in each strata
forval s=1/‘strata’ {
 local term1_‘s’ = ///
  (2 * ‘pbar_‘s’’ * (1 - ‘pbar_‘s’’)) / (‘pa_‘s’’ * ‘pb_‘s’’ * ‘c_‘s’’)

 local term2_‘s’ = ///
  ((1 - ‘rho_‘s’’) * ((‘pa_‘s’’)^2 ///
   + (‘pb_‘s’’)^2)) / (2 * ‘n_‘s’’ * ‘pa_‘s’’ * ‘pb_‘s’’)

 local term3_‘s’ = ///
  ‘rho_‘s’’ * ((((‘pa_‘s’’)^2 + (‘pb_‘s’’)^2) / (2 * ‘pa_‘s’’ * ‘pb_‘s’’))

 local v_‘s’ =  ‘term1_‘s’’ * (‘term2_‘s’’ + ‘term3_‘s’’ - ‘eta_‘s’’)
}

* total variance
local v = 0 // initial value
forval s=1/‘strata’ {
 local v = ‘v’ + ((‘f_‘s’’)^2 * ‘v_‘s’’)
}

* power
local power = (normal(invnormal(‘z_alpha’)-sqrt(((‘betarr’)^2) / (‘v’))) ///
 + normal(invnormal(‘z_alpha’)+sqrt(((‘betarr’)^2) / (‘v’))))*100
```

```
*************************************************************************
* output results
*************************************************************************

di "Relative Risk: `rr'"
di "Two-sided Type 1 error: `alpha'"
di "Power: " round(`power',0.1) "%"
di "Total number of clusters: `c_sum'"

forval s=1/`strata' {
 di "Strata `s'"
 di "Rho: `rho_`s''"
 di "Eta: `eta_`s''"
 di "Event rate in Group A: `pa_`s''"
 di "Event rate in Group B: `pb_`s''"
 di "Harmonic mean of cluster size: `n_`s''"
 di "Number of clusters: `c_`s''"
 di "Fraction of total clusters: `f_`s''"
}

end
```

Appendix G

Sample size formulae for the risk ratio in stratified cluster randomised crossover trials

In this appendix, details for the derivation of the sample size formulae presented in Appendices F are provided. These derivations have not been previously published.

G.1 Model for the probability of an event in a stratified cluster randomised crossover trial

Consider a cluster randomised crossover trial designed to estimate the effect of an intervention on a binary outcome. Assume there are $s = 1, ..., S$ strata with $i = 1, ..., c_s$ clusters in each strata. Each cluster has $j = 1, 2$ periods and $k = 1, ..., n_{sij}$ individuals in each cluster-period in stratum s. There are two interventions $x_{ij} = 0, x_{ij} = 1$. For simplicity, assume there are no fixed period effects. The probability of the event of interest can be modelled with the following marginal binomial-log link model

$$\begin{aligned}
\log P(Y_{sijk} = 1) &= \alpha + \gamma_s + \beta x_{sij}, \\
\text{Corr}(Y_{sijk}, Y_{sijk'}) &= \rho_s, \\
\text{Corr}(Y_{sijk}, Y_{sij'k'}) &= \eta_s, \\
\text{and Corr}(Y_{sijk}, Y_{s'i'j'k'}) &= 0, \text{ for } s \neq s' \text{ or } i \neq i'.
\end{aligned}$$

where γ_s are fixed effects for stratum $s = 1, ...S$, with $\gamma_s = 0$ for identifiability, and β is the fixed effect corresponding to $x = 1$.

Let $\pi_{sl} = P(Y_{sijk} = 1 | X_{sij} = l) = e^{\alpha + \gamma_s + \beta l}$ be the marginal probability of the event of interest in the s'th stratum under treatment condition $X_{sij} = l$. Then, the parameter of interest is the marginal risk ratio $\psi = \pi_{s1}/\pi_{s0} = e^{\beta}$.

G.2 Estimating the intervention effect

The parameter $\beta = \log(\psi)$ can be estimated within each cluster, within each stratum, by the difference in the log of the observed proportions under each treatment condition p_{sij}, such that $\hat{\beta}_{si} = \log(p_{si1}/p_{si0})$.

G.2.1 Stratum-specific estimator of the intervention effect

By analogy with Forbes(2015) [24] we obtain a stratum-specific estimate of β using an unweighted average of cluster-specific estimates

$$\hat{\beta}_s = \frac{\sum_{i=1}^{c_s} \hat{\beta}_{si}}{c_s}$$

We can also define a stratum-specific estimate of β using a weighted average of cluster-specific estimates with

$$\hat{\beta}_{sw} = \frac{\sum_{i=1}^{c_s} w_{si}\hat{\beta}_{si}}{\sum_{i=1}^{c_s} w_{si}}$$

where the cluster-specific weights are given by w_{si}. Following Forbes(2015) [24], if $w_{si} = 1$, then we obtain the unweighted stratum-specific estimate $\hat{\beta}_s$. In the following we consider the unweighted estimator $\hat{\beta}_s$, shown to be near optimal with reasonably large cluster sizes, as is the case in ICUs.

G.2.2 Overall pooled estimator of the intervention effect

We can obtain an estimate of β by combining the unweighted stratum-specific estimates, using a weighted average of the stratum-specific estimates, with

$$\begin{aligned} \hat{\beta} &= \sum_{s=1}^{S} f_s \hat{\beta}_s \\ &= \sum_{s=1}^{S} f_s \frac{\sum_{i=1}^{c_s} \hat{\beta}_{si}}{c_s} \end{aligned} \tag{G.1}$$

where the stratum-specific weights are given by f_s, and $\sum_{s=1}^{S} f_s = 1$.

G.2.3 Variance of the intervention effect estimator $\hat{\beta}_s$

The variance of the intervention effect estimator given in Eqn G.1 is given by

$$\begin{aligned} V(\hat{\beta}) &= \sum_{s=1}^{S} f_s^2 V(\hat{\beta}_s) \\ &= \sum_{s=1}^{S} f_s^2 \frac{\sum_{i=1}^{c_s} V(\hat{\beta}_{si})}{c_s^2} \end{aligned} \tag{G.2}$$

Hence we need to determine the variance of the cluster-specific intervention effect estimators,

$$\begin{aligned} V(\hat{\beta}_{si}) &= V(\log(p_{si1}) - \log(p_{si0})) \\ &= V(\log(p_{si1})) + V(\log(p_{si0})) - 2 \times \mathrm{Cov}(\log(p_{si1}), \log(p_{si0})) \end{aligned} \tag{G.3}$$

Using the multivariate Delta Method to find the variances of the logarithmic terms,

$$\begin{aligned} \log p &= \log \pi + (p - \pi) \log'(\pi) \\ V(\log p) &= V(p \log'(\pi)) \\ V(\log p) &= (\log'(\pi))^2 V(p) \\ V(\log p) &= \frac{1}{\pi^2} V(p) \end{aligned}$$

and therefore,

$$\begin{aligned} V(\log p_{sil}) &= \frac{1}{{\pi_{sl}}^2} V(p_{sil}), \\ \text{and } \mathrm{Cov}(\log(p_{si1}), \log(p_{si0})) &= \frac{1}{\pi_{s1}\pi_{s0}} \mathrm{Cov}\left(p_{si1}, p_{si0}\right) \end{aligned}$$

Following Forbes (2015) [24], the variance $V(p_{sil})$ and covariance $\mathrm{Cov}(p_{si1}, p_{si0})$ are given by

$$\begin{aligned} V(p_{sil}) &= \frac{\pi_{sl}(1 - \pi_{sl})}{n_{sil}} \left(1 + (n_{sil} - 1)\rho_s\right), \\ \text{and } \mathrm{Cov}(p_{si1}, p_{si0}) &= \eta_s \sqrt{\pi_{s1}(1 - \pi_{s1})\pi_{s0}(1 - \pi_{s0})}. \end{aligned}$$

We also make the approximation $\pi_{s0}(1 - \pi_{s0}) \approx \pi_{s1}(1 - \pi_{s1}) \approx \bar{\pi}_s(1 - \bar{\pi}_s)$, where $\bar{\pi}_s = (\pi_{s1} + \pi_{s0})/2$. Furthermore, we assume approximately equal cluster-period sizes within each cluster, i.e., $n_{si1} \approx n_{si0} = n_{si}$. Note that cluster sizes can vary between clusters, $n_{si} \neq n_{si'}$.

Then, Eqn G.3 can be expressed as

$$V(\hat{\beta}_{si}) = \frac{\bar{\pi}_s(1-\bar{\pi}_s)}{n_{si}\pi_{s1}\pi_{s0}}\left[(1+(n_{si}-1)\rho_s)\frac{\pi_{s1}^2+\pi_{s0}^2}{\pi_{s1}\pi_{s0}} - 2n_{si}\eta_s\right]$$

It follows that

$$\begin{aligned}
V(\hat{\beta}_s) &= \frac{\sum_{i=1}^{c_s} V(\hat{\beta}_{si})}{c_s^2} \\
&= \frac{2\frac{\bar{\pi}_s(1-\bar{\pi}_s)}{\pi_{s1}\pi_{s0}}\left(\frac{(1-\rho_s)(\pi_{s1}^2+\pi_{s0}^2)}{2\pi_{s1}\pi_{s0}}\sum_{i=1}^{c_s}\frac{1}{n_{si}} + c_s\left[\frac{\pi_{s1}^2+\pi_{s0}^2}{2\pi_{s1}\pi_{s0}}\rho_s - \eta_s\right]\right)}{c_s^2} \\
&= \frac{2\bar{\pi}_s(1-\bar{\pi}_s)}{\pi_{s1}\pi_{s0}c_s}\left(\frac{(1-\rho_s)(\pi_{s1}^2+\pi_{s0}^2)}{2\bar{n}_{hs}\pi_{s1}\pi_{s0}} + \frac{\pi_{s1}^2+\pi_{s0}^2}{2\pi_{s1}\pi_{s0}}\rho_s - \eta_s\right) \\
&= \frac{\bar{\pi}_s(1-\bar{\pi}_s)}{c_s\pi_{s0}^2}\frac{1+\psi^2}{\psi^2 n_{hs}}\left(1+(n_{hs}-1)\rho_s - \frac{2\psi n_{hs}}{1+\psi^2}\eta_s\right)
\end{aligned}$$

where $\bar{n}_{hs}$ is the harmonic mean cluster-period size in stratum s, defined as $\bar{n}_{hs} = \frac{c_s}{\sum_{i=1}^{c_s}\frac{1}{n_{si}}}$.

By noting that $f_s = c_s/\sum_{s=1}^{S} c_s = c_s/N$, we then obtain

$$V(\hat{\beta}) = \frac{1}{N}\sum_{s=1}^{S} f_s\frac{\bar{\pi}_s(1-\bar{\pi}_s)}{\pi_{s0}^2}\frac{1+\psi^2}{\psi^2 n_{hs}}\left(1+(n_{hs}-1)\rho_s - \frac{2\psi n_{hs}}{1+\psi^2}\eta_s\right) \tag{G.4}$$

where N is the total number of clusters.

G.3 Sample size formulae

The sample size required to detect a difference of $\beta = \log\psi$ between two interventions, with power $1-b$ and significance level α, can be determined by satisfying

$$z_{\alpha/2}\sqrt{V_{h0}(\hat{\beta})} = -z_{1-b}\sqrt{V_{ha}(\hat{\beta})} + \beta$$

where $V_{h0}(\hat{\beta})$ and $V_{ha}(\hat{\beta})$ are the variances of β under the null hypothesis and alternative hypothesis, respectively; and $z_{\alpha/2}$ and z_{1-b} are the standard normal values corresponding to the upper tail probabilities of $\alpha/2$ and b, respectively.

If we further assume that the variance of $\hat{\beta}$ is the same under the null and the alternative

hypotheses, such that $V(\hat{\beta}) = V_{h0}(\hat{\beta}) = V_{ha}(\hat{\beta})$, then we obtain

$$V(\hat{\beta}) = \frac{\beta^2}{(z_{1-b} + z_{\alpha/2})^2}, \tag{G.5}$$

where the variance $V(\hat{\beta})$ is given by Eqn G.4. Equation G.5 is then solved to calculate the required number of clusters N.

G.3.1 Stratified sample size formula for the risk ratio in cluster randomised crossover trials

Using Eqns G.4 and G.5, the required number of clusters, N, to detect a risk ratio with power $1 - b$ and significance level α, is given by

$$N = \frac{(z_{1-b} + z_{\alpha/2})^2}{\log(\psi)^2} \sum_{s=1}^{S} f_s \frac{\bar{\pi}_s(1 - \bar{\pi}_s)}{\pi_{s0}^2} \frac{1 + \psi^2}{\psi^2 n_{hs}} \left(1 + (n_{hs} - 1)\rho_s - \frac{2\psi n_{hs}}{1 + \psi^2}\eta_s\right)$$

G.3.2 Unstratified sample size formula for the risk ratio in cluster randomised crossover trials

An unstratified trial is equivalent to a trial conducted within a single strata $S = 1$. Therefore, the required number of clusters, N, to detect a risk ratio ψ with power $1 - b$ and significance level α is

$$N = \frac{(z_{1-b} + z_{\alpha/2})^2}{\log(\psi)^2} \frac{\bar{\pi}(1 - \bar{\pi})}{\pi_0^2} \frac{1 + \psi^2}{\psi^2 n_h} \left(1 + (n_h - 1)\rho - \frac{2\psi n_h}{1 + \psi^2}\eta\right)$$

where $\bar{n}_h$ is the harmonic mean cluster-period size, defined as $\bar{n}_h = \frac{c}{\sum_{i=1}^{c} \frac{1}{n_i}}$.

G.3.3 Unstratified sample size formula for risk ratio in parallel-group cluster randomised trials

The sample size formula for a parallel-group trial can be determined by assuming $\eta = 0$. The required number of clusters, c, to detect a risk ratio with power $1 - b$ and significance level α is then

$$N = \frac{(z_{1-b} + z_{\alpha/2})^2}{\log(\psi)^2} \frac{\bar{\pi}(1 - \bar{\pi})}{\pi_1} \frac{1 + \psi^2}{\psi^2 n_h} (1 + (n_h - 1)\rho)$$