The Analysis of Clustered Data in Public Health and Healthcare Research

by

Siu Woon Edmond Ng

Thesis submitted for the degree of

Doctor of Philosophy in Health Sciences (by published work)

University of Warwick, Warwick Medical School

October 2021
## Table of Contents

List of tables ................................................................................................................ iii
List of abbreviations ...................................................................................................... iii
Acknowledgements ....................................................................................................... iv
Declaration ..................................................................................................................... iv
Abstract ........................................................................................................................ v

1 Introduction ............................................................................................................... 1

2 Background ............................................................................................................... 3
   2.1 Brief history of clinical trials ......................................................................... 3
   2.2 Cluster randomised trial (CRT): what is it and why use it? ....................... 4
   2.3 Standard statistical methods for group comparisons ................................. 5
   2.4 Key features and challenges in the analysis of CRTs ................................. 7
   2.5 Some appropriate methods for analysing CRT studies ............................... 9

3 The analysis of CRTs and methodological development of appropriate methods for clustered data in public health and healthcare research ......................................................... 18
   3.1 Introduction .............................................................................................................. 18
   3.2 Summary of the publications .............................................................................. 18
   3.3 Methodological development of appropriate methods for analysing CRTs 24

4 Conclusions ............................................................................................................... 29

References .................................................................................................................... 30

Appendix I: List of the published works (in chronological order) ................................ 35
Appendix II: Full list of publications by Siu Woon Edmond Ng (professionally known as Edmond S. W. Ng) .............................................................................................................. 36
Appendix III: Testimony (original in French and a translation in English) regarding Johri, Ng, et al. (2017) ......................................................................................................................... 40
Appendix IV: Co-authors’ statements of candidate’s contribution......................... 42
Appendix V: Copies of the published works (unabridged version only) .............. 50
List of tables

Table 1: Study designs and data sources

List of abbreviations

CEA  Cost Effectiveness Analysis
CRT  Cluster Randomised Trial
CI   Confidence Interval
GEE  General Estimating Equations
GLM  Generalised Linear Model
GP   General Practitioner
MLM  Multilevel Model
MRC  Medical Research Council
ONS  Office for National Statistics
SE   Standard Error
Acknowledgements

I would like to thank my supervisors, Prof. Nigel Stallard and Dr. Nick Parsons, for their advice and guidance that helped bring this thesis to fruition, and Prof. Baron Peter Piot and Prof. James Carpenter for the encouragement and support they provided as I pursued further study.

Declaration

I, Siu Woon Edmond Ng, declare that the publications have not been previously submitted or are currently submitted whether published or in unpublished form, for a degree, diploma, or similar qualification at any university or similar institution.

The work presented was carried out by Siu Woon Edmond Ng.

Siu Woon Edmond Ng

1 October 2021
Abstract

Clustered data arise when data are sampled from clusters of individuals randomized to different interventions in a cluster randomised trial (CRT) or from groups of individuals in a population such as primary sampling units in a multistage clustered survey. Examples include: trials to compare different approaches for managing the wellbeing of elderly people clustered by general practices in the community; surveys on family planning or access to clean water where respondents are geographically clustered; and cost effectiveness studies conducted alongside CRTs.\textsuperscript{1-3}

Although individual randomised trials and simple random sampling are statistically more efficient, CRTs and complex survey designs are pragmatic, and sometimes necessary alternatives in the presence of logistical, financial, or ethical constraints, or a risk of intervention contamination across different arms in a clinical trial. However, the dependence of data from the same cluster violates the critical independence assumption on which most standard statistical methods rely. Such dependence must be accounted for appropriately to enable valid inference.

In this thesis of published work, I draw upon seven publications to highlight the challenges in analysing clustered data, demonstrate the application of a variety of appropriate statistical methods, and show my contribution to developing statistical methodologies for clustered data. They are (i and ii) the analysis of a large CRT, (iii) estimation of intraclass correlation coefficients, [i-iii use data from the same CRT in geriatric research in the UK], (iv) the cost effectiveness analysis (CEA) of a CRT of a non-clinical intervention to reduce caesarean rate, and (v to vii) three pieces of methodological work to improve and extend current methods and models for analysing clustered data.
1 Introduction

Statistical data analysis is integral to any quantitative research including public health, healthcare, and medical research. Statistical methods commonly used in such research rely on an assumption that the data are independent. However, the way in which data are collected for research sometimes violates this important assumption and renders these statistical methods inappropriate for making valid inferences.

Public health and medical researchers are often interested in the effects of different interventions on the health and well-being of people in a population. Randomised controlled trials (RCTs), where interventions are compared by randomly assigning them to individuals, are widely recognised as the gold standard for generating unbiased evidence on the efficacy or effectiveness of the interventions (therapeutic or not). In practice, there are sometimes obstacles that make individual randomisation logistically, financially, or ethically infeasible or unacceptable. A common alternative, with early adoptions since the 1970s, to the RCT overcomes such obstacles by randomising clusters or groups of individuals. This alternative is known as the cluster randomised trial (CRT) design.

Health policy makers and government administrators may be interested in certain characteristics of a population for healthcare provision or planning. Sample surveys are often conducted for such purpose. For example, in a health survey on fertility researchers might like to know the proportion of women using modern contraceptive methods as a means of birth control. The target respondents could be women of reproductive age in the whole country. While simple random sampling is statistically most efficient, it tends to be prohibitive in terms of time and cost in reaching all those selected who may reside in remote corners of the country. Instead, large-scale surveys are typically conducted by multistage cluster sampling (MCS).

In both cases (CRTs and MCS surveys) observations taken from individuals belonging to the same cluster in a CRT or sampled from the same primary sampling unit (PSU) in a survey tend to be more similar than those from other clusters or PSUs. This is primarily due to their exposures to common social, economic, geographical, environmental, and cultural factors found in the same grouping units to which they belong. Observations taken from individuals in the same cluster are typically
dependent. The design and analysis of both study designs are more complex due to the violation of the independence assumption that underpins most conventional statistical methods and hypothesis testing used in public health and medical research.

Data dependence also arises in studies where multiple observations are taken on the same individual (such as in dental research), longitudinal studies when repeated measurements are provided by the same individuals over time, family studies, social or geographical network linked to an index case (in a study of Ebola in West Africa, for example). Such dependencies occur due to persistent characteristics within individuals such as genetic makeup, eating habits, and physical activities among other factors. While some of the statistical approaches that I used or developed in the publications included in this thesis are also applicable to data from longitudinal studies, the primary focus here is on methods for data where correlations exist between observations from different individuals belonging to the same cluster.

In Chapter 2 I briefly describe the history of clinical trials, the key characteristics of CRT, and the statistical issues that beset the analysis of data collected from this study design as well as the pros and cons of several statistical techniques for analysing such data. Chapter 3 provides a summary of the seven publications which sought to address a variety of challenges presented by cluster randomisation in clinical trials as well as clustering by survey design more generally. I conclude in the final chapter with a summary of the contribution by these papers to the analysis of dependent data in public health and healthcare research.
2 Background

2.1 Brief history of clinical trials

The first trial designed to compare the effects of different interventions on the health status of human subjects on record dates back to biblical times. In the “Book of Daniel”, King Nebuchadnezzar believed a diet of only meat and wine offers perpetual health benefit.\textsuperscript{4} When Daniel of Judah, who was taken hostage from the king’s conquered lands, and his friends dissented and insisted instead on a diet of legumes and water, they were allowed their vegetarian diet for 10 days after which they were compared on their ‘countenances’ against those who took the king’s diet. Upon the end of the 10-day trial, the countenances of those on the vegetarian diet “appeared fairer and fatter in flesh” than those on the king’s diet. Almost three centuries later, Grimes (1995) rightly and humorously criticized Daniel’s trial for selection and ascertainment biases as well as confounding by ‘divine intervention’.\textsuperscript{5,6}

Towards the middle of the 18\textsuperscript{th} century the first documented parallel-arm trial was conducted on the British naval ship Salisbury. Sailors on board were afflicted with a common sickness among seafarers – scurvy. By arranging 12 ill sailors with similar symptoms into groups of 2 with each group given a different dietary treatment, the 2 men who were given oranges and lemons got back to their feet after only 6 days while the others remained ‘weak in the knees’.\textsuperscript{4}

However, the 1948 British’s Medical Research Council Trial of Streptomycin for Pulmonary Tuberculosis designed by an advisory committee, whose members included Sir Austin Bradford Hill, is widely recognized as the first RCT and referred to as the “1948 watershed”.\textsuperscript{7-9} Although the trial was neither double-blinded or placebo controlled, it was the first time when the ‘clinical’ and ‘statistical’ reasoning for random allocation came together that led to patients being randomised to different treatment groups, rather than using alternation that was commonly used at the time, as a means to minimize allocation bias.\textsuperscript{7} It has been argued by some that further developments in the RCT design since the 1948 trial are simply refinements to Hill’s proposal.\textsuperscript{4}
2.2 Cluster randomised trial (CRT): what is it and why use it?

The RCT is the recognized gold standard design for generating evidence to inform evidence-based medical and public health practises and decision making. In a RCT individuals from a population of interest are randomised to receive one of two or more interventions (therapeutic or non-therapeutic treatments) being investigated. In reality it is not always feasible or practical to conduct a RCT by randomizing people individually. Ethical considerations, subject acceptability and adherence, costs, intervention contamination, administrative and logistical issues, conflict in war zones and humanitarian crises, and interventions designed to be delivered at group or cluster level are some of the key challenges that sometimes render RCT impractical or unachievable.

When individual randomization is either infeasible or unacceptable, a commonly adopted alternative design is the CRT. For example, it will be infeasible to randomize individuals in a community-based smoking cessation trial where the intervention targets whole communities such as using billboards for the dissemination of public health information. Even if the information is given individually such as randomising households within a community to receive different interventional materials, contamination bias may occur when neighbours receiving different materials confer with each other. The CRT design has gained popularity in recent decades for trials that randomise general practices (GPs), wards, hospitals, schools, worksites, communities as well as other social units to different interventions.

Alternatives to the standard CRT design for comparing multiple interventions (and their interactions) in the same trial, that improve efficiency, or reduce potential biases, have been proposed. These include the factorial design (where effects of multiple interventions and their interactions can be evaluated), stratified design (clusters are randomised separately within each stratum defined by some measure, such as practice size), matched-pairs (clusters are paired and randomised to one of the two interventions), and crossover (where the randomly assigned interventions switch after, typically, a washout period) designs. Other alternatives include stepped-wedge, network-randomised, and “pseudo-cluster” randomised trials.
2.3 Standard statistical methods for group comparisons

Let us consider a simple experimental trial designed to compare the effects of two interventions on a population of interest with $k_t$ clusters in treatment arm $t$ ($t = 1$ for treatment and 0 for control) with cluster $j$ in arm $t$ of size $m_{jt}$ and outcome $Y_{ijt}$ for individual $i$ in arm $t$, $i = 1,...,m_{jt}$, and $j = 1,...,k_t$.

The primary goal of the trial is to compare the means of the outcome of $Y$ in the two arms. In this simple comparison of two group means, statistical methods such as the two-sample t-test as well as linear regression are often used for making inference on the intervention effects.

2.3.1 Standard two-sample t-test

In a standard two-sample t-test, the focus lies in testing the null hypothesis of no difference between the two groups with $H_0: \mu_1 = \mu_0$ against an alternative hypothesis of $H_1: \mu_1 \neq \mu_0$ where $\mu_t$ is the population mean in arm $t$. The population means are estimated by the sample means $\bar{y}_1$ and $\bar{y}_0$ respectively. They are simple averages over all members in the clusters randomised to the respective treatment groups,

$$\frac{1}{M_t} \sum_{j=1}^{k_t} \sum_{i=1}^{m_{jt}} Y_{ijt}$$

where $M_t = \sum_{j=1}^{k_t} \sum_{i=1}^{m_{jt}} 1$ (the number of observations in arm $t$).

The test statistic, $t$, is estimated as follows:

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{s_{\text{pooled}}^2 \left( \frac{1}{M_1} + \frac{1}{M_0} \right)}} \quad [1]$$

where $s_1^2$ and $s_0^2$ are the sample variances with $s_{\text{pooled}}^2 = \frac{(M_1-1)s_1^2 + (M_0-1)s_0^2}{(M_1 + M_0 - 2)}$ and $s_t^2 = \sum_{j=1}^{k_t} \sum_{i=1}^{m_{jt}} \frac{(y_{ijt} - \bar{y}_t)^2}{M_t-1}$. The $p$-value is obtained by referring the resultant t-statistic to the t-distribution with $(M_1 + M_0 - 2)$ degrees of freedom. Where the assumption of equal variances in the two arms of the standard t-test is in question, the Welch’s (or unequal variances) t-test can be adopted instead.

The $p$-value derived from the standard two-sample t-test is generally too low for analysing data from CRTs as it does not account for clustering. This is due to the downwardly biased standard error that failed to account for clustering and the wrong
degrees of freedom used for deriving the $p$-value when there are fewer independent pieces of information than the total number of observations.

The inflation in the magnitude of the test statistic increases with both the intraclass (or intracluster) correlation coefficient (ICC), often denoted by $\rho$, and the average cluster size, $m$.\textsuperscript{13} A formal definition of the ICC is given in 2.4.3. In the presence of clustering the inflation, ignoring the issues of degrees of freedom, is given by the square root of the variance inflation factor, $VIF$ (see page 114 of Donner and Klar),\textsuperscript{13} where

$$VIF = \sum_{t=0}^{1} VIF_{t} \text{ where } VIF_{t} = \sum_{j=1}^{k_{t}} \frac{m_{j}}{M_{t}} [1 + (m_{jt} - 1)\rho]$$ \hspace{1cm} [2]

For constant cluster sizes, $VIF = 1 + (m - 1)\rho$. The $VIF$ is also known as the design effect (DE). Where cluster sizes vary, Eldridge et al. (2006) shows several alternative formulae, including that given by [2], for calculating the DE. They suggest that the effect of adjustment for variable cluster sizes on sample size is negligible when the coefficient of variation (standard deviation of the cluster sizes divided by mean cluster size) is less than 0.23.\textsuperscript{17}

Analysis using a simple two-sample t test also gives no scope for adjusting for any baseline covariates or confounders in case of any baseline imbalance or desire to control for confounders.

2.3.2 Linear models

If data were not clustered, we could assume a simple linear model with an outcome variable $Y_{ijt}$ with variable $X_t$ for treatment assignment as follows:

$$y_{ijt} = \beta_0 + \beta_1 x_{i} + \epsilon_{ijt}$$ \hspace{1cm} [3]

where the residuals $\epsilon_{ijt}$’s are independently and identically distributed (i.i.d.) random variables assumed to follow a Normal distribution with mean 0 and variance $\sigma^2$, with $i, j$, and $t$ as previously defined. The primary interest lies in the regression coefficient, $\beta_1$, representing the treatment effect. Baseline covariates and other confounders can be included in the model for adjusted analysis. Estimation of $\beta_1$ can be performed either by the method of least squares or maximum likelihood estimation. Both methods
yield the same results when the residuals are assumed to follow the Normal distribution.

Where the outcome is assumed to follow a non-Normal distribution, generalised linear models (GLMs) can be used instead for distributions that are part of the exponential family such as Poisson and binomial. The GLMs differ from the normal linear models in that the expectation of the outcome $Y (E[Y] = \mu)$ is transformed by a link function $g()$ in the model $g(\mu) = X\beta$ where $X$ is the design matrix of dimension $(N \times p)$ and $\beta$ the parameter vector of length $p$, and $N$ is the total number of observations. Issues with fitting linear models or GLMs to clustered data will be discussed in 2.4.2.

2.4  Key features and challenges in the analysis of CRTs

2.4.1  Clustered data from CRTs

Data dependence due to clustering arises when values from observations in a cluster are more similar than those in other clusters. In research, such dependence is often a consequence of the study design. CRTs and clustered surveys are examples of such designs.

For example, in a study designed to compare two approaches to assess the health status of elderly people living in the community in the UK, eligible patients from the same participating general practices were randomly assigned to one of the two arms under investigation. Patients registered at the same general practice are generally residents from the same local geographical area in the UK. The size of a catchment area may vary substantially between urban and rural settings. Individuals registered at the same practice are often exposed to similar (and levels of) environmental factors such as air pollution and water quality, public services such as (lack of) access to public facilities and roads. They may also share other similar factors such as age (in the case of English seaside towns with high proportion of retirees), income, among other social-economic factors.
2.4.2 Challenges in analysis of clustered data from CRTs

The primary challenge dependence in the clustered data from CRTs as well as from cluster surveys presents is the violation of the fundamental independence assumption on which many statistical methods rely. This assumption requires that the underlying data are i.i.d. and underpins many statistical methods and hypothesis testing used in public health and medical research for making valid inferences.

Clustering in itself may not be an issue when the unit of randomisation matches the unit of inference (or analysis).\textsuperscript{18} For example, in a trial of GP’s adherence to guidelines for referral where general practices are randomised to receiving or not receiving the guidelines, an appropriate outcome measure could be the proportion of correct (following the clinical guidelines) referrals by each general practice. This is done by pooling the information from all patients in a practice into a single practice-level summary statistic for analysis. The intended level of delivery of an intervention may also affect the choice of the analytical method. For example, an analysis of cluster-level summary measures may be appropriate for an intervention that is designed to be delivered at that level.

However, when inference is intended to be made at the individual level or when research interest focuses on individual level responses to an cluster-level intervention such as in a study that randomised whole general practices to different assessment strategies to ascertain the health status of elderly people,\textsuperscript{1} the study goal was to evaluate the strategies in relation to individual level outcomes in mortality, hospital and institutional admissions.

In the analysis of clustered data from CRTs or clustered surveys, applying conventional statistical methods such as the standard t-test, linear models, or GLMs violates the i.i.d. assumption. This results in downwardly biased SEs leading to overly narrow CIs and type I error (the probability of rejecting the null hypothesis when the null is true) inflation. The effective power to detect a true difference between interventions is also reduced due to the reduction of the effective sample size.\textsuperscript{14,19,20} Statistical methods that are appropriate for modelling clustered data arising from CRTs or by survey designs will be discussed in Section 2.5.
2.4.3 Intraclass correlation coefficient (ICC)

In a study with clustering, the power depends on the ICC, denoted by $\rho$. It can be defined as the correlation between any pair of observations randomly drawn from cluster $j$ with $\rho_j = \frac{\text{cov}(y_{ij}, y_{i'j})}{\sqrt{\text{var}(y_{ij})\text{var}(y_{i'j})}}$ for any individuals $i \neq i'$ while $\text{cov}(y_{ij}, y_{i'j}) = 0$, assuming independence between clusters, for any $j \neq j'$. The subscript $j$ for $\rho$ is often dropped by assuming constant correlation across clusters unless the correlations vary due to some cluster characteristics, e.g., clinical practice variations in a CRT, and one wishes to model such dependence. \(^{21}\)

There are several different ways to estimate the ICC. Among the regression-based methods often used for analysing data from CRTs are marginal models using general estimating equations (GEE)\(^ {22}\) and multilevel (or mixed-effects) models (MLMs) using maximum likelihood.\(^ {23}\) ICCs are estimated somewhat as a by-product for the former method as its focus is on the regression part of the model with clustering treated as a ‘nuisance’ due to the study design.

In the MLMs for CRT the total variance is decomposed into two components attributed to the differences between- and within-clusters. Here the ICC is defined as the proportion of the total variance due to between-cluster variation. Note that the ICCs estimated by MLMs are by construction always non-negative (range from 0 to 1) while there is no such constraint with GEE models. Although the ICCs are generally positive in the context of CRTs. More details of their estimation are given in sections 2.5.5 and 2.5.6.

2.5 Some appropriate methods for analysing CRT studies

2.5.1 Summary measures

One of the simplest approaches to analysing clustered data is to apply standard statistical methods for independent data, such as the standard t-test, to cluster level summary statistics, such as the mean, provided it is reasonable to assume independence between clusters. While this is a valid approach to the analysis of
clustered data, with dependence among observations within clusters effectively removed by the aggregation, there are several shortcomings.

Firstly, when a standard statistical method is applied on the cluster-level summary measures, no consideration is given to the levels of precision with which each of the cluster summaries is estimated if cluster sizes vary. Although a weighted version of the same methods such as a weighted t-test could be performed to take precision into account instead, the aggregation reduces the sample size for hypothesis testing thus results in lower power in general. Furthermore, this approach also does not allow adjustment for baseline imbalance in individual level covariates or confounders, or when treatment effects vary according to age and gender, for example.\textsuperscript{24}

2.5.2 Adjusted two-sample t-test

One of the key features of the ICC is its use for modifying SEs to account for clustering. In a variant of the standard two-sample t-test proposed by Donner and Klar (2000),\textsuperscript{13} the ICC is used to adjust the SE of the t-test that accounts for clustering as follows. The two 1’s inside the bracket in the denominator of Equation [1] are replaced by $VIF_t$ (from [2]) for the corresponding treatment arm $t$.

The test proceeds the same way as the standard two sample t-test except that the test statistic is now based on $K-2$ degrees of freedom where $K$ is the total number of clusters in the whole sample instead. Note that $K-2$ can be substantially smaller than $M_1+M_2-2$ when clustering was ignored (see 2.3.1). When $\hat{\rho} = 0$ suggesting there is no clustering, $VIF_t$ becomes 1 and the adjusted t-test reduces to the standard t-test on $K-2$ degrees of freedom for comparing two independent means.

While the adjusted t-test accounts for clustering, this approach remains limited to the comparison between two groups only. In presence of any baseline imbalance in covariates that are associated with the outcome of interest, skewed outcome, complex variances, and more than two intervention groups for comparison, more flexible alternatives are required.
2.5.3 Robust standard errors

To correct the SE in the presence of clustering, one could simply multiply the SE by the square root of the design effect (DE) calculated using an ICC estimate as discussed earlier.¹³,²⁵ A modified Wald test could be performed for hypothesis testing using the modified SE. An alternative is the robust SE or the square root of the so-called ‘sandwich’ variance estimator as described in more detail below. One of its advantages is that the ICC, which is a critical component of the DE, does not need to be estimated separately before the analysis. Furthermore, the within-cluster dependence does not have to be specified in any way as we shall see below. The application of robust SE in GEE protects the user from misspecification of the within-cluster correlation model.²⁶ GEE models for clustered data will be discussed later.

The sandwich variance estimator has a long history in econometrics and survey sampling, and was independently developed by Huber and White.²⁷,²⁸ According to Rogers (1993) Huber’s original proposal was to produce consistent SEs in the presence of heteroscedasticity, clustered sampling, or when the data are weighted, but not specifically for clustered data.²⁹ The procedure is extended for clustered data by recognizing the dependence of observations within clusters while observations between clusters are independent.²⁹⁻³¹ I will begin by explaining the estimation of the general robust SEs followed by the extension to accommodate clustered data.

The robust (or sandwich) variance estimator is given by

\[ V_R(\hat{\beta}) = \hat{V}(a \sum_{i=1}^{N} u_i'u_i)\hat{V} \]  

where the subscript R denotes robust, a is a finite-sample adjustment factor, \( \hat{V} = -I(\beta) \) where \( I(\beta) = -(\partial^2 \ln L / \partial \beta^2)^{-1} \) (the conventional estimator of variance), and \( u_i \) (a row vector of dimension (1xp)) is the contribution from the \( i^{th} \) observation (with a total of \( N \)) to the scores \( \partial \ln L / \partial \beta \). That is, \( u_i = [\partial \ln L_i / \partial \beta]_{\beta=\hat{\beta}} \). Several formulae for \( a \) exist but all converge to 1 as \( N \to \infty \). Here the \( N \) observations are assumed to be independent. The square root of the diagonal terms in \( V_R(\hat{\beta}) \) are the SEs for the corresponding parameters.
The proposal given by [4] is a very general one which assumes no knowledge or expectation of the underlying structure of the data. In cases where such knowledge exists, such as the within-cluster dependence in CRTs, the estimator in [4] can be extended to accommodate clustered data.\(^{30}\) Assuming the case of a CRT where individual observations within clusters, such as general practices, are not independent. The individual observations are divided into \(K\) independent groups. The \textit{clustered} robust variance estimator is given by

\[
V_c(\hat{\beta}) = \hat{V}(\sum_{t=0}^{1} \sum_{j=1}^{k_t} u_{jt} 'u_{jt})\hat{V}
\]

where the subscript \(C\) denoted ‘clustered’ robust, \(u_{jt} (= \sum_{i=1}^{m_j} u_{ijt})\) is the contribution of the \(j^{th}\) cluster in arm \(t\) to \(\partial lnL / \partial \beta\), and \(m_j\) the size of cluster \(j\) in arm \(t\). Clusters are treated as independent “super-observations” in [5]. The correlated individual-level scores are summed within each cluster before the independent cluster-level totals are multiplied together before the next step in the estimation. Note that the application of robust Ses alone does not affect parameter estimates from, for example, an ordinary regression or GLM.

I demonstrate the use of robust Ses to adjust for clustering in the analysis of a CRT reported in two of my publications in Chapter 3.\(^{1,32}\)

2.5.4 Two-Stage Bootstrap (TSB)

Bootstrap is a common non-parametric re-sampling technique for estimating the uncertainty around a point estimate of a statistic of interest while making minimal assumptions. It serves as an alternative to statistical methods that rely on normality when the assumption is questionable and transformation of the original data is problematic. For example, patient-level cost at an intensive care unit in a CEA is a good example. The primary interest of the health economists is mean cost (whose product with the number of observations yield total cost) while the underlying distribution of this variable is typically highly right-skewed. Transformation of costs by log transformation, for example, is undesirable as the exponential of the averaged log-cost does not provide the arithmetic mean cost. The issue of a highly skewed response provides further challenge to the analysts when coupled with the complex trial design of CRT whose data are increasingly used for CEAs.
In a standard bootstrap $B$ repeated samples of the same size are randomly drawn with replacement from the observed sample. The statistic of interest such as the mean is calculated from each of the $B$ re-samples denoted as $\hat{R}_b^*$ for $b=1…B$. The bootstrap estimates, $\hat{R}_b^*$’s, provides the empirical sampling distribution of the statistic of interest for inference. Davison and Hinkley (1997) proposed an extension to the standard bootstrap that recognizes clustering by resampling clusters and then individuals within the selected clusters. This is referred to as a two-stage bootstrap. While this procedure has been implemented in several health economic studies, researchers from previous studies did not follow Davison and Hinkley’s original suggestion to correct the overestimation of the variance due to resampling at the second stage, unless the number of clusters and individuals within each cluster are both large, by a shrinkage correction. It was suggested that the within-cluster variance is likely to be double-counted at the second stage resampling because the estimation of the cluster means resampled at the first stage would have incorporated both within- and between-cluster variability.

Due partly to the understood software preference in the health economics community, I have implemented Davison and Hinkley’s proposal including the correction for double-counting the within-cluster variance as default in a user-supplied Stata command, tsb. This command is the focus of one of my publications discussed in Chapter 3. In the paper, the use of this command is exemplified with a case study of a CEA using real data from a CRT and the performance of the resultant bootstrap CIs is explored using data from a published simulation study that I co-authored previously.

One of the key limitations of this approach is that it does not allow for any adjustment for covariates or confounders. The next two methods both allow for such adjustments.

2.5.5 Generalized estimating equations (GEE; or marginal-effects models)

The GEE models were proposed by Liang and Zeger as an extension to the GLMs for correlated but, not specifically, clustered data. For clustered data, they work by assuming a ‘working’ correlation matrix, in recognition of the challenge in specifying the ‘correct’ correlation structure, that describes the correlation structure among observations within a cluster. In this approach, clustering is treated as a ‘nuisance’ while the parameters in the regression model are the focus of the modelling exercise.
GEE models are fitted by taking advantage of the less stringent requirement of the quasi-likelihood (QL) methods. Unlike likelihood-based estimation, QL requires only the specification of the relationship between the marginal expectation of the responses and a set of covariates through the regression model and that between the first two moments of the distribution, i.e. mean and variance, without specifying the full parametric distribution.

Let \( y_{jt} \) be the response vector \((y_{1jt}, \ldots, y_{m_{jt}})\)' for cluster \( j \) in arm \( t \) with \( E(y_{jt}) = \mu_{jt} \), \( g(\mu_{jt}) = X_{jt}\beta = \eta_{jt} \) where \( \mu_{jt} \) is the vector of marginal expectations (averaged over the observations with the same covariates), \( g() \) the link function, \( X_{jt} \) the \((m_{jt} \times p)\) design matrix with parameter vector \( \beta \), and the linear predictor \( \eta_{jt} \). The GEE for correlated data are given by:

\[
U(\beta, \alpha) = \sum_{t=0}^{T} \sum_{j=1}^{k_t} D_{jt}' V_{jt}^{-1} (y_{jt} - \mu_{jt}) = 0 \tag{6}
\]

where \( D_{jt} = \partial \mu_{jt}/\partial \beta \). It is sometimes known as the quasi-score equations due to the use of QL. The working variance-covariance matrix for cluster \( j \) is given by:

\[
V_{jt} = \phi A_{jt}^{1/2} R_{jt}(\alpha) A_{jt}^{1/2} \tag{7}
\]

where \( \phi \) is a scale parameter to allow for dispersion, \( A_{jt} \) a \((m_{jt} \times m_{jt})\) diagonal matrix with the variance function \( V(\mu_{jt}) \) as the \( i^{th} \) diagonal element, and \( R_{jt}(\alpha) \) the working correlation matrix with parameter vector \( \alpha \).

Setting initial values for \( R_{jt}(\alpha) \) and \( \phi \) to be the identity matrix and 1 respectively, \( \beta \) is estimated by solving \( [6] \). \( \hat{\beta} \) is then used to estimate the fitted values, \( \hat{\mu}_{jt} \), and, in turn, the residuals, \( y_{jt} - \hat{\mu}_{jt} \). The latter are then used to estimate the parameters \( A_{jt}, \alpha, \phi, \) and \( V_{jt} \) using \( [7] \). These steps are iterated until convergence.

For the analysis of CRTs, the exchangeable correlation structure, that is, \( R_{jt}(\alpha) = 1 \) if \( i = i' \) or \( \rho \) if \( i \neq i' \) for any two individuals, \( i \) and \( i' \), in cluster \( j \) of arm \( t \) is often specified by assuming a common correlation \( \rho \) across all clusters where \( \rho \) is the ICC.

In GEE models, parameters are consistent even if the assumption of the ‘working’ correlation is incorrect provided that there are sufficient clusters (as few as 8 clusters per arm may suffice in general circumstances as demonstrated in \( [37] \)). Hardin and Hilbe (p58) wrote “we gain efficiency in the estimation of the regression parameters by
choosing to formally include a hypothesized structure to the within-panel correlation”. Statistical efficiency can be gained by assuming a working correlation structure that resembles the ‘correct’ correlation structure, especially when individual-level covariates (as opposed to cluster-level ones that do not vary within clusters) are included in the model. Users of GEE models can be further protected from such misspecification with the application of robust SEs (see 2.5.3) since they are robust to any form of within-cluster correlation.

I demonstrate the application of GEE models with robust SEs to adjust for clustering in two of my publications on the analysis of a CRT in Chapter 3.

The GEE models handle correlated data from CRTs with nice statistical properties such as consistent parameter estimates even if the correlation structure is incorrectly specified. However, when cluster-level heterogeneity of the intervention effect is itself of research interest or when individual-level variances and correlations differ across treatment arms, insight into these important characteristics will be lost with GEE models as they are not modelled explicitly. The next approach provides an alternative to address these challenges.

2.5.6 Multilevel models (MLMs; or mixed-effects or cluster-specified models)

In MLMs, the correlations between individuals in the same cluster, and, hence, the within-cluster correlations are assumed to be induced by cluster members sharing a common cluster-level random effect (RE), \( u_{jt} \), for any cluster \( j \) in arm \( t \). This extra level of variation is modelled explicitly in the MLM by assuming a hierarchical data structure with individuals (at level 1) nested within clusters (level 2) while incorporating \( u_{jt} \) into the linear model as defined in section 2.3.2. Due to the presence of both fixed and random effects in the model, MLMs are often referred to as mixed-effects models.

A simple two-level MLM for normally distributed response \( y_{ijt} \) is given by:

\[
y_{ijt} = \beta_0 + \beta_1 x_t + u_{jt} + \epsilon_{ijt} \tag{7}
\]

where the parameters \( \beta_0 \) and \( \beta_1 \) are the intercept and the treatment effect, \( x_t \) the treatment term with \( x_t=1 \) for treatment and 0 for control, \( u_{jt} \) the cluster-level random effect in arm \( t \) with \( u_{jt} \sim N(0, \sigma_u^2) \) and individual-level random effect \( \epsilon_{ijt} \) with \( \epsilon_{ijt} \sim \)
\( N(0, \sigma^2) \). \( u_{jt} \) and \( \varepsilon_{ijt} \) are assumed to be independent. Baseline covariates or confounders can be included in the model. Given this model specification, the observations are known to be conditionally independent on the cluster-specific RE, \( u_{jt} \); that is, \( y_{ijt} \) is i.i.d. conditional on \( u_{jt} \).

Under this model specification the total outcome variance is decomposed into two components. Variation attributable to the differences between clusters is referred to as the between-cluster variance, \( \sigma_u^2 \), while those due to differences within clusters as within-cluster variance, \( \sigma_e^2 \). The ICC is defined as the proportion of total variance due to between-cluster variance, \( \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \). It ranges from 0 when there is no variation in the cluster means to 1 when observations within clusters are identical, i.e., no within-cluster variations.

For normally distributed responses, MLMs can be estimated by maximum likelihood (ML) or residual ML (REML).^23,42 For non-normally distributed response, MLMs had not been widely implemented because fitting a MLM with discrete outcome is mathematically challenging. Parameter estimation involves solving an analytically intractable log-likelihood function that is the product of a discrete (for response) and a continuous (for cluster-level RE) distribution.

The likelihood for the \( m_{jt} \) observations in a cluster is, with subscripts \( j \) and \( t \) omitted for simplicity’s sake, given by:

\[
L(\mathbf{y}; \beta, \sigma_u^2) = \int \prod_{i=1}^{m} h(y_i|u; \beta) f(u; \sigma_u^2) \, du \tag{8}
\]

where the integral is over a normal distribution of the RE. Equation [8] can be generalized for models including multiple REs by replacing \( \sigma_u^2 \) with \( \Sigma \). In [8], \( h() \) is the response distribution conditional on the RE, \( u \), and \( f() \) is the distribution function for \( u \). \( f() \) is replaced by a multivariate version of the distribution when more than one RE exist in the model, e.g., a random intercept and a random slope.^2 When \( h() \) is normal, the integral is tractable and parameters can be estimated by ML. When the response is discrete, such as binary, there may be no closed form solution for [8].

Conjugate distributions such as Poisson distribution for \( ylu \) and gamma for \( u \) in negative binomial models have traditionally been used to make the calculations tractable while numerical integration such as adaptive quadrature methods^43 and
Bayesian Markov chain Monte Carlo methods\textsuperscript{44} have also been proposed. As another alternative I developed an approach known as simulated maximum likelihood for performing Monte-Carlo integration whose results are comparable and faster than those by numerical integration, such as GLLAMM in Stata,\textsuperscript{45} and without the need to specify any prior distribution as in a Bayesian approach.\textsuperscript{2}

In health economics, analysts use MLMs for modelling the bivariate correlations between individual-level costs and effects and the clustering effect simultaneously in CEAs using data from CRTs.\textsuperscript{37} In the next chapter I will show how the MLM can be adapted to accommodate complex variance-covariance structures that may occur in CEAs\textsuperscript{3} and how they were used in a CEA that demonstrates the cost-effectiveness of an audit and feedback intervention targeting health professionals in a hospital-based CRT aiming at reducing caesarean delivery rates in Quebec.\textsuperscript{46}

It is worth noting that for continuous outcomes with an identity link, GEE and MLM methods are mathematically equivalent (see proof on page 97 of Hardin and Hilbe).\textsuperscript{26}
3 The analysis of CRTs and methodological development of appropriate methods for clustered data in public health and healthcare research

3.1 Introduction

The seven publications covered in this thesis include not only examples of the application of appropriate methods in the analysis of CRTs\textsuperscript{1,32,46,47} but also methodological development as well as a statistical programme in Stata for handling different aspects of the complexity attributed to clustering.\textsuperscript{2,3,36}

3.2 Summary of the publications

The seven publications included in this thesis can be found in Appendix I (and V for unabridged version). Appendix II lists all my academic publications. Table 1 shows details of the seven publications, which are briefly described next.
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title of publication</th>
<th>Data source</th>
<th>Study design</th>
</tr>
</thead>
</table>
| Smeeth, Fletcher, Ng, et al. (2002) | Reduced hearing, ownership, and use of hearing aids in elderly people in the UK--the MRC Trial of the Assessment and Management of Older People in the Community: a cross-sectional survey. The Lancet | MRC Trial of the Assessment and Management of Older People in the Community | **General practices randomised** to one of 4 intervention combinations from a 2x2 factorial design, stratified by Jarman score and standardised mortality ratio.  
- Primary comparison of methods of:  
  (i) assessment: universal versus targeted  
  (ii) management: multidisciplinary geriatric assessment team versus primary care team |
| Smeeth, Ng (2002) | Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. Controlled clinical trials | | |
| Ng et al. (2006) | Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Statistical Modelling | • USAsmall simulated data  
• Bangladesh fertility survey | • USAsmall: simulated following same structure as a multistage clustered survey of prenatal health care in a developing country  
• BANG: 1988 Bangladesh fertility survey of contraceptive by a two-stage cluster sample design |
| Ng et al. (2013) | Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data. Stata Journal | Psychological interventions for postnatal depression-randomized controlled trial and economic evaluation (PoNDER) | **General practices randomized** to one of two experimental psychological approaches [person-centred approach (PCA) or cognitive-behavioural approach (CBA)] or usual care (control arm), stratified by number of expected births per year. |
| Ng et al. (2016) | Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice. Statistical Methods in Medical Research | Secondary Prevention of Heart disEase in general practiCE trial (SPHERE) | **General practices randomized** to intervention (practices and patients had access to tailored care plans) or control (patients received usual care). |
| Johri, Ng, et al. (2017) | A cluster-randomized trial to reduce caesarean delivery rates in Quebec: cost-effectiveness analysis. BMC Medicine | The Quality of Care, Obstetrics Risk Management, and Mode of Delivery study (QUARISMA) | **Hospitals** were stratified by level of care and assigned to either the intervention or control group (1:1 allocation ratio) |
3.2.1 The MRC trial of the assessment and management of older people in the community

In 1990 the UK Department of Health introduced a contract of service for GPs requiring them to offer an annual assessment to patients aged 75 or over. Some argued that it was introduced prematurely upon inconclusive evidence of the benefit from underpowered studies of heterogeneous outcomes in disparate healthcare settings, and that some of the earlier studies were flawed due to the use of individual randomization that risked contamination and dilution of the intervention effects. Furthermore, the contract specified only the broad areas of assessment without any guidance on its details. While different methods for the assessment and targeted screening had been proposed, none had been evaluated in an adequately powered randomised trial with a rigorously defined assessment protocol.

The MRC trial was the largest and most comprehensive study designed to evaluate a package of multidimensional assessment and management of elderly people with clearly defined outcomes and assessment protocol in the context of the 1990 contract. It was a CRT with a 2x2 factorial design randomising 106 general practices (33,000 patients) between 4 intervention combinations. The main comparison groups were two assessment (universal versus targeted) and two subsequent clinical management (multidisciplinary geriatric team (GM) versus usual primary care team (PC)) methods. Primary outcomes were mortality, hospital and institutional admissions (nursing or residential home or long stay psychogeriatric hospital), and quality of life.

In this study I took charge of the analysis of one of the main outcomes (institutional admissions), the hearing data for an add-on study, and the estimation and reporting of a large collection of ICCs. I will briefly describe these three papers in turn.

(i) Fletcher, Price, Ng et al. (2004)

This Lancet paper reported results of the analysis of the four main trial outcomes. It was decided at the planning stage of the trial that robust standard errors should be used to account for clustering induced by the cluster-randomised study design. As discussed in section 2.4.2 failure to account for clustering in CRTs often results in...
underestimated standard errors and overestimated test statistics, leading to overly small $p$-values with type I errors inflated above the nominal level.

I implemented the planned statistical analysis by fitting a Poisson regression model for institutional admissions. In this case, robust standard errors were used to account for clustering at practice level and t-tests with $K-L$ degrees of freedom were used to test the significance of regression parameters in the model, where $K$ is total number of PSUs (practices in this case) and $L$ the number of strata (see sections 2.5.2 and 2.5.3).\textsuperscript{56,57}

Overall, the trial did not demonstrate any benefits in the health and quality of life outcomes for people aged 75 or over. The trial led to the withdrawal of a policy for preventative home visits for elderly people in England.\textsuperscript{55}

\textit{(ii) Smeeth, Fletcher, Ng et al. (2002)}\textsuperscript{32}

The aim of this add-on study, reported in the Lancet, was to establish the prevalence of reduced hearing in elderly people in the community and levels of ownership of hearing aids and use in people with poor hearing. I analysed the data on self-reported hearing difficulties, results of the whispered voice test, and the use of hearing aid in the baseline assessment from the trial.

In the presence of clustering induced by the trial design, conventional CIs would have been too narrow as they would have been constructed with underestimated standard errors and the wrong degrees of freedom (see section 2.5.2 for a discussion on degrees of freedom used for an adjusted t-test). I did the analysis using Stata’s survey data suite which uses robust standard errors with adjusted degrees of freedom for any test statistics. The customary degrees of freedom attributed to a test statistic is $d = K-L$ (the number of PSUs minus the number of strata).\textsuperscript{57} Total sample size is usually much larger than the number of PSUs. Hence, the $t$-value, with $d$ degrees of freedom, used for constructing the CIs would generally be larger resulting in a wider and more conservative interval.

In this study high level of prevalence of elderly people with self-reported difficulties in everyday hearing was found with 42% (95% CI 40% to 43%) of participants
reporting at least a little difficulty in hearing. The prevalence of hearing difficulty rose sharply with age.

\textit{(iii) Smeeth and Ng (2002)}^{47}

This paper was motivated by the repeated calls for the publication of ICCs from CRTs to inform the design and planning of future CRTs in the early 2000s.\textsuperscript{11-13,19,25,58-64} This was the first paper that presented the ICCs of a wide range of variables for older people in the community at the level of general practice.\textsuperscript{47} The variables included social variables (such as whether someone has central heating), daily activities and consumptions of alcohol and smoking, as well as self-reported morbidity and functioning.

I used data from the brief assessment from all 106 participating practices and data from the detailed assessment from the 53 practices in the universal arm in which all participants were offered a detailed assessment. Detailed assessment data from the targeted arm were not a representative sample and were not included. I followed the methods by Donner and Klar, which had been used in studies for maternal and child health and adolescent smoking, to obtain ICCs by one-way analysis of variance.\textsuperscript{13,65-67} In this approach the ICC is similarly defined as in section 2.5.6 as the proportion of total variance due to between-cluster variance, \(\sigma_u^2/(\sigma_u^2 + \sigma_e^2)\). I also presented the precision, using the formula by Donner and Koval,\textsuperscript{68} and the variance components (between- and within-cluster variances) of the ICCs as recommended in a framework proposed by Campbell et al.\textsuperscript{69}

A wide range of ICC values were reported from very small (<0.01) to around 0.05. These estimates were comparable to those reported in the literature with similar study design. In community interventional trials where communities were at regional or district level the ICCs were typically <0.01 but their magnitude rose with decreasing cluster size. This is because the within-cluster variation is typically lower with small clusters (such as households) leading to relative larger between-cluster variation. Hence, the ICC tends to be higher for smaller cluster units and vice versa. For example, they were generally <0.05 where communities were defined at postcode level and increased up to 0.3 where community was defined at household level.\textsuperscript{70}
In this study the highest ICC values were observed for some of the social variables, such as whether someone had central heating (0.053) or ‘difficulty making ends meet’ (0.049). Socioeconomic differences in areas served by different practices may explain the large between-practice variation and higher ICCs.

This publication, together with others that report ICCs, has helped improve the quality of CRTs by allowing adequately powered CRTs and clustered surveys that yield more reliable effect estimates of healthcare intervention in similar settings.

3.2.2 The quality of care, obstetrics risk management, and mode of delivery (QUARISMA) trial

The rapid increase in caesarean sections (CSs) worldwide despite its associated short- and long-term risks and increased costs has raised concern.\(^\text{71}\) The QUARISMA trial, a CRT designed to assess whether a non-clinical audit and feedback intervention targeting healthcare professionals would reduce CS rates compared to usual care in Quebec, reported in the New England Journal of Medicine found a statistically significant but clinically small reduction in CSs with no adverse effects.\(^\text{72}\) An economic evaluation was needed to answer the critical question of whether the programme should be scaled up across the province.

(iv) Johri, Ng et al. (2017)\(^\text{46}\)

In July 2015 Prof Mira Johri (Chair of the trial Economic Evaluation Committee) contacted me to enquire about the Stata user-supplied command, \textit{tsb},\(^\text{36}\) I wrote for performing a two-stage bootstrap algorithm that can be used for performing CEA using data from CRT. See Ng et al. (2013) discussed below. I provided Prof Johri with statistical advice for handling the complexity of the cost-effectiveness (CE) data and was subsequently invited to join the trial economic evaluation committee and performed the CEA reported in this paper. In the CEA the intervention was assessed following a difference-in-differences (DID) approach with a focus on the uncertainty in the incremental costs and effects due to the intervention.\(^\text{73}\)
This CEA shares some of the same challenges discussed in sections 2.4 and 2.5.6, such as correlated outcomes (individual-level costs and effects) and dependent data within clusters due to clustering. I modelled the clustered bivariate outcomes jointly by a Bayesian bivariate multilevel linear model (see section 2.5.6 for a discussion of MLM) that accommodates correlations between individual- and hospital-level costs and effects while accounting for the clustering by a hospital-level random effect and is robust to skewed cost data.

Result from the fully adjusted model demonstrated that the intervention was “dominant” in CE terms because the bulk of the joint posterior distribution fell within the quadrant of the cost-effectiveness plane in which the intervention was both effective in reducing CSs and less costly than usual care. In a pre-planned subgroup analysis of low risk women, the posterior probability of reduced costs alone being 99.98% and reduced CSs and costs 99.79%.

The evidence provided by this economic evaluation was instrumental to the funding approval by the Quebec Ministry of Health and Social Services for scaling up the intervention across all 66 hospitals in the province. The findings have also been used to disseminate the QUARISMA programme in other provinces across Canada. See the testimony by the Medical Director of Salus Global's amproOB (Multidisciplinary Approach to Obstetrical Risk Prevention) program, Dr. Guy-Paul Gagné, in Quebec in Appendix III. This was also the only economic evaluation that met the quality assessment criteria for inclusion in the first World Health Organization’s guidelines on interventions to reduce unnecessary caesarean sections published in 2018.

3.3 Methodological development of appropriate methods for analysing CRTs

The next 3 papers demonstrate my contribution to the methodological development in the analysis of clustered data including those from CRTs.

(v) Ng et al. (2006)

Multilevel (or mixed) models with discrete outcome are challenging to fit (see 2.5.6), as there may be no tractable solution to the likelihood function which is the product of
a discrete (outcome) distribution and a continuous (cluster-level random effect) distribution. Quasi-likelihood methods approximate the discrete multilevel likelihood (see, for example, Equation [8] in 2.5.6) by a Gaussian multilevel likelihood using a linearisation (Taylor series expansion) of the multilevel nonlinear (such as logit) model to a multilevel linear model.\textsuperscript{77-79} However, these methods yield downwardly biased (towards zero) parameter estimates, particularly when the data are sparse (few individual-level units per cluster),\textsuperscript{80} and do not produce any estimate of the log-likelihood for comparing nested models. This paper evaluated several simulation-based approaches to bias correction in two-level MLMs with binary outcome. A secondary aim was to use simulated and real data examples to compare and contrast the best simulated-based bias correction method with some alternatives implemented in statistical software.

Among the simulated-based methods, a bootstrap bias correction proposed by Kuk was implemented in MLwiN and the Robbins-Monro stochastic approximation was potentially more efficient.\textsuperscript{81,82} Another contender by Monte-Carlo integration known as Simulated Maximum Likelihood (SML) yielded comparable mean squared errors while providing standard errors of the bias-corrected parameter estimates and an estimate of the log-likelihood at the maximum; thus, allowing hypothesis testing of parameter estimates between nested models. To optimize the search routine for implementing SML in Matlab I supplied it with the analytical form of the first derivative $dI/d\theta$, where $\theta \in (\beta, \Sigma)$ and $\beta$ and $\Sigma$ are the fixed and random parameters respectively, which I manually derived (see Appendix A in Ng et al.\textsuperscript{2}). While all three simulation-based methods were shown to reduce bias and gave similar mean squared errors in the simulation study, the first two did not yield SEs of the biased-corrected parameter estimates.

Using simulated and real data examples, SML was compared with second-order Laplace approximation (implemented in HLM) and numerical integration using quadrature methods (in Stata’s GLLAMM and SAS’s proc NLMIXED).\textsuperscript{83-85} With sparse data the second-order Laplace method yielded markedly lower parameter estimates than SML and quadrature methods. SML required shorter computing times than numerical integration for models with multiple REs.
In all SML performed well against other simulation-based methods for bias reduction and readily provided standard errors of the bias-corrected parameter estimates and an estimate of the log-likelihood at the maximum. Furthermore, it could be extended to handle other link functions, non-normal REs, and higher-level models. This work is referenced in the Stata multilevel mixed-effects reference manual.\textsuperscript{36}

\textit{(vi) Ng et al. (2013)}\textsuperscript{36}

It was reported that up to 95\% of previously published CEAs conducted alongside CRTs failed to account for the complexity of the data from such trials such as clustering, bivariate correlation between costs and effects, and skewness in costs.\textsuperscript{87} In an extensive simulation study that I conducted and reported in a co-authored paper (not included in this thesis), we found a method known as two-stage nonparametric bootstrap (TSB) resampling procedure proposed by Davidson and Hinkley performed favourably even with few clusters, imbalanced cluster sizes as well as skewed costs (Section 2.5.4).\textsuperscript{37}

However, this method was not widely adopted in practice in the community because of its requirement of equal cluster sizes and a lack of readily available software implementation. Although an application of TSB had previously been implemented in R,\textsuperscript{34} the authors did not incorporate a ‘shrinkage’ correction\textsuperscript{35} that was found to be critical in maintaining accurate confidence interval coverage in the simulation study reported in Gomes et al.\textsuperscript{37} Davidson and Hinkley noted that unless the number of clusters and individuals per cluster are both large TSB may overestimate the variance. Resampling at the second stage is likely to double count the within-cluster variance because the cluster means estimated from stage 1 already incorporated both within- and between-cluster variability.\textsuperscript{35,88}

I implemented the TSB algorithm in a new Stata command, \textit{tsb}, described in this Stata Journal paper.\textsuperscript{36} \textit{tsb} was written as a Stata ado-file, that serves as a wrapper function that invokes a number of functions written using the fast Stata programming language Mata to perform the TSB procedure on a user-supplied statistic for clustered data. Although \textit{tsb} was written for CEA in mind, it is applicable to other analyses of clustered data generally. As part of this command suite, I also wrote another command,
tsbceprob, that is more specific for CEA that calculates cost-effective probabilities for the interventions being investigated.

I tested tsb using data from two scenarios in the extensive simulation study reported in Gomes et al. They included a challenging scenario with few (three) clusters per arm, imbalanced cluster sizes, and skewed costs. The results indicated that the CIs constructed by TSB with shrinkage correction gave the most accurate CI coverage.

My intention of making the TSB procedure available through Stata, which is popular among health economists, was to contribute to the improvement in the quality of CEAs alongside CRTs and to help turn research findings into practice. This is evidenced by the 21 citations tracked by Web of Science as of 24 September 2021 suggesting the tsb or tsbceprob command may have been considered or used in 21 studies or more. This work is referenced in the Stata base reference manual.

(vii) Ng et al. (2016)

MLMs provide a flexible modelling framework for CEA of CRTs. They accommodate bivariate correlations between costs and effects at individual and cluster levels, alternative distributions for skewed costs as well as imbalanced cluster sizes (Section 2.5.6). MLMs also allow complex variance-covariance structures that may arise in some CRTs. For example, individual-level costs may have variances, and correlations with effects, that differ between treatment arms or across clusters due to variations in budget constraints, efficiency, clinical practice, or case-mix. However, there was little guidance to model choice among contender MLMs.

In this work I considered a range of Bayesian MLMs (fitted by MCMC in WinBUGS) that assumed increasingly skewed cost distributions in the exponential family (Normal, Gamma, and Inverse Gaussian – variances of the latter two are proportional to increasing power of their means, $\mu^2$ and $\mu^3$, respectively; effects were assumed to be Normally distributed) and with complex variance-covariance structures. I began with a base model with constant variances and correlations across clusters. This was relaxed by allowing individual-level variances to differ first by treatment then by cluster, followed by individual-level correlations to differ in the same ways, and, finally, cluster-level variances and correlations to differ by treatment. The complexity
in the variance-covariance structure was built up cumulatively. Normal distribution was assumed for effects throughout.

An example dataset from the SPHERE study\textsuperscript{91} was used to demonstrate an approach for choosing appropriate MLMs. The aim of the CRT was to assess the effectiveness of a secondary prevention strategy for patients with coronary heart disease. Forty-eight general practices were randomised to intervention (tailored care plans) or control (usual care). The two endpoints were health service costs and health-related quality of life. The primary statistics of interest are the incremental net benefit (INB) estimated by the MLMs.

The approach to model choice was based on four criteria: data characterisation (visual inspection and descriptive statistics), literature informed model pre-specification, diagnostic plots, and model assessment by the Deviance Information Criterion (DIC)\textsuperscript{92}.

Given the same variance-covariance specification, normal plots of deviance residuals differentiated MLMs with different cost distributions while indicating Inverse Gaussian for skewed costs in our example data. However, these plots showed little variations between the MLMs with different variance-covariance structures given the same cost distribution. The DIC differentiated models with different cost distributions as well as different variance-covariance structures. It showed that the best fitting model was the Inverse Gaussian MLM with individual-level variances and correlations that differed by cluster. It is worth noting though that all model yields similar INB estimates with largely overlapping credible intervals that include the null value of zero suggesting no net benefit of the intervention.

This work provided a potential framework to aid model choice of MLMs for CEA of CRTs. However, the results presented were based on one example dataset and may not be generalisable for all CEAs of CRTs.
4 Conclusions

The pragmatic nature of CRT and clustered surveys has made them the designs of choice when individual randomisation or simple random sampling is infeasible or undesirable. However, the data dependency induced by these designs invalidates most standard statistical methods and requires methods that handle such dependency appropriately for valid inference.

Through the seven publications presented in this thesis I have demonstrated the key challenges involved in the analysis of clustered data arising from CRTs or clustered surveys. The papers served as exemplars to those tasked with the analysis of clustered data including CEA of CRTs, e.g., the application of robust standard errors to account for clustering and constructing confidence intervals for CRTs, estimation of ICCs for sample size calculation for future studies, and MLMs for CEA of CRTs.\textsuperscript{1,3,2,4,6,47} The last of these demonstrated the cost-effectiveness of a non-clinical audit programme to reduce unnecessary caesarean sections (CSs) which led to its scale-up across all obstetric hospitals in Quebec. It was the only CEA included in the first WHO’s guidelines on interventions to reduce unnecessary CSs. The papers also contributed to the development of statistical methodology for analysing clustered binary outcomes, the software implementation of a two-stage bootstrap algorithm for clustered data as well as the development of MLMs, with a guidance of model choice, that handle bivariate correlations, skewed costs and complex variance-covariance structures.\textsuperscript{2,3,36}

Looking beyond this thesis, the amount of clustered health data arisen from CRTs is set to rise partly due to the increasing availability of nationally representative electronic health databases\textsuperscript{93} around the world and the advent of pragmatic CRTs that use routinely collected electronic records.\textsuperscript{94-98} While this has the potential to reduce the cost and effort in the management and data collection of CRTs, it is important that researchers and analysts are aware of the issues associated with clustered data and be knowledgeable of some of the appropriate methods as I have illustrated in this thesis. Recent advances in missing data methodology have already expanded the applicability of some of the modelling techniques presented in these papers\textsuperscript{99} and should continue to do so for more challenging missing data mechanisms.\textsuperscript{100}
References


95. Groenwold RHH, Dekkers OM. Designing pragmatic trials—what can we learn from lessons learned? J Clin Epidemiol 2017; 90: 3-5.
Appendix I: List of the published works (in chronological order)

Unabridged version of the seven published works is included Appendix V.


Appendix II: Full list of publications by Siu Woon Edmond Ng (professionally known as Edmond S. W. Ng)


28. Smeeth L, Ng ESW. Intraclass correlation coefficients for cluster randomized trials in primary care - data from the MRC Trial of the Assessment and Management of Older People in the Community. Controlled Clinical Trials 2002; 23(4); 409-21.


1 My surname was erroneously omitted in the authorship list due to an editorial error. An erratum (Int J Geriatr Psychiatry 2002: 17(6): 592) regarding this error was published in a later issue that year.

Book chapter

Conference proceedings
39. Ng ESW, Carpenter J., Goldstein H. Estimation in Generalised Linear Mixed Models; contributed paper presented by Ng ESW in the International Biometric Society 2003 British Region Conference at Reading, UK.
40. Smeeth L, Ng E, Fletcher A. Reduced hearing, and ownership and use of hearing aids among elderly people in the United Kingdom; presented at The Royal National Institute for Deaf people (UK) Research Conference 2001.
41. Leung KS, Ng SWE, Cheung WK, Lee KM. Changes in the BMD in the diaphyses of long bone with age – an attempt to explain the low incidence of diaphyseal fractures in aged population; oral presentation by Ng SWE in the 19th Annual Congress of the HKOA 1999.
42. Cheng JCY, Qin L, Cheung SK, Guo X, Lee KM, Ng E. Low volumetric bone mineral density in adolescent idiopathic scoliosis; presented in Award Paper presentation of the 19th Annual Congress of the Hong Kong Orthopaedics Association (HKOA) 1999.
43. Lau J, Thomas J, Cheung A and Ng E. Cross border sexual networking and its potential impact on low seroprevalence in Hong Kong; presented in the 4th International Congress on AIDS in Asia and the Pacific (ICAAP) at Manila on 26th Oct. 97.

Reports
Appendix III: Testimony (original in French and a translation in English) regarding Johri, Ng, et al. (2017)

Le 2 février 2021

À l’attention de la Professeure Mira Johni
École de santé publique - Université de Montréal
Centre de recherche hospitalier de l’Université de Montréal (CRCHUM)

OBJET : Témoignage concernant la publication « A cluster-randomized trial to reduce caesarean delivery rates in Quebec: cost-effectiveness analysis; John M, Edmond S. W. Ng et al. BMC Medicine (2017) 15:96 »

Chère Pr. Mira Johni,

À titre de directeur médical du programme amproOB (Approche Multidisciplinaire en Prévention des Risques Obstétricaux) de Salus Global au Québec, je désire témoigner de la grande valeur scientifique, ainsi que de l’intérêt de l’article cité en objet.

En 2017 nous avons commencé les démarches auprès du Ministère de la Santé et des services sociaux du Québec (MSSS) afin d’obtenir le financement pour déployer le programme d’audit et rétroaction QUARISMA auprès des 66 hôpitaux obstétricaux de la province. Les représentants du MSSS voulaient une démonstration objective de l’efficacité économique du programme. Le financement a été accordé principalement sur la base des résultats de l’analyse coût-bénéfice présenté dans cet article du BMC.

Il s’agit d’un article de recherche clé qui démontre la rentabilité d’un programme spécifique de gestion sécuritaire du mode d’accouchement. Il éclaire les décisions des gestionnaires et favorise le transfert des connaissances acquises en recherche vers la pratique clinique. Nous utilisons maintenant cette preuve de rentabilité pour diffuser le programme QUARISMA dans les autres provinces canadiennes.

Très cordialement

Guy-Paul Gagné MD
Obstétricien Gynécologue
Directeur amproOB Québec
guy-paul.gagne@salusglobal.com
Le 04 février, 2021

ENGLISH TRANSLATION

RE: Testimony regarding the publication "A cluster-randomized trial to reduce caesarean delivery rates in Quebec: cost-effectiveness analysis; Johri M, Edmond S. W. Ng et al. BMC Medicine (2017) 15:96".

Dear Prof. Mira Johri,

As Medical Director of Salus Global’s amproOB (Multidisciplinary Approach to Obstetrical Risk Prevention) program in Quebec, I wish to testify to the great scientific value and interest of the article cited above.

In 2017 we began the process of approaching the Quebec Ministry of Health and Social Services (MSSS) to obtain funding to deploy the QUARISMA audit and feedback program in the province’s 66 obstetrical hospitals. MSSS representatives wanted an objective demonstration of the program’s economic efficiency. Funding was granted mainly on the basis of the results of the cost-benefit analysis presented in this BMC article.

This is a key research article that demonstrates the cost-effectiveness of a specific program for the safe management of childbirth. It informs management decisions and promotes the transfer of knowledge gained from research to clinical practice. We are now using this evidence of cost-effectiveness to disseminate the QUARISMA program in other Canadian provinces.

With kind regards
Guy-Paul Gagné MD
Obstetrician Gynaecologist
Director amproOB Quebec
guy-paul.gagne@salusglobal.com

https://www.moreob.com/

Translated by Mira Johri, PhD MPH
Professeure titulaire
DG EPS
Appendix IV: Co-authors’ statements of candidate’s contribution

Scanned signed letters follow detailing the contribution of Siu Woon Edmond Ng to the published works.

Statements are signed by the candidate and the first author of each paper or the last living author where the candidate is the first author.
‘Reduced hearing, ownership, and use of hearing aids in elderly people in the UK—the MRC Trial of the Assessment and Management of Older People in the Community: a cross-sectional survey.’ by Smeeth L, Fletcher AE, Ng ES, et al. (2002)¹

This paper reports the hearing component of the results of the MRC Trial of the Assessment and Management of Older People in the Community. I joined the study in 2000 as a research fellow at the host institute, the London School of Hygiene & Tropical Medicine, where the project was based. For the trial as a whole, I first took part in data management by ensuring the accuracy of the data from the scanned questionnaires. Subsequently I took charge of the analysis of the hearing component of the trial results. The key results reported in this paper include the prevalence of hearing loss and the level of use of hearing aid based on one of the largest representative sample of people aged 75 or over in the UK in the early 2000s.

The results confirmed a high level of potential treatable hearing loss among elderly people in the community in the UK as well as low level of use of hearing aids among those who could have benefitted from wearing them. The findings of this paper contributed to the evidence base in geriatric care for informing important health policy to reduce the disabling effect of, often treatable, hearing loss in the UK and other countries in high-income setting.

I co-wrote the statistical section with the first author and presented the results of the in tables throughout the paper. All authors contributed to the discussion and approved the final content of the manuscript.

I, Liam Smeeth (print and sign) hereby certify that the statement above reflects accurately of the candidate's contribution to the manuscript described above.

I, Edmond Ng (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.
'Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community.' by Smeeth L and Ng ES. (2002)\(^1\)

A major problem facing all investigators designing a cluster-based intervention study is the need to estimate the intraclass correlation coefficient (ICC). The ideal solution is to have ICCs available from previous studies that were large enough and had sufficient clusters to generate reasonably accurate estimates of the ICC for the variable of interest. There had been calls for the publication of ICCs to aid the design of future studies. Previous papers presented ICCs for different cluster units and populations.

This was the first paper to present ICCs for a range of outcomes at the general practice level. To aid the sample size calculation for future cluster-based trials of the elderly population in high income setting, I estimated one of the broadest collection of ICCs using data from the MRC Trial of the Assessment and Management of Older People in the Community.

In this paper I provided the formulae for estimating ICC, prepared all of the tables, and performed the data analysis to produce all of the ICCs of a host of variables for elderly people living in the community in a high income setting. I co-wrote the statistical section with the first author. Both authors contributed to the discussion and approved the final content of the manuscript.

I \text{Liam Smeeth} \hspace{2cm} (print and sign) hereby certify that the statement above reflects accurately of the candidate's contribution to the manuscript described above.

I \text{Edmond Ng} \hspace{2cm} (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.

Reference


This paper reports the analysis of the four principal outcomes of mortality, hospital and institutional admissions, and quality of life of the MRC Trial of the Assessment and Management of Older People in the Community. It was a cluster randomised trial designed to compare a universal versus a targeted approach to assessment of the elderly in the community in relation to the three important outcomes.

I joined the study in 2000 as a research fellow at the host institute, the London School of Hygiene & Tropical Medicine, where the project was based. I was first part of a data management team working to ensure the accuracy of the data scanned from pen-and-paper questionnaires as well as getting the data into a suitable format for analysis. For the final analysis of the trial results, I was in charge of the statistical analysis of the principal outcome of institutional admissions.

I estimated unadjusted and adjusted risk ratios using Poisson regression for comparing the two methods of assessment on the intent-to-treat population. The analysis accounted for clustering (by general practice) of the study design by the use of robust standard errors. I also prepared the tables and figures throughout. Although fewer institutional admissions were found in the universal than the targeted arm, the difference was not found to be statistically significant.

The findings of this paper contributed to the evidence base in primary and geriatric care for the assessment and management of the elderly people living in the community in the UK and other high-income country settings.

I co-wrote the statistical section and presented the results in various tables throughout the paper. All authors contributed to the discussion and approved the final content of the manuscript.

I Astrid Fletcher (print and sign) hereby certify that the statement above reflects accurately of the candidate’s contribution to the manuscript described above.

I Edmond Ng (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.

References
‘Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood’ by Ng et al. (2006)

In this work I began with a literature review of the current methods for fitting multilevel or mixed models (MLMs) to discrete outcome data. Noting the deficiencies of some of these methods such as biased estimation and prohibitive running time with the given computing power, I explored the theoretical basis of a number of alternative approaches and successfully implemented them; these included the Robbins-Munro approximation, Kuk’s bias correction and simulated maximum likelihood method (SML).

In particular, I derived a general mathematical expression for the first derivative of the log-likelihood function with respect to model parameters. This was critical to the implementation of the SML approach for efficient estimation in the computer programme Matlab.

I compared the relative performances of the alternative approaches in a detailed literature-informed simulation study as well as applying them to real datasets from a number of published studies. Through these explorations, I was able to demonstrate the dual advantages of the SML approach for bias correction as well as relative improvement in computing time over the quadrature method and Bayesian estimation via MCMC.

I developed the original idea of this paper with Professors James Carpenter and Harvey Goldstein. I wrote the paper under the supervision of the Principal Investigator, Prof James Carpenter, of the ESRC project whose funding supported this work. I performed the simulation study in R and Matlab, the data analyses in various statistical software packages including Stata and HLM, collated all the results and prepared the first draft of manuscript. All authors contributed to the discussion and approved the final content of the manuscript.

I James Carpenter (print and sign) hereby certify that the statement above reflects accurately of the candidate’s contribution to the manuscript described above.

I Edmond Ng (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.

References

‘Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data’ by Ng et al. (2013)\(^1\)

In this work I implemented a stratified two-stage nonparametric bootstrap sampling (TSB) procedure for use with clustered data by creating a new user-supplied command in the statistical package Stata. I described the details of the implementation in this paper. This work was motivated by the lack of options in mainstream software for applying the bootstrap procedure which had been shown in an earlier paper I co-authored to yield estimates with a number of good statistical properties.\(^2\) This procedure was previously shown to be well suited to cost effectiveness analysis conducted alongside cluster randomized trials. Stata is known to be a popular package among health economists.

In this paper I explained the theory behind the bootstrap sampling procedure, gave a detailed description of the algorithm, and demonstrated its use through a number of illustrative examples using real data from published studies. I included Stata syntax for reproducing all of the examples which also serves as a useful tutorial for the command. I performed a simulation study to demonstrate the performance of the method using data simulated from a number of challenging scenarios.

All authors contributed to the conception of the original idea of the paper. All authors contributed to the discussion and approved the final content of the manuscript.

I, James Carpenter, hereby certify that the statement above reflects accurately of the candidate’s contribution to the manuscript described above.

I, Edmond Ng, hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.

References

1. Ng ESW, Grieve R, Carpenter JR. Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data. *Stata Journal* 2013; 13(1): 141-64.
‘Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice’ by Ng et al. (2016)¹

The primary focus of this work was to develop new flexible multilevel models (MLMs) that account for clustering, skewed costs, correlated bivariate outcomes as well as complex variance-covariance structure often observed in cost effectiveness analysis conducted alongside cluster randomise trials. A framework for choosing among candidate MLMs with different levels of variance-covariance structures was also developed in the absence of guidance in the literature.

I began this work by performing a literature search that informed the scope of the research. This includes the alternative distributions that accommodate skewed costs as well as the levels of model complexity. The scope of this research was finalized with agreement of the project steering committee.

I successfully developed a series of flexible MLMs with increasingly level of complexities and demonstrated the use of the framework for model choice. I performed the analyses described in the paper that compare the treatment and the control groups and contrasted the MLMs with different levels of complexities in their variance-covariance structures. I wrote the code to fit the MLMs using the software Winbugs and compared their relative performances using diagnostic plots and the deviance information criterion. I also derived all of the summary statistics and produced all of the graphs in the paper using R.

All authors contributed to the conception of the original idea of the paper. All authors contributed to the discussion and approved the final content of the manuscript.

I James Carpenter  (print and sign) hereby certify that the statement above reflects accurately of the candidate’s contribution to the manuscript described above.

I Edmond Ng  (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.

References

A cluster-randomized trial to reduce caesarean delivery dates in Quebec: cost-effectiveness analysis' by Johri, Ng, et al. (2017)\(^5\)

This paper describes a cost-effectiveness analysis (CEA) conducted alongside the QUARISMA clustered randomized trial (CRT). The primary aim of the main trial was to study the clinical effectiveness of an audit and feedback intervention targeting healthcare professionals to reduce caesarean delivery rates in Quebec, Canada. The challenge of the CEA centres around the clustering and bivariate nature of the individual patient level data from the CRT.

Based on my publications on advancing CEA methods to handle clustered data, I was invited by the Chair, Prof Mira Johri (first author of the paper), of the QUARISMA Economic Evaluation Committee to join the Committee to plan and conduct the CEA using data from the original trial. Prof Johri and I led the planning of the analysis while I developed the bivariate multilevel models needed for this CEA. Thirty-two participating hospitals were assigned to either the intervention or control (usual care) group using blocked randomization resulting in equal numbers of hospital in each arm. The trial adopted a variant of the difference-in-difference approach to estimate the intervention effect. Differences in the bivariate outcomes (delivery mode (caesarean section or virginal) and hospital-based costs) between the pre-intervention and the post-intervention cohorts of pregnant women in the two arms were used for studying the intervention effect. Planned sub-group analyses were performed for women with high-risk or low-risk pregnancies. I had no part in the original clinical trial.

In this CEA, I adapted the bivariate multilevel model framework that I developed previously in Gomes, Ng, et al. (2012).\(^3\) In this modelling framework, correlations between individual-level effects and costs were modelled explicitly while clustering within hospitals was accounted for by cluster-level random effects. The bivariate multilevel models were first fitted by restricted unbiased iterative generalized least squares (RIGLS) for crude and adjusted (for other patient level covariates) intervention effect estimates on each of the outcomes.

The joint distribution of incremental effects and costs was estimated by Bayesian Markov chain Monte Carlo methods. Incremental costs and effects from the joint posterior distribution (stored in the Markov chains) were plotted in cost-effectiveness planes to visually present the joint uncertainty of incremental costs and effects.

I co-wrote the statistical sections with the first author and presented the results of the CEA in tables and figures throughout the paper and in the Appendix. All authors contributed to the discussion and approved the final content of the manuscript.

I ________________ (print and sign) hereby certify that the statement above reflects accurately of the candidate's contribution to the manuscript described above.

I ________________ (print and sign) hereby certify that the statement above reflects accurately of my contribution to the manuscript described above.
Appendix V: Copies of the published works (unabridged version only)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smeeth, Fletcher, Ng et al. (2002)</td>
<td>51-55</td>
</tr>
<tr>
<td>Smeeth and Ng (2002)</td>
<td>56-68</td>
</tr>
<tr>
<td>Fletcher, Price, Ng et al. (2004)</td>
<td>69-79</td>
</tr>
<tr>
<td>Ng et al. (2006)</td>
<td>80-99</td>
</tr>
<tr>
<td>Ng et al. (2013)</td>
<td>100-123</td>
</tr>
<tr>
<td>Ng et al. (2016)</td>
<td>124-140</td>
</tr>
<tr>
<td>Johri, Ng et al. (2017)</td>
<td>141-149</td>
</tr>
</tbody>
</table>
A cluster-randomized trial to reduce caesarean delivery rates in Quebec: cost-effectiveness analysis

Mira Johri1,2,3*, Edmond S. W. Ng4, Clara Bermudez-Tamayo5,6,7, Jeffrey S. Hoch8,9,10, Thierry Ducruet11 and Nils Chaillet12

Abstract

Background: Widespread increases in caesarean section (CS) rates have sparked concerns about risks to mothers and infants and rising healthcare costs. A multicentre, two-arm, cluster-randomized trial in Quebec, Canada assessed whether an audit and feedback intervention targeting health professionals would reduce CS rates for pregnant women compared to usual care, and concluded that it reduced CS rates without adverse effects on maternal or neonatal health. The effect was statistically significant but clinically small. We assessed cost-effectiveness to inform scale-up decisions.

Methods: A prospective economic evaluation was undertaken using individual patient data from the Quality of Care, Obstetrics Risk Management, and Mode of Delivery (QUARISMA) trial (April 2008 to October 2011). Analyses took a healthcare payer perspective. The time horizon captured hospital-based costs and clinical events for mothers and neonates from labour onset to 3 months postpartum. Resource use was identified and measured from patient charts and valued using standardized government sources. We estimated the changes in CS rates and costs for the intervention group (versus controls) between the baseline and post-intervention periods. We examined heterogeneity between clinical subgroups of high-risk versus low-risk pregnancies and estimated the joint uncertainty in cost-effectiveness over 20,000 trial simulations. We decomposed costs to identify drivers of change.

Results: The intervention group experienced per-patient reductions of 0.005 CS (95% confidence interval (CI): −0.015 to 0.004, \(P = 0.09\)) and $180 (95% CI: $−277 to $−83, \(P < 0.001\)). Women with low-risk pregnancies experienced statistically significant reductions in CS rates and costs; changes for the high-risk subgroup were not significant. The intervention was “dominant” (effective in reducing CS and less costly than usual care) in 86.08% of simulations. It reduced costs in 99.99% of simulations. Cost reductions were driven by lower rates of neonatal complications in the intervention group (−$190, 95% CI: −$255 to −$125, \(P < 0.001\)). Given 88,000 annual provincial births, a similar intervention could save $15.8 million (range: $7.3 to $24.4 million) in Quebec annually.

Conclusions: From a healthcare payer perspective, a multifaceted intervention involving audits and feedback resulted in a small reduction in caesarean deliveries and important cost savings. Cost reductions are consistent with improved quality of care in intervention group hospitals.

Trial registration: International Clinical Trials Registry Platform, ISRCTN95086407. Registered on 23 October 2007

Keywords: Randomized controlled trial, Cost-benefit analysis, Caesarean section/utilization, Pregnancy outcomes, Medical audit, Guideline adherence, Multilevel analysis, Female, Adult, Adolescent, Infant, Newborn

*Correspondence: mirajohri@umontreal.ca

© The Author(s). 2017 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

The use of caesarean sections has increased to unprecedented levels worldwide; 18.6% of global and 32.3% of North American births now occur by caesarean section [1]. With the exception of the African region, where caesarean section rates remain low, substantial increases in caesarean deliveries have been documented in all global regions [1]. Caesarean sections are important in reducing maternal and perinatal mortality and morbidity when medically justified; however, non-medically necessary caesareans have no documented benefit [2]. In many jurisdictions, caesarean section rates in excess of recommended thresholds have sparked concerns about potential risks to mothers and infants and escalating healthcare costs [1, 3, 4]. Evidence concerning effective approaches to reduce unnecessary caesareans is currently limited [3, 5].

The Quality of Care, Obstetrics Risk Management, and Mode of Delivery (QUARISMA) study was a cluster-randomized controlled trial conducted from 2008 to 2011 in 32 hospitals in Quebec, Canada [6]. The trial assessed whether a multifaceted audit and feedback intervention targeting health professionals involved in labour, delivery, and postpartum care would reduce caesarean section rates as compared to usual care for pregnant women [6]. The idea motivating the QUARISMA trial was that an effective knowledge translation intervention could improve adherence to clinical practice guidelines and quality of care, thereby reducing unnecessary caesareans and optimizing resource use.

The main trial analysis found a statistically significant but clinically small reduction in caesarean sections (adjusted absolute risk difference of −1.8% (95% CI: −3.8 to −0.2, \(P = 0.04\)) with no adverse effects on maternal or neonatal health outcomes [6]. The reduction in caesareans was observed among women with low-risk pregnancies but not among women with high-risk pregnancies [6]. Although resources required to deliver the intervention are modest, the anticipated magnitude of clinical benefit is small, leaving unanswered questions about whether the intervention should be offered at scale. A cost-effectiveness analysis is required to complete information from the trial and enable decision makers to interpret results for policy and practice.

We conducted a cost-effectiveness analysis of the QUARISMA trial using individual patient data. Our primary objective was to compare the impact on caesarean delivery rates and costs of a multifaceted audit and feedback intervention targeting health professionals versus usual care for pregnant women. Subgroup analyses examined cost-effectiveness in high-risk versus low-risk pregnancies, the clinical subgroups established a priori for the trial.

Methods

The QUARISMA trial was conducted in the province of Quebec, Canada. In 2014, Quebec’s population was 8.2 million, and life expectancy at birth was 79 for males and 83 for females [7]. Gross domestic product per capita measured in US dollar purchasing power parity was $36,216 in 2013 [8]. The QUARISMA economic evaluation adopted the perspective of the publicly funded healthcare system in Quebec. Analysis of effects focused on the primary hypothesized trial outcome of reduction in caesarean sections, while cost analysis examined direct costs to the public healthcare system. Cost variables capture the full spectrum of resource use related to labour and delivery from hospital admission for labour until 3 months postpartum. As delivery mode (caesarean or vaginal) was the primary trial endpoint, a short-term hospital stay is sufficient to capture all meaningful differences in costs and clinical outcomes between intervention and control arms. Costs and health outcomes were discounted at an annual rate of 0% due to the short (<1 year) period for results assessment [9, 10].

All participating hospital centres granted research ethics approval for the trial.

Trial overview

Details of the QUARISMA trial have been reported [6]. Briefly, QUARISMA employed a cluster-randomized design with a 1:1 allocation ratio. Clusters were 32 public hospitals in the province of Quebec, Canada. Hospitals were stratified by level of care and assigned to either the intervention or control group using computer-generated blocked randomization within each stratum. Hospital centres were eligible to participate if they had at least 300 deliveries in the year prior to study initiation, a caesarean section rate of at least 17%, and no concurrent program to reduce caesareans. Based on these criteria, 40 hospitals were eligible, 38 agreed to participate, and 32 were randomly selected for inclusion. All women who gave birth at a participating hospital during the study period and whose newborns met criteria related to gestational age (≥24 weeks) and birth weight (≥500 g) were included in the analysis.

The study comprised three phases: a 1-year pre-intervention (baseline) period, a 1.5-year intervention period, and a 1-year post-intervention period. The baseline period involved onsite training and capacity building to improve caesarean delivery and intrapartum care. During the 1.5-year intervention period, hospital audit committees implemented four 3-month audit cycles using local data to assess the appropriateness of caesarean delivery, engage in collective learning, provide feedback to clinicians, and implement best practices based on the results. Analyses for the main trial and the economic evaluation compared outcomes in the 1-year baseline period (1 April 2008 to 31 March 2009) to those in the 1-year post-intervention period (1 November 2010 to 31 October 2011). The QUARISMA trial captured
more than 65% of all deliveries in Quebec province during the study period [11]. No hospital or woman was lost to follow-up.

Sample size
QUARISMA was designed and powered to detect differences between treatment groups in the primary clinical endpoint (caesarean sections averted). Sample size calculations were not designed to test cost-effectiveness hypotheses.

Effects
We recorded clinical events from patient charts. Trained data collectors extracted data concerning caesarean or vaginal delivery as well as secondary outcomes including major and minor maternal complications and major and minor neonatal complications [6]. Trained research nurses or medical archivists abstracted in-hospital data from the medical records of mothers and newborns 3 months after delivery. Data collectors were aware of randomization assignments but were not involved in outcomes assessment. Given the nature of the clinical condition and the short time horizon, we did not include preference-based measures of health-related quality of life.

Resource use and costs
We considered resource use and costs associated with delivery and complications recorded in the QUARISMA trial (Additional file 1: Section 3; Table S1) [6]. Medical procedure costs (Additional file 1: Table S2 and Table S3) were calculated as the sum of inpatient (hospital) costs and physician fees. For all trial participants contributing data to either the baseline or post-intervention period, we used inpatient chart data to identify clinical events generating resource use. To estimate inpatient procedure costs, we applied unit costs from the 2013 Canadian Institute for Health Information (CIHI) Patient Cost Estimator (PCE) for the jurisdiction of Quebec to resource use categories [12]. Published annually by the Canadian government, PCE costs represent the product of the costs of a standard hospital stay in a specific jurisdiction multiplied by a resource intensity weight reflecting resource use for a specific case mix and age group [12]. The 2013 PCE release includes 2010–2011 financial information [12], which corresponds to the QUARISMA post-intervention period. PCE costs exclude physician fees, which were taken from Canada’s National Physician Database (NPD), 2011–2012 [13]. Total costs also included resource requirements and costs associated with delivery of the QUARISMA intervention (Additional file 1: Table S4), which were estimated directly from QUARISMA trial records. Protocol-driven costs were excluded [9]. All costs are reported in 2013 Canadian dollars. In 2013, one US dollar was worth on average 1.03 Canadian dollars [14].

Statistical analysis
Analyses for the QUARISMA economic evaluation were based on the intention-to-treat principle and carried forward strategies from the main trial analysis [6]. We adopted a variant of the difference-in-differences approach to estimate the intervention effect while controlling for unobserved characteristics of individual patients or program placement in hospitals that might lead to selection bias [15]. Specifically, to assess intervention impact, we studied changes in caesarean delivery rates and costs in the two study groups between the 1-year baseline period and the 1-year post-intervention period using adjusted regression coefficients (with their 95% confidence intervals) for the interaction between group (intervention or control) and time period (post-intervention or baseline) [6, 16].

Main analysis
We modelled costs and effects jointly using bivariate multilevel linear models that explicitly recognize potential correlations between the bivariate outcomes (caesarean sections and costs) at individual and cluster (hospital) levels. In our models, individual patients (level 1) were assumed to be nested within hospitals (level 2), and correlations were estimated at both hierarchical levels. We used cluster-level random effects to handle clustering of observations within hospitals [17, 18]. The analysis included all trial participants. Crude models studied the intervention effect by estimating the interaction between study group and time period. Adjusted models estimated the same interaction term while including additional baseline covariates (pregnancy risk, parity, current smoking, birth weight, and hospital type) used in the QUARISMA trial analysis [6]. Models were estimated by restricted unbiased iterative generalized least-squares estimation implemented in MLwiN (version 2.35) [19] within Stata 14.1 using the runmlwin [20] command.

We had intended to calculate the incremental cost-effectiveness ratio (ICER) statistic as the ratio of additional costs per additional caesarean sections averted. However, interpretation is problematic when ICER estimates are negative, as in this case where the intervention reduces costs [21]. Incremental costs and effects are therefore presented separately.

Uncertainty and heterogeneity of the cost-effectiveness measures
We applied Bayesian Markov chain Monte Carlo methods to the bivariate multilevel linear models used in the main analysis to ascertain the joint uncertainty of the estimands, incremental costs, and effects [22].
posterior distributions of the incremental costs and effects were given by 20,000 iterations of the Markov chains after a burn-in of 5000 iterations. We used cost-effectiveness planes plotting the values of incremental costs and effects stored in the Markov chains to present the joint uncertainty of the estimands. To explore potential heterogeneity in the results, we prospectively planned to repeat all analyses over patient risk subgroups [21].

Other analyses
We adapted the bivariate multilevel linear models from the main analysis to consider two further issues. (1) NPD estimates for physician billings are uncertain [13] and can substantially influence total costs. We prospectively planned sensitivity analyses using alternative physician fee estimates from the Quebec medical specialists billing manual [23]. (2) To gain insight into cost drivers, we decomposed total costs into costs associated with delivery, maternal complications, and neonatal complications. Costs were grouped into these three categories because statistical power was insufficient to examine cost differences on a per-complication basis. Analyses of cost components were empirically motivated.

An independent team not involved in the main trial analysis conducted the economic analysis. All reported \( P \) values are two-sided. \( P \) values lower than 0.05 were considered statistically significant. Additional file 1: Section 2 expands on the statistical analysis; Tables S8 and S9 describe data clustering.

Results
Participants
All (100%) of the 105,351 women who delivered in the baseline or post-intervention period were included in the QUARISMA main trial analysis [6] and the economic evaluation. Analysis of baseline hospital, cost, and patient characteristics (Additional file 1: Table S5) revealed no significant differences between study groups, with the exception of maternal parity, which was included as a covariate in the adjusted analyses.

Incremental costs and effects
Table 1 presents main analysis results concerning the incremental change in effects and costs due to the QUARISMA intervention. Analyses including all patients showed a small non-significant reduction in caesarean sections in the intervention group as compared to controls and an important reduction in costs, yielding adjusted estimates (Table 1, Model 2) of a group as compared to controls and an important reduction in costs. Sensitivity analyses using alternative cost estimates predicted lower estimated savings but similar conclusions (Additional file 1: Table S6).

Uncertainty of incremental costs and effects
Figures 1 and 2 present the joint uncertainties in cost and effectiveness estimates from the Bayesian Markov chain Monte Carlo analyses on the cost-effectiveness plane. Analyses including all patients (Fig. 1) revealed that the intervention reduced caesarean sections and costs in 86.0% of the Monte Carlo iterations, and reduced costs in 99.99% of iterations. For women with low-risk (high-risk) pregnancies (Fig. 2), the intervention reduced caesarean sections in 99.81% (15.97%) of the iterations, reduced costs in 99.98% (90.36%), and reduced both outcomes in 99.79% (15.24%). Sensitivity analyses using alternative cost estimates generated similar results (Additional file 1: Figure S1 and Figure S2).

Cost subcategories
Table 2 presents changes in per-patient costs by subcategories of delivery costs, maternal complications, and neonatal complications. Findings suggest that the intervention to reduce caesarean sections resulted in improved quality along the continuum of care. Adjusted analyses including all patients revealed that the major driver of cost reductions was management of neonatal complications (−$190, 95% CI: −$255 to −$125, \( P < 0.001 \)), which were less frequent in the intervention group [6]. Reductions in management costs for neonatal cardiopulmonary complications were especially important (−$150, 95% CI: −$197 to −$103, \( P < 0.001 \)); costs for other neonatal complications were also reduced (−$41, 95% CI: −$80 to −$1, \( P < 0.047 \)). Results using alternative cost assumptions were similar (Additional file 1: Table S7).

Discussion
The QUARISMA trial found that an intervention involving clinical audits, feedback, and implementation of best practices resulted in a statistically significant but small reduction in caesarean sections without adverse effects on maternal or neonatal outcomes [6]. In this economic evaluation conducted alongside the main trial, the analysis including all trial participants found that the intervention also conferred significant average cost savings of $180 (95% CI: $83–$277) per patient, equivalent to 3.0% (95% CI: 1.4–4.5%) of mean per-patient total costs. Patterns differed by patient risk. Women with low-risk pregnancies experienced statistically significant reductions in caesarean sections and costs, while women with high-risk pregnancies did not experience significant changes in either outcome.
Sensitivity analyses exploring uncertainty in cost-effectiveness estimates over 20,000 trial simulations demonstrated that QUARISMA was the “dominant” intervention (effective in reducing caesarean sections and less costly than usual care) 86.08% of the time. The intervention reduced costs 99.99% of the time. Cost savings largely reflect the reduced costs of managing neonatal complications, which occurred less frequently in the intervention group [6].

Our findings are important for clinicians and policymakers interested in the care of pregnant women and neonates. The QUARISMA trial tested a multifaceted strategy involving audits and feedback to enable groups of health professionals to improve the quality of labour and delivery care at participating hospitals. In addition to achieving a statistically significant but clinically small benefit in the primary outcome of reducing caesarean sections, QUARISMA also demonstrated benefits for a variety of secondary clinical endpoints consistent with improvements in the standard of care implemented in intervention group hospitals [6]. Secondary trial outcomes revealing statistically significant improvements

### Table 1: Impact of the QUARISMA intervention on caesarean sections and total direct medical costs

<table>
<thead>
<tr>
<th>Model 1: BMLM (crude)</th>
<th>Model 2: BMLM (adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef. (β)</td>
<td>Std. Err.</td>
</tr>
<tr>
<td><strong>All participants</strong>&lt;sup&gt;c&lt;/sup&gt; (N = 105,351 patients; 32 hospitals)&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Effects (CS)</td>
<td>-0.009</td>
</tr>
<tr>
<td>Costs ($)</td>
<td>-185</td>
</tr>
<tr>
<td><strong>Low-risk subgroup</strong>&lt;sup&gt;d&lt;/sup&gt; (N = 49,281 patients; 32 hospitals)</td>
<td></td>
</tr>
<tr>
<td>Effects (CS)</td>
<td>-0.013</td>
</tr>
<tr>
<td>Costs ($)</td>
<td>-210</td>
</tr>
<tr>
<td><strong>High-risk subgroup</strong>&lt;sup&gt;d&lt;/sup&gt; (N = 56,070 patients; 32 hospitals)</td>
<td></td>
</tr>
<tr>
<td>Effects (CS)</td>
<td>0.009</td>
</tr>
<tr>
<td>Costs ($)</td>
<td>-102</td>
</tr>
</tbody>
</table>

BMLM bivariate multilevel linear model, CS caesarean section
<sup>a</sup>All costs given in 2013 Canadian dollars = (0.94 USD) [30]
<sup>b</sup>Total costs calculated using CIHI National Physician Database 2011–2012 physician fees for the Province of Quebec [13]
<sup>c</sup>Adjusted models for all participants included the following covariates: parity, smoking, birth weight, hospital type, and women’s risk level
<sup>d</sup>Adjusted subgroup models included the following covariates: parity, smoking, birth weight, and hospital type

---

**Fig. 1** Incremental cost-effectiveness of the QUARISMA intervention versus routine care. Cost-effectiveness (CE) plane for the covariate-adjusted cost-effectiveness analysis of the QUARISMA intervention versus routine care. Incremental cost-effectiveness results were based on 20,000 Markov chain Monte Carlo iterations including all trial participants (N = 105,351). An ellipse containing 95% of the joint posterior distribution of incremental costs and effects is used to represent uncertainty on the CE plane. The centre of the ellipse represents the point estimate of incremental effects and costs, i.e. a per-patient reduction of 0.005 caesarean sections and $180 saved. Percentages represent the distribution of points by quadrant.
included minor neonatal morbidity (i.e. cardiopulmonary morbidity, moderate acidosis (pH ≥7 and <7.1), minor trauma, and neonatal infection/sepsis) and major neonatal morbidity (i.e. major trauma, use of invasive mechanical ventilation, and intrapartum and neonatal deaths) [6]. Economic evaluation offers a framework within which complex changes can be synthesized to aid in policymaking. Our findings (Table 1) suggest that if a similar intervention were to be delivered at scale in the Province of Quebec, for an annual provincial birth cohort of approximately 88,000 [11] and a per-patient reduction in caesareans of 0.005 (95% CI: −0.015 to 0.004), one could anticipate a reduction of roughly 440 (1320 to 0) caesarean sections. For the same annual birth cohort of 88,000 [11] and a per-patient reduction in costs of $180 (95% CI: −$277 to −$83), this translates to a cost savings of $15.8 million (range: $7.3 to $24.4 million), achieved without increasing maternal and neonatal morbidity and mortality.

These results have important implications for the scientific literatures evaluating audit and feedback approaches for quality improvement and non-clinical interventions to reduce caesarean sections. The audit and feedback strategy is widely used to promote quality improvement. A recent Cochrane systematic review found that audit and feedback generally leads to small but potentially important improvements in professional practice [24], which coheres well with the QUARISMA trial finding that the intervention conferred a modest reduction in caesarean sections [6]. Intervention impact on healthcare costs is often an outcome of central interest for quality improvement interventions. Yet, we found only one economic evaluation of an audit and feedback intervention based on data from a randomized trial, published by Fretheim and colleagues in 2006 [25]. This study is a well-conducted evaluation of a small, individually randomized trial in Norway, which found potential cost savings due to the intervention [25]. Analytic methods have advanced considerably in the interim; notably, the Fretheim and colleagues study was not based on individual patient data.

Important studies of non-clinical interventions to reduce unnecessary caesareans exist [5]; however, only one has published an economic evaluation [26]. Hollinghurst and colleagues conducted a cost-consequences analysis of a small, individually randomized trial of 742 pregnant women in the UK with a prior caesarean section [26], to evaluate two computer-based decision aids to reduce repeat caesarean sections [27]. One decision aid reduced repeat caesareans and was likely cost-neutral [27].

Ours is the first economic evaluation of a non-clinical intervention to reduce caesarean section rates based on a cluster-randomized trial. The QUARISMA trial was...
Table 2 Changes in per-patient costs due to the intervention, by clinical category\(^a, b\)

<table>
<thead>
<tr>
<th>Cost component</th>
<th>BMLM (crude)</th>
<th></th>
<th>BMLM (adjusted)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef. ($)</td>
<td>Std. Err. ($)</td>
<td>95% CI ($)</td>
<td></td>
</tr>
<tr>
<td>All participants (N = 105,351 patients; 32 hospitals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery (caesarean or vaginal)(^d)</td>
<td>−8.3</td>
<td>8.2</td>
<td>(−24.4 to 7.7)</td>
<td>0.310</td>
</tr>
<tr>
<td>Maternal complications(^e)</td>
<td>19.9</td>
<td>30.0</td>
<td>(−38.9 to 78.8)</td>
<td>0.507</td>
</tr>
<tr>
<td>Neonatal complications(^f)</td>
<td>−196</td>
<td>33.8</td>
<td>(−262 to −130)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low-risk group (N = 49,281 patients; 32 hospitals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery (caesarean or vaginal)(^d)</td>
<td>−17.2</td>
<td>9.4</td>
<td>(−35.7 to 1.3)</td>
<td>0.069</td>
</tr>
<tr>
<td>Maternal complications(^e)</td>
<td>22.0</td>
<td>38.4</td>
<td>(−53.2 to 97.2)</td>
<td>0.566</td>
</tr>
<tr>
<td>Neonatal complications(^f)</td>
<td>−215</td>
<td>42.0</td>
<td>(−297 to −133)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>High-risk group (N = 56,070 patients; 32 hospitals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery (caesarean or vaginal)(^d)</td>
<td>17.6</td>
<td>12.2</td>
<td>(−6.3 to 41.6)</td>
<td>0.149</td>
</tr>
<tr>
<td>Maternal complications(^e)</td>
<td>17.6</td>
<td>45.3</td>
<td>(−71.2 to 106)</td>
<td>0.698</td>
</tr>
<tr>
<td>Neonatal complications(^f)</td>
<td>−136</td>
<td>51.2</td>
<td>(−237 to −36.0)</td>
<td>0.008</td>
</tr>
</tbody>
</table>

BMLM: bivariate multilevel linear model, CI: confidence interval
\(^a\)Model coefficients (ß), standard errors, and 95% confidence intervals present costs given in 2013 Canadian dollars = (0.94 USD) [30]
\(^b\)Total costs calculated using CIHI National Physician Database 2011–2012 physician fees for the Province of Quebec [13]
\(^c\)The P value is the probability of observing a result at least as extreme as the z-statistic, for the null hypothesis that the difference in the rate changes in costs for the intervention group from baseline to post-intervention versus the control group from baseline to post-intervention does not differ from zero
\(^d\)Delivery costs include: caesarean delivery (primary or secondary, with or without induction, with or without uterine scar, pediatric or adult); vaginal delivery (with or without anaesthetic, with or without intervention, assisted or unassisted, pediatric or adult)
\(^e\)Maternal complication costs include: maternal death, hysterectomy, symptomatic uterine rupture, thromboembolic disease, internal organ injuries, perineal tear (grades 3–4), puerperal infection/sepsis, postpartum hospital stay ≥7 days, admission to intensive care unit, readmission to hospital after postpartum discharge, blood transfusion
\(^f\)Neonatal complication costs include: intrapartum/neonatal death, Apgar score (<4; 4–7); major and minor acidosis (pH <7; pH 7–7.1), major and minor trauma, intraventricular haemorrhage, seizure at less than 24 h, invasive and non-invasive mechanical ventilation, necrotizing enterocolitis, hypoxic-ischaemic encephalopathy, cardiopulmonary morbidity, neonatal infection/sepsis, blood transfusion

Well designed and implemented and captured 65% of provincial births occurring in the trial period, plausibly representing the general population of pregnant women in Quebec province. Prospective data collection improved data completeness, accuracy, and coherence, while the use of standardized government sources for costing enhanced generalizability. The economic analysis was based on the intention-to-treat principle; it used complete data from individual patients and appropriate statistical methods that accounted for clustering and correlated outcomes to derive methodologically robust estimates of between-group differences in costs and effects.

Several limitations should be considered. (1) We assigned unit costs using Canadian government estimates computed annually through established methodologies. Because physician payment mechanisms in Canada are heterogeneous, physician cost estimates are less reliable than those for inpatient costs [13]. We repeated all analyses using an alternative data source representing billable fees per procedure, which provides a lower bound for physician costs and conservatively estimates cost savings.

(2) The economic evaluation used a linear model to study mean costs and effects. This enabled us to address the key policymaking question of interest to decision makers, which concerns the absolute change in effects per change in costs conferred by the intervention. Linear regression models may result in incorrect standard errors. While our bivariate multilevel linear models account for clustering and bivariate correlations at hospital and individual levels, the use of Markov chain Monte Carlo methods to generate joint posterior distributions of incremental costs and effects yielded sample means and estimates of uncertainty robust to skewed costs [17, 28]. The main trial analysis appropriately modelled caesarean sections as binary, and its results, presented on the odds ratio scale, should be considered the definitive estimates of intervention impact [6].

(3) The central limit theorem, upon which valid inference of sample means is based, may not hold with small sample sizes [29]. Our dataset has 105,351 patients from 32 independent clusters with a median cluster size of 2644 patients (range: 638 to 9608). (4) Sample size calculations for the QUARISMA trial were not designed to test cost-effectiveness hypotheses. Methodological guidance suggests that economic evaluation should focus on characterizing differences between study group, rather than formal hypothesis testing [21]. This trial had a large sample size, and overall and stratified cost-effectiveness analyses were able to detect a difference between study groups at a significance level of 0.05, suggesting that the sample size was adequate. (5) Our analysis was based on a health system perspective. Patients and providers may have personal preferences for caesarean sections, positive or negative, not
captured by this analysis. (6) The main trial analysis identified a possible random increase in complications in the control group. Some proportion of the cost savings conferred by the intervention might be attributable to random variation; this cannot be ascertained from the trial data. (7) Hospitals eligible for inclusion in this trial had a minimum annual delivery volume of 300; intervention performance in smaller hospitals is unknown, and this may affect the scope for achieving benefits at scale. Returns from the QUARISMA trial itself may range from $0.5 million to $5 million in the 1-year post-intervention period alone (Additional file 1: Table S10). (8) This study was done in a single jurisdiction; differences in health system organization and financing may limit transferability of results to other settings. A national budget impact analysis is being conducted to translate these results for all Canadian provinces. (9) The relatively short trial time horizon does not elucidate how the intervention impact may evolve over time.

A similar program could be beneficial in other regions with similar or higher caesarean section rates [6]. Cost-effectiveness results can be adapted to other jurisdictions by adjusting for local baseline caesarean rates, health system costs, and proportions of high- and low-risk patients.

Conclusions
From a healthcare payer perspective, a multifaceted intervention involving audits and feedback resulted in a small reduction in caesarean deliveries and an important reduction in per-patient costs. These results were achieved without increasing neonatal and maternal morbidity and mortality.

In keeping with the theory of change for the QUARISMA intervention, which attempted to optimize medical practice by reducing unnecessary caesareans, our study found that women with low-risk pregnancies experienced statistically significant reductions in caesarean sections and costs, while changes for the high-risk subgroup were not significant. We also found evidence of improved quality along the continuum of care. Cost reductions were driven principally by lower rates of neonatal complications and corresponding lower use of resources within the intervention group. These changes are consistent with improvements in the quality of care in intervention group hospitals.

Findings from our study provide critical new evidence concerning a safe and possibly sustainable strategy to reduce unnecessary caesarean sections and shed new light on the potential for audit and feedback interventions to improve quality of care while controlling costs. Delivery of the intervention at wider scale accompanied by further research is required to assess whether the improvements in clinical practice driving reductions in caesarean sections and cost savings are sustainable over time.

Additional file

**Additional file 1: Supplementary material for a cluster-randomized trial to reduce caesarean delivery rates in Quebec: cost-effectiveness analysis.** (PDF 716 kb)

**Abbreviations**
CIHI: Canadian Institute for Health Information; CS: Caesarean section; ICER: Incremental cost-effectiveness ratio; NPD: National Physician Database; PCE: Patient Cost Estimator; QUARISMA: Quality of Care, Obstetrics Risk Management, and Mode of Delivery (trial)

**Acknowledgements**
The authors gratefully acknowledge Drs. François Audibert, Anne-Monique Nuyt, Jean-Charles Pasquier, and Annie Li for expert guidance on care pathways and costs for maternal and neonatal complications, and the editor and reviewers for thoughtful guidance that has improved the scientific quality of the paper.

**Funding**
The Canadian Institutes of Health Research (grants 200702MCT-171307-RFA-CFCF-153236 and MOP 81275) funded this study. The funding sources played no role in the study design, data collection, analysis and interpretation, writing of the report, or the decision to submit the paper for publication.

**Availability of data and materials**
The data that support the findings of this study are available on reasonable request from the senior author (NC), subject to approval of the local institutional review boards. The data are not publicly available because they contain information that could compromise privacy and consent for participating centres and individuals.

**Authors’ contributions**
MJ, JSH, and NC conceived of the idea for the study and developed the study design. TD and NC provided the data. Statistical analysis was lead by ESWN in collaboration with MJ and CBT; all authors contributed to interpretation of results. MJ drafted the manuscript, and ESWN, CBT, JSH, TD, and NC revised it for important intellectual content. All authors read and approved the final manuscript. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

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

**Consent for publication**
Not applicable.

**Ethics approval and consent to participate**
The economic evaluation plan was pre-specified in the trial protocol and updated to reflect recent advances in analytic methods and reporting. On behalf of the 32 participating hospital centres, the following 32 local institutional review boards granted research ethics approval for the QUARISMA trial: (1) Comité d’éthique de la recherche du Centre intégré de santé et de services sociaux du Bas-Saint-Laurent; (2) Comité d’éthique de la recherche du Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean (Hôpital-Dieu de Roberval); (3) Directeur des services professionnels du Hôtel-Dieu d’Alma; (4) Comité d’éthique de la recherche, Centre de santé et de services sociaux de Chicoutimi; (5) Comité d’éthique de la recherche du Centre hospitalier universitaire de Québec - Université Laval; (6) Comité d’éthique mixte du Centre Hospitalier du Centre-de-la-Mauricie; (7) Comité d’éthique de la recherche multi-clientèle du Centre intégré universitaire de santé et de services sociaux de la Mauricie-et-du-Centre-du-Québec; (8) Comité d’éthique de la recherche du Centre Hospitalier Universitaire de Sherbrooke; (9) Comité d’éthique de la recherche Centre Hospitalier de St. Mary; (10) Directeur des services professionnels, Hôpital général du Lakeshore; (11) Directeur des services professionnels, Hôpital St-Luc (Centre hospitalier de l’Université de Montréal); (12) Comité d’éthique de la recherche du Centre hospitalier universitaire Sainte-Justine; (13) Comité d’éthique de la recherche du Centre intégré universitaire de santé et de services sociaux de l’Est-de-l’Île-de-Montréal (Hôpital Maisonneuve-Rosemont); (14)
Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1 Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Tour Saint-Antoine, Porte 503-910, 850, rue St-Denis, Montréal, Québec H2X 0A9, Canada. 2 Department of Health Management, Evaluation and Policy, School of Public Health, University of Montreal, Montréal, Québec, Canada. 3 Department of Maternal, Neonatal, Child and Adolescent Health, World Health Organization, Geneva, Switzerland. 4 Director's Office, London School of Hygiene and Tropical Medicine (LSHTM), London, UK. 5 Department of Obstetrics and Gynecology, Centre Hospitalier Universitaire (CHU) de Sherbrooke, Sherbrooke, Québec, Canada. 6 Andalusian School of Public Health, Granada, Spain. 7 CIBER Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain. 8 Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada. 9 Centre for Excellence in Economic Analysis and Research (CLEAR), Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada. 10 Department of Public Health Sciences, University of California, Davis, California, USA. 11 Department of Biostatistics, Centre hospitalier universitaire Sainte-Justine, Université de Montréal, Montréal, Québec, Canada. 12 Department Obstétrique et Gynécologie, Centre Hospitalier de l'Université Laval (CHUL), Québec, Québec, Canada.

Received: 11 December 2016 Accepted: 20 April 2017
Published online: 22 May 2017

References
9. Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technologies. Ottawa: Canadian Agency for Drugs and Technologies in Health (CADTH); 2006.