COVARIATE ADJUSTMENT IN CLINICAL TRIALS

STATISTICAL AND METHODOLOGICAL ISSUES ON COVARIATE ADJUSTMENT IN CLINICAL TRIALS

By

RONG (RACHEL) CHU, B.Sc. (Honours), M.Sc. (Statistics)

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

McMaster University DOCTOR OF PHILOSOPHY (2012) Hamilton, Ontario

(Health Research Methodology – Biostatistics Specialization)

TITLE: Statistical and Methodological Issues on Covariate Adjustment in Clinical Trials

AUTHOR: Rong (Rachel) Chu B.Sc. (McMaster University),
M.Sc. (University of British Columbia)

SUPERVISOR: Professor Lehana Thabane

NUMBER OF PAGES: xi, 158

ABSTRACT

Background and objectives

We investigate three issues related to the adjustment for baseline covariates in late phase clinical trials: (1) the analysis of correlated outcomes in multicentre RCTs, (2) the assessment of the probability and implication of prognostic imbalance in RCTs, and (3) the adjustment for baseline confounding in cohort studies.

Methods

Project 1: We investigated the properties of six statistical methods for analyzing continuous outcomes in multicentre randomized controlled trials (RCTs) where within-centre clustering was possible. We simulated studies over various intraclass correlation (ICC) values with several centre combinations.

Project 2: We simulated data from RCTs evaluating a binary outcome by varying risk of the outcome, effect of the treatment, power and prevalence of a binary prognostic factor (PF), and sample size. We compared logistic regression models with and without adjustment for the PF, in terms of bias, standard error, coverage of confidence interval, and statistical power. A tool to assess sample size requirement to control for chance imbalance was proposed.

Project 3: We conducted a prospective cohort study to evaluate the effect of tuberculosis (TB) at the initiation of antiretroviral therapy (ART) on all cause mortality using Cox

proportional hazard model on propensity score (PS) matched patients to control for potential confounding. We assessed the robustness of results using sensitivity analyses.

Results and conclusions

Project 1: All six methods produce unbiased estimates of treatment effect in multicentre trials. Adjusting for centre as a random intercept leads to the most efficient treatment effect estimation, and hence should be used in the presence of clustering.

Project 2: The probability of prognostic imbalance in small trials can be substantial. Covariate adjustment improves estimation accuracy and statistical power, and hence should be performed when strong PFs are observed.

Project 3: After controlling for the important confounding variables, HIV patients who had TB at the initiation of ART have a moderate increase in the risk of overall mortality.

PREFACE

This dissertation is a “sandwich thesis”, which is composed of three individual projects prepared for publication in peer-reviewed journals. The following are the contributions of R. Chu in all of the papers included in the dissertation: developing the research ideas and research questions; designing the studies; developing the analysis and simulation plans; conducting all statistical analyses and simulations; produced all figures and tables; interpreting the results; writing all of the manuscripts; submitting the manuscripts; and responding to reviewers’ comments. My co-authors contributed to the acquisition of the example datasets, provision of clinical expertise, and critical revision of the manuscripts. The work of this thesis was conducted between September 2008 and July 2012.

The first and second papers have been published and the remaining one will be submitted to a peer-reviewed journal in the near future.

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my thesis supervisor, Dr. Lehana Thabane. He has been providing outstanding mentorship, guidance, support and encouragement throughout my PhD study at McMaster University. I am thankful for the tremendous advice and opportunities he provided in the last four years. I hope he will be a life-long mentor and friend.

I am grateful to my thesis committee, including Dr. Lehana Thabane, Dr. Eleanor Pullenayegum, and Dr. PJ Devereaux, for their continuing guidance and collaboration during my training at McMaster. A special thank goes to Dr. Stephen Walter, who was extremely supportive and provided invaluable and timely contributions to the paper on prognostic imbalance in the thesis.

I would like to thank Dr. Gordon Guyatt, Dr. Anne Holbrook, Dr. Ed Mills, and Dr. Joseph Beyene for their collaborations and inspiring discussions, and for providing me with access to the datasets used in the thesis. I appreciate the friendship and support from my colleague students and help from the administrative staff at McMaster University and St. Joseph’s Healthcare Hamilton. Without you, life would not have been the same.

Finally, I am grateful for the continuing support from my parents, Renzhen Chu and Fenghua Hu, throughout this unforgettable journey. The thesis would not have been possible without their encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

Page

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

Page

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 1

INTRODUCTION

Clinical trials are widely used to compare the effects of medical interventions in humans (1-5). Although it can be difficult to propose a universal framework to categorize clinical trials evaluating various types of interventions in different disease areas, in the development of therapeutic drugs, clinical trials are conventionally grouped into four phases (1,2,6). Phase I studies look for the best dose of a drug from a pharmacological perspective. Phase II studies aim to generate preliminary data on the safety and efficacy of fixed doses of a drug and assess study feasibility. Phase III studies are randomized controlled trials (RCTs) that provide confirmative evidence on the efficacy and severe side effects of an experimental intervention relative to control therapy. Such trials, if conducted rigorously, can produce comparability of the known and unknown baseline characteristics between intervention groups, and provide valid estimates of intervention effects (1,2,7). Phase IV studies are usually long-term surveillance studies (interventional or observational in nature) to help understand the effectiveness and safety of the new intervention in real-world situations.

Statistics, as an essential tool to make inference in the designing and analysis of clinical studies, has advanced considerably over the past decades (7-10). Many methodological challenges in statistical modeling of late phase trials have been identified and studied in the literature (2,11,12). However, different philosophies of statistical reasoning sometimes lead to the use of alternative approaches to address a clinical question. The diversified perspectives are likely to promote the development and prosperity of the field of statistics. On the other hand, a lack of standardization of statistical procedures can yield different estimates of an intervention effect for a given research question. Disagreement exists on many practical issues in planning and analyzing clinical trials (2,13). Comprehensive investigation on the methodological and practical challenges demands efforts of generations of statisticians.

The objective of this thesis is to address some of the challenges around baseline covariate adjustment in phases III and IV clinical trials, through the lens of Monte Carlo simulations and sensitivity analyses, and to provide directions for future research. Three specific statistical issues are investigated: (1) the analysis of correlated outcomes in multicentre RCTs, (2) the assessment of the probability and implication of chance imbalance of a baseline prognostic factor in simple RCTs, and (3) the adjustment for baseline confounding in large prospective cohort studies.

Issue 1: Intracentre correlation in multicentre RCTs

Multicentre RCTs can improve patient accrual rate and increase the applicability of a study (14). In such experiments, investigators often use a stratified randomization design to achieve balance over centre level differences in the study population or the management team, and to improve the precision and efficiency of the statistical analysis (10). Clustering within a centre emerges when the outcomes observed from patients managed by the same centre (practice or physician) are more similar than the outcomes from different centres. The extent of clustering is often quantified by intraclass (or intracentre) correlation (ICC), a number ranging between 0 and 1 (15). A large ICC value indicates that individual observations within a centre contain less unique information.

In practice, different strategies of handling centre effects are carried out to analyze data from multicentre RCTs (16-20,20-22). While some completely ignore any possible centre variation, others account for centre difference as fixed or random effects using a regression or meta-analytic framework. To date, only a few studies have been conducted to compare the performance of statistical methods that are commonly used to analyze continuous outcomes from multicentre RCTs, using the Monte Carlo simulation technique (16,17,23). No consensus has been reached on what are the best methods to analyze such data. The association of model performance (in terms of accuracy, precision and efficiency) and trial characteristics, including the number of participating centres, the number of patients per centre, the variation of centre size, and the value of ICC, has not been thoroughly studied in the literature.

Issue 2: Prognostic imbalance in RCTs

Randomization is the most important feature of RCTs, because it minimizes selection bias, and on average balances the known and unknown baseline prognostic factors (PFs) between treatment groups (1). Despite randomization, imbalance in PFs as a result of chance may still arise in an individual trial, and with small to moderate sample sizes such imbalance may be substantial (24,25). A biased estimate of the treatment effect may result from ignoring a large chance imbalance in key PFs between treatment groups (26,27). When the PFs are well known and measured at baseline, the consequences of prognostic imbalance can be controlled by stratified statistical analyses, sometimes in conjunction with design techniques such as stratified randomization (28,29). The balancing of unknown PFs between treatment groups entirely relies on the randomization and play of chance, for which the sample size plays a critical role. Sample size calculations often assume a balance of prognosis between the treatment groups regardless of sample size, yet the distribution of the possible unobserved PFs is impossible to examine based on the observed data.

In the process of grading the quality of medical evidence resulted from RCTs, an important issue is the lack of understanding of the likelihood of chance imbalance of the known and unknown PFs, and its implication on the estimation of treatment effects. The sample size required to minimize the probability and impact of chance imbalance in RCTs is lacking in clinical trial literature. The knowledge gap encumbers the assessment of the quality of evidence and the strength of recommendations in healthcare research.

Issue 3: Baseline confounding in the assessment of the effect of tuberculosis on mortality among HIV patients

The Human Immunodeficiency Virus (HIV) is one of the world's leading infectious diseases. The total number of people living with HIV reached 34 million in 2010, among whom 60% live in Sub-Saharan Africa (30). Tuberculosis (TB) is a leading cause of mortality among HIV infected individuals and accounts for one half of the AIDS deaths worldwide (31). TB at the initiation of antiretroviral therapy (ART) is expected to importantly affect the likelihood of survival among HIV-co-infected patients. Yet, the magnitude of increased mortality is poorly understood, particularly in populations with lower TB prevalence settings, such as Uganda, Africa (32-34).

Observational studies constitute an important tool in HIV/AIDS research. They may be the only feasible method to address a clinical or epidemiological question similar to the one stated above, for ethical or practical reasons. In the absence of randomization, attributing causality is a major challenge. The conventional statistical approaches of adjustment may not provide adequate control against a large number of potential confounders when the sample size or the number of outcome events is small. The balancing distribution of confounding variables between the intervention or risk groups after conventional stratification or covariate adjustment is difficult to examine. A powerful alternative that has been increasingly used to control for baseline confounding is the propensity score (PS) methods (35-37). This approach can estimate the average exposure effects on the whole population or subpopulations using the observed datasets. In the PS methods, the vector of potential confounding variables reduces to a single score that reflects one's propensity of being exposed to an intervention or a risk factor. Conditional on PS, the exposure is independent of confounders being included in the PS model, and the actual exposure effects can be estimated (35,36). The PS methods are more advantageous when numerous confounders need to be accounted for in studying the association between a rare outcome and a common exposure variable. The relative performance of different methods of controlling for PS in the outcome model is likely to vary between studies. Missing data is another major problem in HIV/AIDS research. Missing data may occur because of data entry errors, missing visits, loss to follow-up, or additional reasons which may or may not related to the exposure or the outcome. Ignoring incomplete information is likely to lead to invalid or unreliable results when the missing is not completely at random (38).

Summary of Chapters

This is a sandwich thesis of three papers, each matched to one of the issues described above. The papers are separated into three chapters beginning with Chapter 2.

Chapter 2 deals with intracentre correlation (ICC) among continuous outcomes in multicentre RCTs. We compared six statistical methods for analyzing correlated continuous outcomes in multicentre RCTs using Monte Carlo simulations. The methods under investigation include simple linear regression, fixed-effects regression, random-effects regression, generalized estimating equation (GEE), and fixed- and random-effects centre-level analysis. We considered a wide spectrum of ICC values, and varying numbers of centres and centre size in balanced and unbalanced designs in the simulation study, assuming the absence of treatment by centre interaction. Model performance was evaluated using bias, precision, mean squared error of the point estimator of the treatment effect, empirical coverage of the 95% confidence interval, and statistical power of the procedure.

Chapter 3 discusses the assessment and implication of prognostic imbalance in RCTs evaluating a binary outcome. We confined our attention to one binary baseline prognostic factor (PF), and varied five trial design parameters in this simulation study, including the frequency of the outcome event in the control group; the effect of treatment on the outcome; the strength of the association between the PF and the outcome; the prevalence of the PF; and the sample size. First, we evaluated the probabilities of various levels of imbalance in the binary PF between two treatment groups. Second, we investigated the impact of prognostic imbalance on the estimation of treatment effect, by comparing statistical performances of the two logistic regression models with and without adjustment for the PF. Finally we examined the effect of sample size on the probability and impact of prognostic imbalance. Our simulation study was intended to provide information on what constitutes an adequate sample size to control against potential impact of prognostic imbalance in simple RCTs.

In Chapter 4, we aim to estimate the effect of tuberculosis (TB) at the initiation of antiretroviral therapy (ART) on all cause mortality in HIV co-infected patients who received ART, using a large prospective HIV cohort in Uganda, Africa. We applied propensity score (PS) matching method to account for potential baseline confounding when assessing the impact of TB on patient survival using a Cox proportional hazard model. We inspected the comparability of the potential confounders using numerical and graphical tools. We used multiple imputation to handle missing covariate information at baseline. In addition, we examined the sensitivity of study results by comparing estimates from different PS methods (matching on PS, stratifying on PS, adjusting for PS as regression covariate) with the conventional multivariable Cox regression model.

Chapter 5 summarizes the key findings of Chapters 2 to 4, and discusses the implications and limitations of the thesis. The common goal of all three papers is to advance our understanding on the analytical strategies involving baseline covariates in clinical research. Results of the individual projects will also shed light on the design of efficient and rigorous clinical trials.

References

(1) Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 4th ed. New York, NY: Springer; 2010.

(2) Piantadosi S. Clinical Trials: A Methodologic Perspective. New York, NY: John Wiley & Sons, Inc.; 1997.

(3) Yusuf S, Cairns J, Camm J, Fallen EL, Gersh BJ. Evidence-Based Cardiology. 3rd ed.: John Wiley & Sons, Inc.; 2011.

(4) Girling DJ, Parmer MKB, Stenning SP, Stephens RJ, Stewart LA. Clinical Trials in Cancer: principles and practice. New York, NY: Oxford University Press; 2003.

(5) Finkelstein DM, Schoenfeld DA editors. AIDS Clinical Trials. New York, NY ed.: John Wiley & Sons, Inc.; 1995.

(6) International Harmonised Tripartite Guideline: General Considerations for Clinical Trials:E8. 1997; Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E8/Step4/E8_Guideline.pdf. Accessed 06/01, 2012.

(7) Cook TD, DeMets DL. Introduction to Statistical Methods for Clinical Trials. Boca Raton FL: Chapman & Hall/CRC; 2007.

(8) Armitage P, Colton T editors. Encyclopedia of Biostatistics. 2nd ed. New York, NY: John Wiley & Sons, Inc.; 2005.

(9) Collett D. Modelling Survival Data in Medical Research. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2003.

(10) Lachin JM. Biostatistical methods: the assessment of relative risks. 1st ed. New York: John Wiley and Sons; 2000.

(11) Senn S. Consensus and Controversy in Pharmaceutical Statistics. Journal of the Royal Statistical Society: Series D (The Statistician) 2001;49:135-176.

(12) Pocock SJ. Current controversies in data monitoring for clinical trials. Clin Trials 2006;3(6):513-521.

(13) Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, et al. Randomized clinical trials. Perspectives on some recent ideas. N Engl J Med 1976 Jul 8;295(2):74-80.

(14) International Conference on Harmonisation E9 Expert Working Group. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Stat Med 1999 Aug 15;18(15):1905-1942.

(15) Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. BMJ 1998 May 9;316(7142):1455.

(16) Moerbeek M, van Breukelen GJ, Berger MP. A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. J Clin Epidemiol 2003 Apr;56(4):341-350.

(17) Pickering RM, Weatherall M. The analysis of continuous outcomes in multi-centre trials with small centre sizes. Stat Med 2007 Dec 30;26(30):5445-5456.

(18) Worthington H. Methods for pooling results from multi-center studies. J Dent Res 2004;83 Spec No C:C119-21.

(19) Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.

(20) Whitehead A. Meta-analysis of controlled clinical trials. 1st ed. Chichester: John Wiley and Sons; 2002.

(21) Gould AL. Multi-centre trial analysis revisited. Stat Med 1998 Aug 15-30;17(15-16):1779-97; discussion 1799-800.

(22) Fleiss JL. Analysis of data from multiclinic trials. Control Clin Trials 1986 Dec;7(4):267-275.

(23) Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. Stat Med 1998 Aug 15-30;17(15-16):1767-77; discussion 1799-800.

(24) Senn S. Testing for baseline balance in clinical trials. Stat Med 1994 Sep 15;13(17):1715-1726.

(25) Wang SJ, O'Neill RT, Hung HJ. Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. Clin Trials 2010 Oct;7(5):525-536.

(26) Brower RG, Lanken PN, MacIntyre N, Matthay MA, Morris A, Ancukiewicz M, et al. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. N Engl J Med 2004 Jul 22;351(4):327-336.

(27) Meade MO, Cook DJ, Guyatt GH, Slutsky AS, Arabi YM, Cooper DJ, et al. Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive

end-expiratory pressure for acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. JAMA 2008 Feb 13;299(6):637-645.

(28) Yu LM, Chan AW, Hopewell S, Deeks JJ, Altman DG. Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. Trials 2010 May 18;11:59.

(29) Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002 Oct 15;21(19):2917-2930.

(30) World Health Organization. HIV/AIDS Fact sheet N°360. 2011; Available at: http://www.who.int/mediacentre/factsheets/fs360/en/index.html. Accessed 06/01, 2012.

(31) Centers for Disease Control and Prevention. TB and HIV/AIDS. CDC HIV/AIDS facts. 2008; Available at: http://www.cdc.gov/hiv/resources/factsheets/PDF/hivtb.pdf. Accessed November/15, 2009.

(32) Harries AD, Zachariah R, Corbett EL, Lawn SD, Santos-Filho ET, Chimzizi R, et al. The HIV-associated tuberculosis epidemic--when will we act? Lancet 2010 May 29;375(9729):1906-1919.

(33) Gandhi NR, Nunn P, Dheda K, Schaaf HS, Zignol M, van Soolingen D, et al. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 2010 May 22;375(9728):1830-1843.

(34) Lonnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet 2010 May 22;375(9728):1814-1829.

(35) Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984;79:516-524.

(36) Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55.

(37) Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007 Jan 15;26(1):20-36.

(38) Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2002.

CHAPTER 2

Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: A simulation study

Rong Chu[12], Lehana Thabane[124], Jinhui Ma[12], Anne Holbrook[134], Eleanor Pullenayegum[124], Philip James Devereaux[1]

Affiliations:

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Health Sciences Centre, Room 2C7, 1200 Main Street West, Hamilton ON, Canada, L8N 3Z5

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

[3]Division of Clinical Pharmacology, Department of Medicine, McMaster University, Hamilton ON, Canada

[4]Centre for Evaluation of Medicine, St Joseph's Healthcare Hamilton, Hamilton ON, Canada

Corresponding author:

Rong Chu
Email: chur@mcmaster.ca
Telephone: (905) 522 1155 ext. 34918
Fax: (905) 308 7212

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

ABSTRACT

Background

Multicentre randomized controlled trials (RCTs) routinely use randomization and analysis stratified by centre to control for differences between centres and to improve precision. No consensus has been reached on how to best analyze correlated continuous outcomes in such settings. Our objective was to investigate the properties of commonly used statistical models at various levels of clustering in the context of multicentre RCTs.

Methods

Assuming no treatment by centre interaction, we compared six methods (ignoring centre effects, including centres as fixed effects, including centres as random effects, generalized estimating equation (GEE), and fixed- and random-effects centre-level analysis) to analyze continuous outcomes in multicentre RCTs using simulations over a wide spectrum of intraclass correlation (ICC) values, and varying numbers of centres and centre size. The performance of models was evaluated in terms of bias, precision, mean squared error of the point estimator of treatment effect, empirical coverage of the 95% confidence interval, and statistical power of the procedure.

Results

While all methods yielded unbiased estimates of treatment effect, ignoring centres led to inflation of standard error and loss of statistical power when within centre correlation was present. Mixed-effects model was most efficient and attained nominal coverage of 95% and 90% power in almost all scenarios. Fixed-effects model was less precise when the number of centres was large and treatment allocation was subject to chance imbalance

within centre. GEE approach underestimated standard error of the treatment effect when the number of centres was small. The two centre-level models led to more variable point estimates and relatively low interval coverage or statistical power depending on whether or not heterogeneity of treatment contrasts was considered in the analysis.

Conclusions

All six models produced unbiased estimates of treatment effect in the context of multicentre trials. Adjusting for centre as a random intercept led to the most efficient treatment effect estimation across all simulations under the normality assumption, when there is no treatment by centre interaction.

Background

A multicentre randomized control trial (RCT) is an experimental study "conducted according to a single protocol but at more than one site and, therefore, carried out by more than one investigator"[1]. Multicentre RCTs are usually carried out for two main reasons. First, they provide a feasible way to accrue sufficient participants to achieve reasonable statistical power to detect the effect of an experimental treatment compared with some control treatment. Second, by enrolling participants of more diverse demographics from a broader spectrum of geographical locations and various clinical settings, multicentre RCTs increase generalizability of the experimental treatment for future use [1].

Randomization is the most important feature of RCTs, for on average it balances known and unknown baseline prognostic factors between treatment groups, in addition to minimizing selection bias. Nevertheless, randomization does not guarantee complete balance of participant characteristics especially when the sample size is moderate or small. Stratification is a useful technique to guard against potential bias introduced by imbalance in key prognostic factors. In multicentre RCTs, investigators often use a stratified randomization design to achieve balance over key differences in study population (e.g. environmental, socio-economic or demographical factors) and management team (e.g. patient administration and management) at centre level to improve precision of statistical analysis [2]. Regulatory agencies recommend that stratification variables in design should usually be accounted for in analysis, unless the potential value of adjustment is questionable (e.g. very few subjects per centre) [1].

The current study was motivated by the COMPETE II trial which was designed to determine if an integrated computerized decision support system shared by primary care providers and patients could improve management of diabetes [3]. A total number of 511 patients were recruited from 46 family physician practices. Individual patients were randomized to one of the two intervention groups stratified by physician practice using permuted blocks of size 6.The number of patients treated by one physician varied from 1 to 26 (interquartiles= 7.25, 11, 15; mean=11; standard deviation [SD]=6). The primary outcome was a continuous variable representing the change of a 10-point process composite score based on eight diabetes-related component variables from baseline to a mean of 5.9 months' follow-up. A positive change indicated a favourable result. During the study, the possibility of clustering within physician practice and its consequence on statistical analysis was a concern to the investigators. The phenomenon of clustering emerges when outcomes observed from patients managed by the same centre, practice or physician are more similar than outcomes from different centres, practices or physicians. Clustering often arises in situations where patients are selective about which centre they belong to, patients in a centre or practice are managed according to the same clinical care paths, or patients influence each other in the same cluster [4]. Intraclass (or intracentre) correlation (ICC) is often used to quantify the average correlation between any two outcomes within the same cluster [5]. It is a number between zero and one. A large value indicates that within-cluster observations are similar relative to observations from other clusters and each observation within cluster contains less unique information. This implies that the independence assumption which many standard statistical models are based on is violated. An ICC of zero indicates that individual observations within the same clusters are uncorrelated and different clusters on average have similar observations.

Through a literature review, we identified six statistical methods that were sometimes employed to analyze continuous outcomes in multicentre RCTs: A. simple linear regression (two sample t-test), B. fixed-effects regression, C. mixed-effects regression, D. generalized estimating equations (GEE), E-1. fixed-effects centre-level analysis, and E-2. random-effects centre-level analysis. The first four methods use patient as unit of analysis, yet address centre effects differently [6-8]. Simple linear regression completely ignores centre effects that are likely to arise from two sources: (1) possible differences in environmental, socio-economic or treatment factors between centres, and (2) potential correlation among patients within centres. Although stratified randomization attempts to minimize the impact of centre on standard error of the treatment effect by ensuring that the treated and control groups are largely balanced with respect to centre, failure to control for stratification in analysis will likely inflate variance of the effect estimate. The fixed-effects model treats each participating centre as a fixed intercept to control for possible population or environmental differences among centres. This model assumes that study subjects from the same centre have independent outcomes, i.e. the intraclass correlation is fixed at zero. The mixed-effects model incorporates dependence of outcomes within a centre and treats centres as random intercepts. Proposed by Liang and Zeger [9], the generalized estimating equation (GEE) model extends generalized linear regression with continuous, categorical or count outcomes to correlated observations within cluster. Under a commonly used and perhaps oversimplified assumption, that the degree of similarity between any two outcomes from a centre is equal, an exchangeable correlation structure can be used to assess treatment effect in Model C and D. Though the within- and between-centre variances (σ_e^2 and σ_b^2) are estimated differently in these two models. Method E-1 and E-2 are routinely employed to combine information from different studies in meta-analysis [10]. One can also apply them to aggregate treatment effects over multiple centres [11-13]. The overall effect is obtained as the average within-centre effect differences over centre, using inverse-variance weighting.

To date, only a few studies have been carried out to compare the performance of statistical models in analyzing multicentre RCTs using Monte Carlo simulation [6, 7, 14], whereas many studies assessed the impact of ICC in cluster randomization trials. Moerbeek et al [6] compared the simple linear regression model, fixed-effects regression and fixed-effects centre-level analysis with equal centre size. Pickering et al [7] examined the bias, precision and power of three methods: simple regression, fixed-effects and mixed-effects regression assuming block randomization of size 2 or 4 on a continuous outcome. In the presence of imbalance and non-orthogonality, they found ignoring centres or incorporating them as random-effects led to greater precision and smaller type II error compared with treating centres as fixed effects. Performance of the GEE approach and centre-level methods were not investigated in that work. Jones et al [14] compared the fixed-effects and random-effects regression models to a two-step Frequentist procedure as well as a Bayesian model, in the presence of treatment by centre interaction, and recommended fixed-effects weighted method for future analysis of multicentre trials. The investigation was further expanded to assessing correlated survival outcomes from large multicentre cancer trials. A series of random-effects approaches were proposed to account for centre or treatment by centre heterogeneity in proportional hazards models [15, 16].

A lack of definitive evidence on which models perform the best in various situations led to this comprehensive simulation study to examine the performance of all six commonly used models with continuous outcomes. The objective was to assess their comparative performance in terms of bias, precision (simulation standard deviation (SD) and average estimated SE), and mean squared error (MSE) of the point estimator of the treatment effect, empirical coverage of the 95% confidence interval (CI) and the empirical statistical power, over a wide spectrum of ICC value and centre size. We did not consider treatment by centre interaction this study, partly because clinicians and trialists have been making efforts to standardize the conduct and management of multicentre trials via, for instance, uniform patient selection criteria, staff training, and trial monitoring and auditing to reduce heterogeneity of treatment effects among centres. Furthermore it is uncommon to find clinical trials designed with sufficient power to detect treatment by covariate interactions.

In this paper, we survey six methods to investigate the effect of a treatment in multicentre RCTs in detail. We outline the design and analysis of an extensive simulation study, and report how model performance varies with ICC, centre size and the number of centres. We also present the estimated effect of the computer-aid decision support system on management of diabetes using different methods.

Methods

Approaches assessing treatment effects

We investigated six statistical approaches to evaluating effect of an experimental treatment on a continuous outcome compared with the control, for multicentre RCTs. Assuming baseline prognostic characteristics are approximately balanced between the treatment and control groups via randomization, we do not consider covariates other than centre effects in the models. The first four approaches use individual patient as unit of analysis, while centre is the unit of analysis in the last two approaches.

Simple linear regression (Model A)

This approach models the impact of treatment (X) on outcome (Y) via regression technique (Equation 1). In the context of a two-arm trial, this approach is the same as a two-sample t-test [6].

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + e_{ij}, \qquad \text{(Equation 1)}$$

where Y_{ij} is the outcome of the i-th patient in the j-th centre, X_{ij} stands for the treatment assignment (X_{ij} = 1 for the treatment, X_{ij} = 0 for the control), and e_{ij} is the random error assumed to follow a normal distribution with mean 0 and variance σ_e^2. The intercept, β_0, represents the mean outcome for the control group in all participating centres, and the slope β_1 represents effect of the treatment on the mean outcome.

Fixed-effects regression (Model B)

This model (Equation 2) allows a separate intercept for each centre (β_{0j}) as a fixed effect by restricting the scope of statistical inference to the sample of participating centres in a RCT. Interpretation for β_1 remains the same as in Model A. Model A and B were fitted using the linear model procedure 'lm()' in **R**.

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + e_{ij} \qquad \text{(Equation 2)}$$

Mixed-effects regression (Model C)

Similar to Model B, the mixed-effects regression model assumes that the intercept $\beta_{0j} = \beta_0 + b_{0j}$ follows a normal distribution N(β_0 , σ_b^2), and is thus random effect. In Equation 3, b_{0j} is the random deviation from the mean intercept β_0 , specific for each centre.

$$Y_{ij} = \beta_0 + b_{0j} + \beta_1 X_{ij} + e_{ij} \qquad \text{(Equation 3)}$$

Similar to the previous models, the within-centre variability is reflected by σ_e^2. The variability of outcome between-centre is captured by σ_b^2 in Model C. The variance and covariance of outcomes in the same or different centres can be expressed as: Var(Y_{ij}) =

$\sigma_b^2 + \sigma_e^2$, Cov($Y_{ij}, Y_{i'j}$) = σ_b^2, Cov($Y_{ij}, Y_{i'j'}$) = 0. The intraclass correlation that measures the correlation among outcomes within centre is given by $\frac{\sigma_b^2}{\sigma_e^2 + \sigma_b^2}$, assumed equal across all centres. We fitted Model C in **R** via linear mixed-effects procedure 'lme()' using the restricted maximum likelihood (REML) method [17,18].

Generalized estimating equations (Model D)

The GEE method has gained increasing popularity among health science researchers for its availability in most statistical software. As opposed to the mixed-effects method that estimates treatment difference between arms and individual centre effects, the GEE approach models the marginal population-average treatment effects in two steps: 1) it fits a naïve linear regression assuming independence between observations within and across centres, and 2) it estimates parameters of the working correlation matrix using residuals in the naïve model and refit regression model to adjust standard error and confidence interval for within-centre dependence [19]. As a result, the estimated impact of treatment on the outcome in GEE model reflects the "combined" within- and between-centre relationship. GEE employs quasi-likelihood to estimate regression coefficients iteratively, and a working correlation needs to be supplied to approximate the within centre correlation. When the working correlation is mis-specified, the sandwich-based covariance estimator will lead to a robust yet less efficient estimate of treatment effect in GEE model [9]. Recently, statisticians found that variance of the estimated treatment effect could be underestimated when the number of centres was small [20]. We therefore assessed the efficiency of GEE models using procedure 'gee()' in library(gee) in **R**. As in the mixed-effects model, an exchangeable correlation structure was assumed in fitting Model D.

Centre-level fixed-effects model (Model E – 1)

The centre level model is a stratified analysis performed on the mean difference in outcome between the treatment and control arms within centre. The overall treatment effect is estimated by a weighted average of individual mean differences across all centres. The principle of inverse-variance weighting is often used (Figure 1). This model is essentially a centre-level inverse-variance weighted paired t-test (i.e. the treatment arm is paired to the control arm in the same centre) to account for within centre correlation [10]. In the absence of intraclass correlation and under the assumption of equal sampling variation at patient level, the inverse-variance weight reduces to $\frac{n_{tj}n_{cj}}{n_{tj}+n_{cj}}$ for the j-th centre, which can be further simplified as the size of centre $n_j = n_{tj} + n_{cj}$, given equal numbers of patients in two arms. Here n_{tj} and n_{cj} represent the number of patients in the treatment and control group, respectively, in the j-th centre. This form of the weighted analysis (without adjustment for covariates) was discussed extensively by many researchers [21-23]. We implemented Models E – 1 using the fixed-effects method for meta-analysis provided by the ‘metacont()’ procedure in **R.**

Centre-level random-effects model (Model E – 2)

A random-effects approach is used to aggregate mean effect differences not only across all participating centres but also across a population of centres represented by the sample. This model factors heterogeneity of treatment effect among centres (i.e. random treatment by centre interaction) into its weighting scheme and captures within- and between-centre variation of the outcome. One should not confuse this method with the mixed-effects model using patient-level data (Model C). For Model E-2, the underlying true treatment effects are not a fixed single value for all centres, rather they are considered random effects, normally distributed around a mean treatment effect with between-centre variation. Model C, on the other hand, treats centres as random intercepts and postulates the same treatment effect across all centres. Model E-2 does not serve as a fair comparator to the alternatives listed here which assume no treatment by centre interaction. Preliminary investigation suggested E-2 could outperform E-1 in some situations; we therefore included E-2 in the study to advance understanding of these models. Details of centre level models are provided in Figure 1. Model E – 2 was carried out using DerSimonian-Laird random-effects [24] method using the ‘metacont()’ procedure in **R**. The confidence interval for Model E – 2 was constructed based on the within- and between-centre variances.

Study data simulation

We used Monte Carlo simulation to assess performance of statistical models to analyze parallel group multicentre RCTs with a continuous outcome. We simulated outcome, *Y*, using the mixed-effects linear regression model (Model C): $Y_{ij} = \beta_0 + b_{0j} + \beta_1 X_{ij} + e_{ij}$ for the i-th patient in the j-th centre, where X_{ij} (=0, 1) is the dummy variable for treatment allocation (i = 1… m_j, j = 1… J). We generated random error, *e*, from N(0, σ_e^2

=1). We set the true treatment effect (β_1) to be 0.5 residual standard deviation (σ_e), an effect size suggested by the COMPETE II trial. This corresponds to a medium effect size according to Cohen's criterion [25]. To simulate centre effects, we employed the relationship between ICC and σ_b^2: ICC $= \dfrac{\sigma_b^2}{\sigma_e^2 + \sigma_b^2}$. To fully study the behaviour of candidate models at various ICC levels, we considered the following values of ICC for completeness: 0.00, 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50 and 0.75. This in turn set the corresponding σ_b^2 values to be 0, 1/99, 1/19, 1/9, 3/17, 1/4, 1/3, 3/7, 7/13, 2/3, 9/11, 1 and 3. However, we focused interpretation of the results on lower values of ICC as they were more likely to occur in practice [26-28].

The original sample size was determined to be 84 per arm using a two-sided two-sample t-test (Model A) to ensure 90% power to detect a standardized effect size of 0.5 at 5% type I error rate. We increased the final sample size to 90[1] per arm to accommodate more combinations of the number and size of participating centres. We assumed patients were randomly allocated to two groups with a ratio of 1:1, the most common and efficient choice. We generated data in nine scenarios (Table 1) to assess model performance in three designs: (a) balanced studies where equal numbers of patients are enrolled from study centres and the numbers of patients in the two arms are the same (fixed by design); (b) unbalanced studies where equal numbers of patients are enrolled from study centres but the numbers of patients in two arms within centre may be different due to chance yet

[1] Power increases to 91.8%

remain 1:1 allocation ratio; and (c) unbalanced studies where the numbers of patients enrolled vary among centres, and block randomization of size 2 or 4 is used to reduce chance imbalance. For designs (a) and (b), we considered three combinations of centre size and number of centres: J = 45 centres, 4 patients per centre; J = 18 centres, 10 patients per centre; and J = 6 centres, 30 patients per centre. Design (c) mimicked a more realistic scenario for multicentre RCTs. For the first setup of design (c), we grouped 180 patients to 17 centres. It was constructed so that the centre composition and degree of allocation imbalance were analogous to the COMPETE II trial but at a smaller sample size: the number of patients per centre varying from 1 to 28; quartiles = 5, 10, 15; mean=11; SD=8; percentage of unbalanced centres between 47% and 70% depending on block size.

To compare results from various models in analyzing the COMPETE II trial, and assess accuracy and precision of the effect estimates, we included an additional scenario in design (c) to imitate this motivating example more closely, with respect to sample size and centre composition (scenario 9). We generated treatment allocation (X1) and outcome (Y) for 511 patients in 46 centres, where the number of patients per centre was set exactly the same as observed in the COMPETE II trial (Table 2). In particular, three centres recruiting only one patient was simulated. Analogously to COMPETE II, a fixed block size of 6 was used to assign patients to treatments. The same simulation model was employed as in previous scenarios yet a separate set of parameters based on results of the COMPETE II trial were used (Table 3): β_0= 1.34, β_1 = 1.26, σ_b^2 = 1, σ_e^2=7, ICC = 0.125.

We generated 1000 simulations for each of the 13 ICC values under each of the first eight scenarios and 1000 simulations for the specified ICC value for the ninth scenario. Separate sets of centre effects were simulated for each scenario and each simulation 1-1000. We chose to simulate 1000 replicates so that the simulation standard deviation for the empirical power at a nominal level of 90% in the absence of clustering was controlled at 1%. This also ensured that standard deviations of the coverage of the confidence interval and the empirical power not exceed 1.6%.

Comparison of analytic models

We applied six statistical models to each simulated dataset. For each model, we calculated the bias, simulation standard deviation (SD), average of estimated standard error (SE) and mean squared error (MSE) of the point estimator of treatment effect (i.e. β_1), empirical coverage of the 95% confidence interval around β_1 and the empirical statistical power. We constructed confidence intervals based on t-test for Models A – C, and Wald interval based on normal approximation for Models D and E. We estimated bias as the difference between the average estimate of β_1 over 1000 simulated datasets and the true effect. The simulation or empirical SD was calculated as the standard deviation of the estimated β_1s across simulations, indicating precision of the estimator. We also obtain average of the estimated SEs from 1000 simulations to assess accuracy of variance estimator from each simulation dataset. The overall error rate of the point estimator was captured by the estimated MSE, enumerated by the average squared difference between the estimated β_1 and true value across the 1000 datasets. Furthermore, we reported performance of the interval estimators in each model. The empirical coverage was estimated as the proportion of 95% confidence intervals that covered the true β_1, and the empirical power was the proportion of confidence intervals that rejected a false null hypothesis, i.e. zero lies outside CI. All datasets were simulated and analyzed in **R** version 2.4.1[29].

Results

Analysis of COMPETE II trial data

We applied all six models to the COMPETE II data and reported results in Table 3. Approximately equal numbers of patients were randomized to the intervention and control groups within each family doctor, leading to 253 and 258 patients in the intervention and control group, respectively. Among 46 family physicians, 11 physicians (24%) treated equal numbers of patients in two arms, 24 physicians (52%) treated one more patient in the intervention or control arm, 10 physicians (22%) managed 2 more patients in either arm, and one physician (2%) managed 3 more patients in one arm compared with the other.

All baseline characteristics were roughly balanced between arms [3]. The analyses using patient-level data produced similar estimates for β_1 and the effect size was around 0.5 times the corresponding residual standard deviation. The standard error of the estimated β_1 reduced from 0.25 (Model A) to 0.23 (Models B, C) then 0.19 (Model D) when centre effects were adjusted, leading to narrower CIs around estimated β_1 in Models B – D. The intraclass correlation was estimated 0.138 in Model C and 0.124 in Model D. The two centre-level analyses returned slightly larger estimates of β_1 than those from the individual patient-level models. In fact the minimal variance between physicians indicated no noticeable heterogeneity between physicians (τ^2=0, I^2=0), resulting in same estimates from E-1 and E-2. Zero was not contained in the 95% confidence intervals, therefore all models led to the conclusion that the experimental intervention significantly improved patient management over usual care based on the change of composite process score.

Balanced design with equal centre size

Properties of point estimates

Table 4 summarizes descriptive statistics of the point estimator of treatment effect in Models A – E for three values in the lower range of the spectrum of ICC, in the balanced design. The point estimates of β_1 were unbiased in all six models for all ICC values. Upon review, it was surprising that the point estimates in Model A, ignoring stratification and clustering, were invariant of ICC, and that the same estimates were returned by four patient-level models for each simulation. In fact, when treatments are allocated in same proportion in all centres, centre has no association with the treatment allocation, hence adjusting for centre effect or not has little impact on point estimate of the treatment – response relationship given a continuous response variable. For this reason, different ways to incorporate between-centre information (Models B -D) led to same estimates of treatment contrast in a balanced design. Same point estimates led to same empirical SD and overall error rate (measured by MSE) of the estimator in Models A – D regardless of ICC. Across different ICC values and scenarios 1 – 3, Models B and C yielded accurate estimates of the standard error of $\hat{\beta}_1$ that approximated the empirical SD and the true standard deviation, 0.149, calculated using the best linear unbiased estimator of the simulation model, i.e. Model C [18]. From Table 4, we found that the standard error of $\hat{\beta}_1$, in Model A increased with ICC in each scenario, deviating from the corresponding empirical SD. The standard error could be slightly underestimated in Model D when the number of centres was small (Table 4, scenario 2 and 3 comparing empirical SD and average SE). This agreed with previous work concerning small sample properties of the GEE model [20].

The centre-level analyses produced larger empirical SE and MSE for $\hat{\beta}_1$ compared with the patient-level analyses given small or moderate centre sizes (Table 4). The difference reduced as centre size increased. When only a few patients were enrolled per centre, the fixed-effects centre-level point estimator in Model E – 1 had large sampling variation that was severely underestimated at all ICC values. The random-effects model (E – 2) based on DerSimonian-Laird method on the other hand seemed to yield valid SE for $\hat{\beta}_1$ that was on average greater than SEs from the patient-level models. The average estimate of SE for $\hat{\beta}_1$ over all simulations in Model E – 2 was always larger than estimates of SE in Models B and C, followed by the SE estimated in Model E – 1 across different combinations of centre size and number of centres. In this study, although datasets were generated so that the treatment effects were homogeneous among centres (i.e. no treatment by centre interaction), random-effects analysis using centre-level data outperformed the fixed-effects analysis when the centre size was small, for Model E – 2 took into account the observed "heterogeneity" due to imprecise estimation of the centre mean difference and the associated standard error.

Properties of interval estimates

The empirical coverage of confidence intervals (CIs) and the statistical power in balanced studies are displayed in Table 5. Models B and C produced similar coverage close to the nominal value of 95% over different ICC values and centre composition. Model A provided conservatively high coverage increasing with ICC, illustrating that for moderate to large ICC values, CIs in Model A were abnormally wide due to overestimated SE for $\hat{\beta}_1$. The empirical coverage of CIs from Model D or E – 1 on average was farther down from 95% compared with Models B and C. This is likely caused by underestimation of the standard error in Models D and E-1, and is associated with an apparent increase of power in the first three scenarios. For Model D, the coverage dropped to below 90% when the number of centres reduced to six in scenario 3. The coverage of Model E – 1 was too low to be useful when studies were conducted at many smaller centres (scenario 1). However, coverage increased gradually with centre size and approached 95% when there were 30 patients per centre (scenario 3). Model E-2 presented similar coverage pattern to E-1, although the coverage was closer to 95%. Models B and C largely maintained nominal power of 91.8% regardless of ICC value. Power of Model A decreased dramatically as ICC departed from 0, indicating that the model failed to adjust for between-centre variation or within-centre correlation in the outcome measure. The nominal type II error rate (8%) was maintained in Models D and E – 1 in scenarios 1 – 3. Model E – 2 generally had lower power to detect the true treatment effect due to a larger standard error that reflects both the within-centre variability and treatment by centre interaction. Interestingly, this power rose as the number of centres reduced and approached 88% in scenario 3.

Overall, Models B and C had very close performance that outweighed other models in balanced design. Models C and D converged to a solution in all simulations.

Design with equal centre size and chance imbalance

Properties of point estimates

Performance of different models in multicentre studies of equal centre sizes, 1-to-1 allocation ratio and chance imbalance is displayed in Tables 6 and 7. Similar results were observed as in the balanced design, though a few differences emerged. The unbalanced allocation of patients into treatment arms due to pure within-centre variation introduced chance imbalance (in both directions) into treatment – response relationship, hence ignoring centre effects completely (as in Model A) led to unbiased yet less efficient estimates for large ICC values. Model B could be less precise than Model A given small to moderate ICC values, a phenomenon previously reported by Pickering and Weatherall [7]. As in the balanced design, the fixed- and random-effects models performed comparably for various ICC values, largely because the fixed and random intercepts for study centres cancelled out in estimating effect contrast when we fit Models B and C, and had little impact on the estimation of the fixed effect contrast across centres. However, the fixed-effects model produced larger empirical standard deviation and average standard error in scenario 4, a study being composed of many centres each managing a few patients. Adjusting for between-centre variation as random effects in Model C or using population-averaged analysis in Model D allowed to borrow information across centres and resulted in greater precision.

Properties of interval estimates

Similar results were observed relative to the balanced design. Patient – level models A – C guaranteed nominal coverage of confidence intervals at different ICC values, whereas the other models were likely to produce lower coverage under certain conditions. Among all models, Models C and D achieved the best empirical power that was closest to the nominal value of 91.8% across different centre sizes. When centre size was small and the number of centre was large (scenario 4), power for Models C and D also decreased with ICC, a pattern that was less obvious in scenarios 5 and 6. Models C and D achieved convergence in analyzing all simulated datasets.

Design with unequal centre sizes and chance imbalance

The properties of point and interval estimates in the scenarios 7 and 8 (with unequal centre sizes and chance imbalance) were close to results in the previous two designs. In particular, the comparative performance of six models lay in the middle ground between scenarios 2 and 5, as the level of imbalance between two treatments was no more than half of the block size within centres. As similar results were observed for block sizes 2 and 4, summary statistics based on block size 4 were plotted in Figures 2, 3, 4 and 5. Results suggested that unequal centre size had little impact on model performance, yet it was associated with slight enlargement of the empirical variance of $\hat{\beta}_1$ in Model E – 1. To summarize, although all six models produced unbiased point estimates, the fixed- and mixed-effects models using patient-level data provided the most accurate estimates of the standard error of $\hat{\beta}_1$ given large ICC values, hence should be used in the analysis of multicentre trials when the ICC was nontrivial or unknown to control type I and type II error rates. For studies consisting of a large number of centres with only a few patients per centre, adjusting for centre as mixed effects produced most precise point estimate of treatment effect, hence were more preferable. The information sandwich method appeared to slightly underestimate the actual variance when patients were recruited from 17 centres in scenarios 7 or 8. Due to varying centre sizes, Model D did not converge for all simulated datasets (number varied between 1 and 93 out of 1000 simulations) after 2000 iterations, when ICC was less than or equal to 0.1 or greater than 0.4 for block size of 2 or 4. Such datasets were excluded for all models and extra data were simulated to attain a total number of 1000 simulations for any ICC value. In most cases, the non-convergence of GEE occurred due to a non-positive definite working correlation matrix.

In scenario 9, as a result of mimicking the particular centre composition of the COMPETE II trial, on average, three centres out of 46 contained no patients in one of the treatment groups per simulation. These centres were removed from the fixed-effects model (Model B), as no comparison patients in the same centre were available. About six centres out of 46 recruited less than two patients per treatment arm for each simulation. These centres were dropped from the centre-level analyses, as the standard error for treatment difference per centre could not be obtained as input variables for 'metacont()'. Performance of six models in scenario 9 was similar to that in scenarios 7 and 8, although point estimates from all models appeared to be marginally biased toward the null (Table 8). Estimates from patient-level models were more precise and closer to 0.230, the best linear unbiased estimate of standard error based on the simulation model. Once again, the standard error was slightly biased upward in Model A and marginally biased downward in Model D. This resulted in wider and conservative interval estimates from Model A and slightly narrower intervals from Model D. Models B and C performed comparably, probably because on average only three centres each containing one patient were dropped from Model B, which did not affect the variance estimation much. Models C and D achieved convergence for all 1000 simulations in this scenario.

Discussion

In this paper, we investigated six modelling strategies in a Frequentist framework to study the effect of an experimental treatment compared to the control treatment in the context of multicentre RCTs with a continuous outcome. We focused on three designs with equal or varying centre sizes and a treatment allocation ratio of 1:1 in the absence of treatment by centre interaction. Results of this simulation study showed that, when the proportion of patients allocated to the experimental treatment was the same in each centre or subject to chance imbalance only, models using patient-level and centre-level data yielded unbiased point estimates of treatment effect across a wide spectrum of ICC values. Ignoring stratification by centre or within-centre correlation did not bias the estimated treatment effects even when ICC was large. In fact, Parzen et al showed that mathematically the usual two-sample t-test, naively assuming independent observations of the response within centre was asymptotically unbiased in this context [30].

The simulation study also indicated that these models produced different standard errors of $\hat{\beta}_1$, and the properties of interval estimates were affected by several factors: whether and how centre effects were incorporated in analysis, the combination of centre size and number of participating centres, and the level of non-orthogonality of the observed data . Treating centre as a random intercept resulted in the most precise estimate, and nominal values of coverage and power were attained in all circumstances. The fixed-effects model had extremely similar performance compared with the mixed-effects model in balanced design, but was slightly less efficient when the number of centres was large (J>20) in an unbalanced design. Pickering and Weatherall observed the same pattern in their simulation study comparing three patient-level models with small ICC values [7]. The GEE model using information sandwich covariance method tended to underestimate the standard error across centre effects when the sample of centres was small, a property noticed by researchers [20, 31]. This resulted in higher statistical power. That is, the treatment effect estimate was more likely to be significant with a smaller standard error, but was associated with a lower coverage of the conference interval. Marray et al suggested that at least 40 centres should be used to ensure reliable estimate of standard error in the context of cluster randomized trials [32]. Our simulation results suggested that such cut value was also applicable to multicentre RCTs. Failure to control for centre effects in any form resulted in inflation of standard error, falsely high interval coverage and sizable drop of power, as ICC increased. Parzen et al quantified the impact of correlation among observations within centre on the variance of $\hat{\beta}_1$ in Model A as 1/(1-ICC) [30]. Alternatively, one may consider a variant of robust variance estimation or a GEE model with an independent working correlation to control for the impact of ICC on variance estimation using t-test. Centre-level models generally produced larger standard errors, lower coverage or power than the patient-level models. Centre-level random-effects model incorporated variability of the treatment effect over centres, and was not a fair comparator to other models. Interestingly, this model seemed to fare better than the centre-level fixed-effects model in terms of precision and coverage even though the simulated datasets contained no treatment by centre interaction. Despite that the random-effects centre-level model may be a reasonable alternative for patient-level models when the number of patients per centre is large (≥30), centre-level models cannot adjust for patient-level covariates, a potential fatal drawback in the presence of patient prognostic imbalance.

Statisticians have different viewpoints on treating centre effects and treatment by centre interaction as fixed or random effects when analyzing multicentre RCTs [12, 13, 21, 33]. Our simulation results demonstrated the advantage of treating centres as random intercepts in the absence of treatment by centre interaction. When many centres enrol a few patients and allocation is unbalanced, the random intercept models can give more precise estimates of the treatment effect than the fixed intercept models, because they recover inter-centre information in unbalanced situations. For instance, in a multicentre RCT consisting of 45 centres each recruiting 4 patients, the empirical variance of the estimator of the treatment effect resulting from the fixed-effects model was 24.8% and 26.0% greater than that from the random-effects model when the ICC was 0.01 and 0.05, respectively. In the sentence alluded to, we need to compare the empirical variance of 0.162^2 with the value of 0.145^2 for ICC = 0.01, and 0.174^2 to 0.155^2 for ICC = 0.05 (Table 6, scenario 4). We therefore take the same position as Grizzle [33] and Agresti and Hartzel [12] that, “Although the clinics are not randomly chosen, the assumption of random clinic effect will result in tests and confidence intervals that better capture the variability inherent in the system more realistically than clinical effects are considered fixed”.

Our results have some implications for the design of multicentre RCTs in the absence of treatment by centre interaction. First, regardless of the pre-determined allocation ratio, permutated block randomization (of relatively small block sizes) should be used to maintain approximate balance or orthogonality (i.e. same treatment allocation proportion across centres [7]) between treatments and centres, so that their individual effects can be evaluated independently. Variable block sizes can be used to strengthen allocation concealment. Second, for a given sample size, the number of patients randomized in majority of centres should be sufficiently large to ensure reliable estimate of within-centre variation. Third, it is essential for investigators to obtain a rough estimate of ICC for within-centre responses, through literature review or a pilot study. To reach nominal power of 80% or 90% (in the absence of clustering), centre effects should be taken into consideration in sample size assessment. When centre effects are included without treatment by centre interaction, the analysis becomes more powerful than a two-sample t-test. One method to assess sample size is to start with a two sample t-test for continuous outcomes (ignoring centre effect) then multiple the original estimated error variance by an variation inflation factor of 1/(1-ICC). This factor would have the effect of increasing the required sample size. Ignoring centre effects results in the larger sample size in the absence of interaction. Sample size determined using information sandwich covariance of GEE model could lead to slight loss of power, when the number of centres is small (≥40) and no proper adjustment is done. Lastly, there is no particular reason to require equal numbers of patients being enrolled in all participating centres and this is seldom the case in practice. Throughout the simulations, we observed similar results for studies of equal and varying centre sizes. In the study, we considered three scenarios representing the particular centre composition of the COMPETE II trial. For discussion on potential impact of enrolment patterns on the point and interval estimates of treatment effect, readers can refer to the publications on random enrolment verse determined enrolment, and relative efficiency between equal and unequal cluster sizes in the reference list [34, 35].

The current ICH E9 guideline recommends that researchers investigate treatment effect using a model that allows for centre differences in the absence of treatment by centre interaction [1]. However, it is implausible or impractical to include centre effects in statistical modelling or stratify randomization by centre, when it is anticipated from the start that trials may have very few subjects per centre. As it is acknowledged in the document, these recommendations are based on fixed-effects models. Mixed-effects models on the other hand may also be used to explore the centre and centre by interaction effects, especially when the number of centres is large [1]. Our simulation results indicated that when a considerable number of centres contains only a few patients, adjusting for centre as a fixed effect may lead to reduced precision (depending on distribution of patients between arms) compared with the naïve unadjusted analysis. Our work complements the ICH E9 guideline, by studying the impact of intraclass correlation on the assessment of treatment effects – a challenge that is seldom discussed, although routinely faced by investigators in reality. Our investigation suggests that, (1) ignoring centre effects completely may cause substantial overestimation of the standard error, faulty increase of coverage of the confidence interval and reduction of power; and (2) mixed-effects models and GEE models, if employed appropriately, can produce accurate and precise effect estimates, regardless of the degree of clustering. We recommend consider these methods in developing future guidelines.

When the number of patients per centre is very small, it is not practical to include centre as a fixed effect to analyze patient-level data, as centre effects cannot be reliably estimated, and precision of the treatment effect will be compromised. In fact for extremely small centres, all patients may be allocated to the same treatment group, and such centres will be ignored by the fixed-effects model [36-39]. The alternatives include collapsing all centres to perform a two-sample t-test, collapsing smaller centres to create an artificial centre and treating it as a fixed effect, and exploring other models discussed above. The mixed-effects model utilizes small centres more efficiently by "borrowing" information from larger centres. The GEE approach models the average treatment difference across all centres and adjusts for centre effects through a uniform correlation structure. This is an intuitively more efficient model which unfortunately does not always converge when the number of patients per centre was highly variable (simulation scenarios 7 and 8). In the current study, non-convergence problems were more likely to arise for very small or large ICC values (less than 0.1 or greater than 0.4 for block size 2 or 4) due to non-positive definite working correlation matrices, and the frequency could be as big as 10% after 2000 iterations. Conversely, convergence problems did not occur for the mixed-effects models in any scenarios. Our results show that analysis of trials consisting of very small centres (i.e. those containing less than 2 patients per arm) using centre-level models may not be an optimal strategy, because the within-centre standard deviation of treatment difference cannot be estimated for such centres, and consequently these very small centres are excluded from the analysis.

Results of two large empirical studies and one systematic review of cluster RCTs in primary care clinics suggested that most ICC values on physical, functional and social measures were less than 0.10 [26-28]. The estimated ICC in the COMPETE II trial using GEE and linear mixed-effects model, on the other hand, was 0.124 and 0.138, respectively. We chose to include rare yet possible large ICC values (0-0.75) in this simulation to examine the overall trend of model performance by ICC, and for the purpose of completeness and generalizability. Readers should anticipate the ICC values likely to emerge from their studies when interpreting these results. Throughout the work, we quantified correlation among subjects within centre using ICC, the most commonly used concept to assess clustering in biomedical literature. As indicated in previous sections, ICC reflects the interplay of two variance components in multicentre data: the between-centre variance and within-centre variance. These variance components are relatively easy to interpret for analysis of continuous outcomes using linear models. For analysis of binary or time-to-event data from multicentre trials using generalized mixed and frailty models, interpretation of centre heterogeneity can present challenges because random effects are linked to the outcome via nonlinear functions [40]. Reparameterization of the probability density function may be used to assess the impact of within- and between-centre variance. Interested readers can refer to Duchateau and Janssen [40] for more details.

A major limitation of the study is that it did not address model performance when the treatment by centre interaction exists. The interactions may be due to different patient populations or variable standard of care. Interested readers may read Moerbeek et al [6] for formulas of variance of $\hat{\beta}_1$ in different models and Jones et al [14] for simulation results. Future studies addressing interaction effects in multicentre RCTs are needed. Datasets in the current paper were generated based on a moderate treatment effect reflected by the standardized mean difference between the treatment and control group. More or less prominent treatment effects are also likely to occur in clinical studies and similar findings are expected. The current study investigated on continuous outcomes in two groups from a Frequentist perspective. The models discussed above can be naturally extended to compare three or more treatments. Agresti and Hartzel [12] surveyed different methods to evaluate treatments for binary outcomes in multicentre RCTs. Non-parametric approaches and Bayesian methods are also available to obtain treatment contrast. Interested readers can refer to Aitkin [41], Gould [11], Smith et al [42], Legrand et al [16], and Louis [43], to name a few.

Conclusions

We used simulations to investigate the performance of six statistical approaches that have been advocated to analyze continuous outcomes in multicentre RCTs. Our simulation study showed that all six models produced unbiased estimates of treatment effect in individual patient randomization multicentre trials. Adjusting for centre as random effects resulted in more efficient effect estimates in all scenarios over a wide spectrum of ICC values and various centre compositions. Fixed-effects model performed comparably to the mixed-effects model under most circumstances but lost efficiency when many centres contained a relatively small number of patients. The GEE model underestimated standard error of the effect estimates when a small number of centres were involved, and did not always converge when the centre size was variable for very large or small ICC values. Two-sample t-test severely overestimated standard error given moderate to large ICC values. The relative efficiencyof statistical modelling of treatment contrasts was also affected by ICC, distribution of patient enrolment, centre size and the number of centres.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RC participated in the design of the study, simulation, analysis and interpretation of data, and drafting and revision of the manuscript. LT contributed to the conception and design of the study, interpretation of data and revision of the manuscript. JM contributed to the design of the study and revision of the manuscript. AH contributed to acquisition of data and critical revision of the manuscript. EP and PJD advised on critical revision of the manuscript for important intellectual content. All authors have read and approved the final manuscript.

Acknowledgements

We thank the reviewers and Dr. Werner Vach for their constructive comments that have helped us to improve this work. We thank Ms. Ruthanne Cameron for proofreading the manuscript. Dr. Thabane is a clinical trials mentor for the Canadian Institutes of Health Research. Rong Chu was supported in part by funding from the CANNeCTIN (Canadian Network and Centre for Trials Internationally) training award, Father Sean O'Sullivan Research Center (FSORC) Studentship award and the Canadian Institute of Health Research (CIHR) doctoral research award.

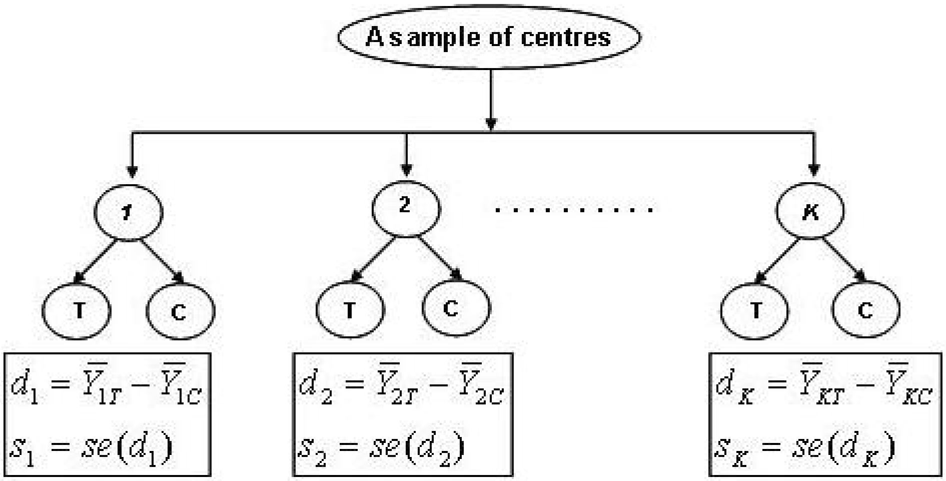
References

1. International Conference on Harmonisation E9 Expert Working Group. **ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials.** *Stat Med* 1999, **18**(15):1905-1942.
2. Lachin JM: *Biostatistical methods: the assessment of relative risks:* 1st ed. New York: John Wiley and Sons; 2000.
3. Holbrook A, Thabane L, Keshavjee K, Dolovich L, Bernstein B, Chan D, Troyan S, Foster G, Gerstein H, COMPETE II Investigators: **Individualized electronic decision support and reminders to improve diabetes care in the community: COMPETE II randomized trial.** *CMAJ* 2009, **181**(1-2):37-44.
4. Donner A, Klar N: *Design and analysis of cluster randomization trials in health research:* 1st ed. London: Arnold; 2000.
5. Kerry SM, Bland JM: **The intracluster correlation coefficient in cluster randomisation.** *BMJ* 1998, **316**(7142):1455.
6. Moerbeek M, van Breukelen GJ, Berger MP: **A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies.** *J Clin Epidemiol* 2003, **56**(4):341-350.
7. Pickering RM, Weatherall M: **The analysis of continuous outcomes in multi-centre trials with small centre sizes.** *Stat Med* 2007, **26**(30):5445-5456.
8. Worthington H: **Methods for pooling results from multi-center studies.** *J Dent Res* 2004, **83 Spec No C**:C119-21.
9. Liang KY, Zeger SL: **Longitudinal data analysis using generalized linear models.** *Biometrika* 1986, **73**:13-22.
10. Whitehead A: *Meta-analysis of controlled clinical trials:* 1st ed. Chichester: John Wiley and Sons; 2002.

11. Gould AL: **Multi-centre trial analysis revisited.** *Stat Med* 1998, **17**(15-16):1779-97; discussion 1799-800.
12. Agresti A, Hartzel J: **Strategies for comparing treatments on a binary response with multi-centre data.** *Stat Med* 2000, **19**(8):1115-1139.
13. Fleiss JL: **Analysis of data from multiclinic trials.** *Control Clin Trials* 1986, **7**(4):267-275.
14. Jones B, Teather D, Wang J, Lewis JA: **A comparison of various estimators of a treatment difference for a multi-centre clinical trial.** *Stat Med* 1998, **17**(15-16):1767-77; discussion 1799-800.
15. Glidden DV, Vittinghoff E: **Modelling clustered survival data from multicentre clinical trials.** *Stat Med* 2004, **23**(3):369–388.
16. Legrand C, Ducrocq V, Janssen P, Sylvester R, Duchateau L: **A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model.** *Stat Med* 2005, **24**(24):3789-3804.
17. Brown HK, Kempton RA: **The application of REML in clinical trials.** *Stat Med* 1994, **13**(16):1601-1617.
18. McLean RA, Sanders WL: **Approximating the degrees of freedom for SE's in mixed linear models**. Proceedings of the Statistical Computing Section of the American Statistical Association. New Orleans, Louisiana; 1988.
19. Twisk J: *Applied longitudinal data analysis for epidemiology: a practical guide.* Cambridge: Cambridge University Press; 2003.
20. Ukoumunne OC, Carlin JB, Gulliford MC: **A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials.** *Stat Med* 2007, **26**(18):3415-3428.
21. Senn S: **Some controversies in planning and analysing multi-centre trials.** *Stat Med* 1998, **17**(15-16):1753-65; discussion 1799-800.
22. Lin Z: **An issue of statistical analysis in controlled multi-centre studies: how shall we weight the centres?** *Stat Med* 1999, **18**(4):365-373.
23. Kallen A: **Treatment-by-centre interaction: what is the issue.** *Drug Info J* 1997, **31**:927-936.-936.
24. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7**(3):177-188.
25. Cohen J: *Statistical Power Analysis for the Behavioral Sciences:* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
26. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ: **Patterns of intra-cluster correlation from primary care research to inform study design and analysis.** *J Clin Epidemiol* 2004, **57**(8):785-794.
27. Parker DR, Evangelou E, Eaton CB: **Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART).** *Contemp Clin Trials* 2005, **26**(2):260-267.
28. Smeeth L, Ng ES: **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.** *Control Clin Trials* 2002, **23**(4):409-421.

29. R Core Development Team: *R: A language and environment for statistical computing:* 1st ed. Vienna: R Foundation for Statistical Computing; 2005.
30. Parzen M, Lipsitz S, Dear K: **Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial?** *Biomtrc J* 1998, **40**:385-402.
31. Mancl LA, DeRouen TA: **A covariance estimator for GEE with improved small-sample properties.** *Biometrics* 2001, **57**(1):126-134.
32. Murray DM, Varnell SP, Blitstein JL: **Design and analysis of group-randomized trials: a review of recent methodological developments.** *Am J Public Health* 2004, **94**(3):423-432.
33. Grizzle JE: **Letter to the editor.** *Control Clin Trials* 1987, **8**(4):392-393.
34. van Breukelen GJ, Candel MJ, Berger MP: **Relative efficiency of unequal cluster sizes for variance component estimation in cluster randomized and multicentre trials.** *Stat Methods Med Res* 2008, **17**(4):439-458.
35. Fedorov V, Jones B: **The design of multicentre trials.** *Stat Methods Med Res* 2005, **14**(3):205-248.
36. Localio AR, Berlin JA, Ten Have TR, Kimmel SE: **Adjustments for center in multicenter studies: an overview.** *Ann Intern Med* 2001, **135**(2):112-123.
37. Breslow NE, Day NE: *Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies.* Lyon: International Agency for Research on Cancer; 1980.
38. Cox DR, Hinkley DV: *Theoretical Statistics:* London: Chapman & Hall; 1974.
39. Graubard BI, Korn EL: **Regression analysis with clustered data.** *Stat Med* 1994, **13**(5-7):509-522.
40. Duchateau L, Janssen P: **Understanding heterogeneity in generalized mixed and frailty models.** *The American Statistician* 2005, **59**(2):143-146.
41. Aitkin M: **A general maximum likelihood analysis of variance components in generalized linear models.** *Biometrics* 1999, **55**(1):117-128.
42. Smith TC, Spiegelhalter DJ, Thomas A: **Bayesian approaches to random-effects meta-analysis: a comparative study.** *Stat Med* 1995, **14**(24):2685-2699.
43. Louis TA: **Using empirical Bayes methods in biopharmaceutical research.** *Stat Med* 1991, **10**(6):811-27; discussion 828-9.

Figure 1 A schematic of fixed- and random-effects centre-level models


A sample of centres
1
2
. . . .
K
T
C
T
C
T
C
$d_1 = \overline{Y}_{1T} - \overline{Y}_{1C}$
$s_1 = se(d_1)$
$d_2 = \overline{Y}_{2T} - \overline{Y}_{2C}$
$s_2 = se(d_2)$
$d_K = \overline{Y}_{KT} - \overline{Y}_{KC}$
$s_K = se(d_K)$

Fixed effects: (E -1) $$\hat{D}_{FW} = \frac{\sum_k w_k d_k}{\sum_k w_k}$$, where weight $w_k = \frac{1}{s_k^2}$

Random effects: (E - 2) $$\hat{D}_{RW} = \frac{\sum_k u_k d_k}{\sum_k u_k}$$, where weight $u_k = \frac{1}{s_k^2 + \textbf{heterogeneity}}$

Figure 2 Empirical standard deviation (SD) across 1000 simulations by ICC for scenario 8 (block size = 4)

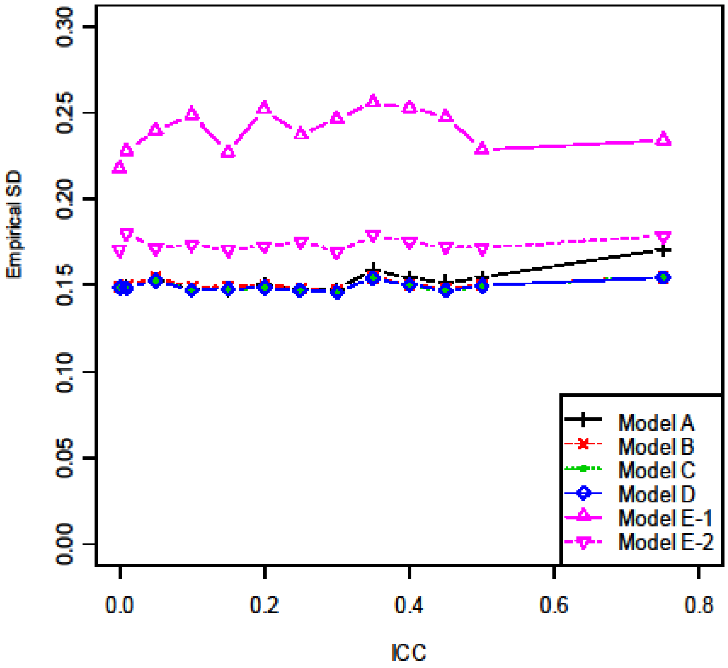


Figure 3 Average of standard error (SE) across 1000 simulations by ICC for scenario 8 (block size = 4)

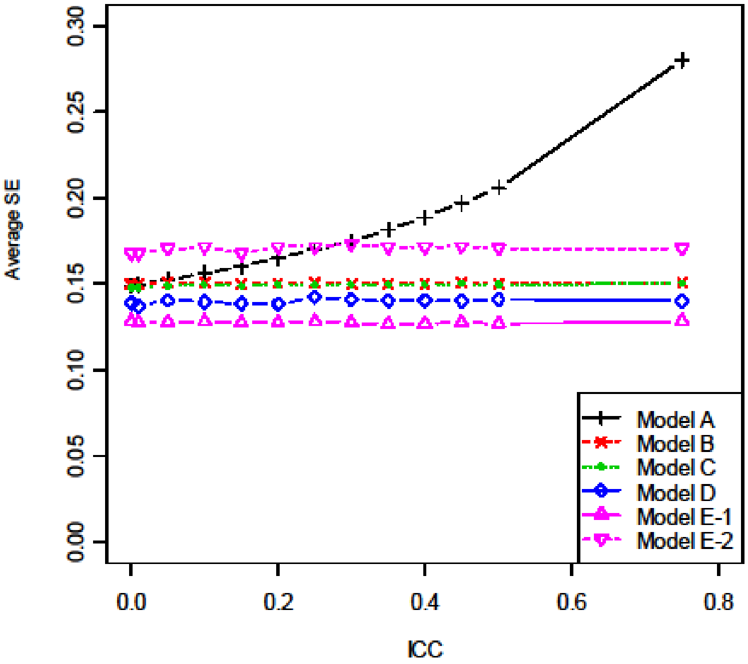


Figure 4 Coverage of 95% CI by ICC for scenario 8 (block size = 4)

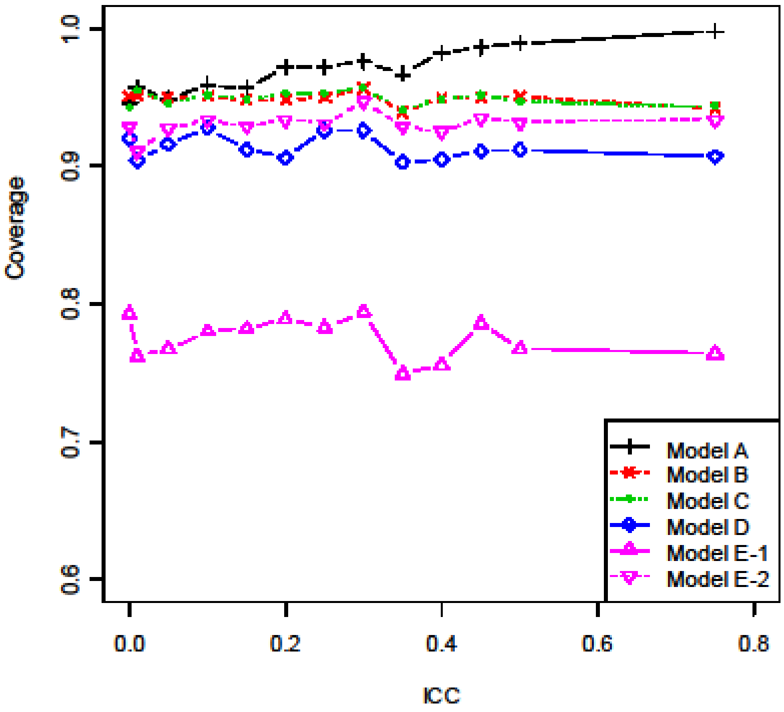


Figure 5 Empirical power by ICC for scenario 8 (block size = 4)

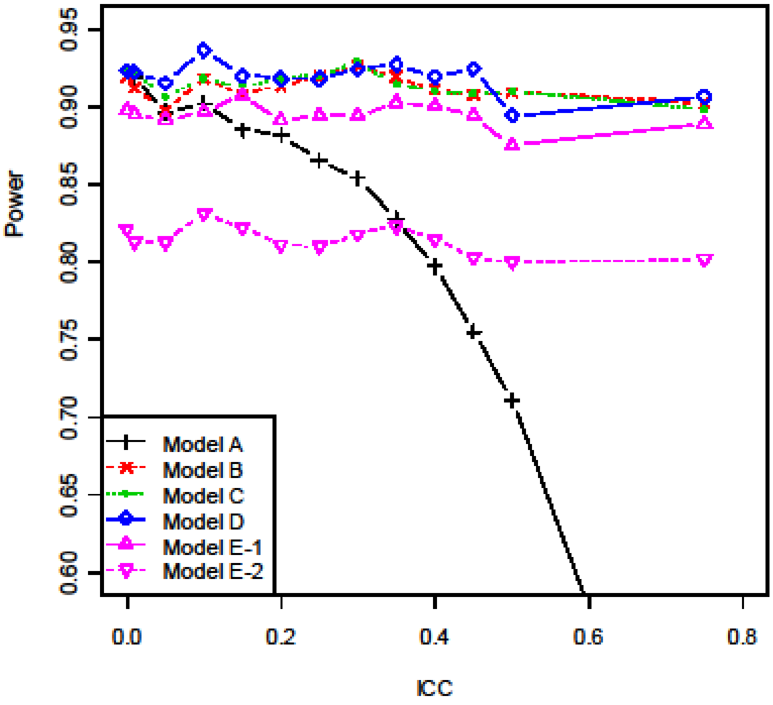


Table 1 Catalogue of simulation designs

Design	**Scenario**	**Number of centres**	**Centre size**	**ICC**
Balance	1	45	4	0 – 0.75
	2	18	10	
	3	6	30	
Chance imbalance	4	45	4	
	5	18	10	
	6	6	30	
Blocking (size = 2)	7	17	1, 1, 4, 5, 5, 5, 8, 8, 10, 10, 10 10, 15, 15, 20, 25, 28	
Blocking (size = 4)	8	17		
Blocking (size = 6)	9	46	Same as Table 2	0.125

ICC: Intraclass (intracentre) correlation

Table 2 Centre composition of the COMPETE II trial

Number of patients per centre	**Number of Centres**
1	3
2	0
3	1
4	4
5	1
6	1
7	2
8	3
9	4
10	3
11	5
12	3
13	2
14	0
15	3
16	3
17	0
18	2
19	2
20	1
21	0
22	1
23	1
24	0
25	1

Table 3 Estimates of intervention effects in COMPETE II trial

Model	Estimate of intervention effect	SE	95% CI	Variance component
A: Simple linear regression	1.270	0.246	(0.787, 1.753)	σ_e^2=7.712
B: Fixed-effects regression	1.291	0.231	(0.836, 1.745)	σ_e^2=6.682
C: Mixed-effects regression†	1.263	0.230	(0.811, 1.714)	σ_e^2=6.678 σ_b^2=1.069
D: GEE‡	1.263	0.193	(0.884, 1.641)	
E – 1: centre-level Fixed-effects model	1.397	0.219	(0.967, 1.826)	
E – 2: centre-level Random-effects model	1.397	0.219	(0.967, 1.826)	

SE: standard error; CI: confidence interval; σ_e^2 *: within-centre variance;* σ_b^2 *: between-centre variance; ICC: intraclass (intracentre) correlation*
† *ICC=0.138*
‡ *ICC=0.124*

Table 4 Properties of point estimates of the treatment effect from Models A – E in scenarios 1 to 3

	ICC = 0.01			ICC=0.05			ICC=0.20		
Model	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE
Scenario 1 – balanced design, 45 centres each with 4 subjects									
A	0.496 (0.148)	0.149	0.022	0.499 (0.146)	0.152	0.021	0.497 (0.151)	0.167	0.023
B	0.496 (0.148)	0.148	0.022	0.499 (0.146)	0.148	0.021	0.497 (0.151)	0.149	0.023
C	0.496 (0.148)	0.147	0.022	0.499 (0.146)	0.148	0.021	0.497 (0.151)	0.149	0.023
D	0.496 (0.148)	0.146	0.022	0.499 (0.146)	0.146	0.021	0.497 (0.151)	0.147	0.023
E-1	0.496 (0.494)	0.066	0.244	0.491 (0.454)	0.066	0.206	0.506 (0.447)	0.065	0.200
E-2	0.499 (0.163)	0.172	0.027	0.497 (0.166)	0.170	0.027	0.494 (0.162)	0.170	0.026
Scenario 2 – balanced design, 18 centres each with 10 subjects									
A	0.490 (0.149)	0.150	0.022	0.504 (0.155)	0.152	0.024	0.498 (0.145)	0.166	0.021
B	0.490 (0.149)	0.149	0.022	0.504 (0.155)	0.149	0.024	0.498 (0.145)	0.149	0.021
C	0.490 (0.149)	0.148	0.022	0.504 (0.155)	0.148	0.024	0.498 (0.145)	0.149	0.021
D	0.490 (0.149)	0.142	0.022	0.504 (0.155)	0.143	0.024	0.498 (0.145)	0.142	0.021
E-1	0.490 (0.178)	0.130	0.032	0.501 (0.180)	0.130	0.032	0.498 (0.171)	0.130	0.029
E-2	0.492 (0.164)	0.154	0.027	0.503 (0.165)	0.155	0.027	0.498 (0.158)	0.153	0.025

Table 4 (*continued*)

	ICC = 0.01			ICC=0.05			ICC=0.20		
Model	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE
Scenario 3 – balanced design, 6 centres each with 30 subjects									
A	0.496 (0.149)	0.149	0.022	0.492 (0.149)	0.152	0.022	0.504 (0.151)	0.164	0.023
B	0.496 (0.149)	0.149	0.022	0.492 (0.149)	0.149	0.022	0.504 (0.151)	0.149	0.023
C	0.496 (0.149)	0.149	0.022	0.492 (0.149)	0.149	0.022	0.504 (0.151)	0.149	0.023
D	0.496 (0.149)	0.130	0.022	0.492 (0.149)	0.130	0.022	0.504 (0.151)	0.149	0.023
E-1	0.497 (0.153)	0.144	0.023	0.491 (0.154)	0.144	0.024	0.508 (0.156)	0.144	0.024
E-2	0.497 (0.151)	0.163	0.023	0.491 (0.151)	0.163	0.023	0.507 (0.153)	0.161	0.023

SD: empirical standard deviation; Ave. SE: average estimated SE; MSE: mean squared error; ICC: intraclass (intracentre) correlation

Table 5 Coverage of the 95% interval estimate of the treatment effect and statistical power of Models A – E in scenarios 1 to 3

	ICC = 0.01		ICC = 0.05		ICC = 0.20	
Model	Cover. of CI	Power	Cover. of CI	Power	Cover. of CI	Power
Scenario 1 – balanced design, 45 centres each with 4 subjects						
A	0.952	0.901	0.953	0.912	0.973	0.862
B	0.947	0.905	0.945	0.924	0.951	0.899
C	0.947	0.907	0.944	0.924	0.951	0.899
D	0.941	0.911	0.936	0.931	0.933	0.902
E-1	0.286	0.902	0.294	0.920	0.320	0.912
E-2	0.933	0.810	0.921	0.818	0.938	0.821
Scenario 2 – balanced design, 18 centres each with 10 subjects						
A	0.955	0.899	0.941	0.903	0.973	0.881
B	0.954	0.906	0.935	0.906	0.954	0.916
C	0.951	0.908	0.935	0.906	0.954	0.916
D	0.929	0.909	0.902	0.919	0.940	0.924
E-1	0.845	0.904	0.835	0.917	0.857	0.924
E-2	0.921	0.868	0.905	0.886	0.938	0.875
Scenario 3 – balanced design, 6 centres each with 30 subjects						
A	0.953	0.905	0.949	0.901	0.966	0.888
B	0.948	0.907	0.947	0.906	0.952	0.918
C	0.948	0.907	0.946	0.906	0.952	0.918
D	0.860	0.915	0.854	0.931	0.867	0.929
E-1	0.939	0.918	0.931	0.910	0.926	0.927
E-2	0.952	0.867	0.949	0.846	0.953	0.880

Cover. of CI: coverage proportion of 95% confidence interval; ICC: intraclass (intracentre) correlation

Table 6 Properties of point estimates of the treatment effect from Models A – E in scenarios 4 to 6

	ICC = 0.01			ICC=0.05			ICC=0.20		
Model	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE
Scenario 4 – chance imbalance, 45 centres each with 4 subjects									
A	0.502 (0.146)	0.150	0.021	0.511 (0.154)	0.154	0.024	0.494 (0.168)	0.166	0.028
B	0.506 (0.162)	0.172	0.026	0.510 (0.174)	0.172	0.030	0.496 (0.180)	0.172	0.032
C	0.502 (0.145)	0.149	0.021	0.511 (0.155)	0.152	0.024	0.496 (0.165)	0.159	0.027
D	0.501 (0.146)	0.146	0.021	0.511 (0.155)	0.149	0.024	0.496 (0.165)	0.155	0.027
E-1	0.492 (0.525)	0.122	0.275	0.504 (0.544)	0.126	0.296	0.481 (0.482)	0.127	0.232
E-2	0.506 (0.274)	0.265	0.075	0.515 (0.284)	0.269	0.081	0.490 (0.285)	0.260	0.081
Scenario 5 – chance imbalance, 18 centres each with 10 subjects									
A	0.495 (0.148)	0.150	0.022	0.498 (0.150)	0.153	0.023	0.497 (0.169)	0.166	0.028
B	0.494 (0.156)	0.157	0.024	0.498 (0.152)	0.157	0.023	0.500 (0.161)	0.157	0.026
C	0.495 (0.148)	0.150	0.022	0.498 (0.149)	0.151	0.022	0.499 (0.159)	0.153	0.025
D	0.494 (0.148)	0.142	0.022	0.498 (0.150)	0.144	0.022	0.499 (0.159)	0.148	0.025
E-1	0.488 (0.206)	0.130	0.042	0.498 (0.199)	0.130	0.039	0.503 (0.204)	0.130	0.042
E-2	0.490 (0.177)	0.163	0.031	0.501 (0.172)	0.162	0.030	0.501 (0.178)	0.164	0.032

Table 6 (*continued*)

	ICC = 0.01			ICC=0.05			ICC=0.20		
Model	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE	Mean effect (SD)	Ave. SE	MSE
Scenario 6 – chance imbalance, 6 centres each with 30 subjects									
A	0.499 (0.153)	0.149	0.023	0.502 (0.150)	0.153	0.022	0.510 (0.165)	0.164	0.027
B	0.499 (0.155)	0.150	0.024	0.501 (0.150)	0.151	0.022	0.507 (0.151)	0.152	0.023
C	0.499 (0.153)	0.149	0.024	0.503 (0.149)	0.150	0.022	0.507 (0.151)	0.151	0.023
D	0.498 (0.154)	0.129	0.024	0.503 (0.150)	0.131	0.022	0.508 (0.151)	0.134	0.023
E-1	0.498 (0.159)	0.146	0.025	0.502 (0.156)	0.146	0.024	0.507 (0.157)	0.146	0.025
E-2	0.498 (0.157)	0.165	0.025	0.502 (0.153)	0.164	0.023	0.507 (0.154)	0.167	0.024

SD: empirical standard deviation; Ave. SE: average estimated SE; MSE: mean squared error; ICC: intraclass (intracentre) correlation

Table 7 Coverage of the 95% interval estimate of the treatment effect and statistical power of Models A – E in scenarios 4 to 6

	ICC = 0.01		ICC = 0.05		ICC = 0.20	
Model	Cover. of CI	Power	Cover. of CI	Power	Cover. of CI	Power
Scenario 4 – chance imbalance, 45 centres each with 4 subjects						
A	0.954	0.910	0.949	0.918	0.942	0.845
B	0.966	0.846	0.946	0.822	0.931	0.810
C	0.954	0.912	0.945	0.917	0.934	0.870
D	0.948	0.924	0.934	0.926	0.934	0.878
E-1	0.411	0.782	0.417	0.793	0.424	0.745
E-2	0.897	0.468	0.900	0.501	0.887	0.468
Scenario 5 – chance imbalance, 18 centres each with 10 subjects						
A	0.954	0.898	0.949	0.900	0.942	0.843
B	0.946	0.874	0.959	0.891	0.937	0.891
C	0.952	0.898	0.951	0.904	0.939	0.895
D	0.922	0.916	0.932	0.911	0.910	0.900
E-1	0.776	0.868	0.794	0.890	0.791	0.895
E-2	0.905	0.810	0.918	0.834	0.918	0.839
Scenario 6 – chance imbalance, 6 centres each with 30 subjects						
A	0.950	0.897	0.953	0.905	0.961	0.860
B	0.949	0.892	0.954	0.907	0.961	0.916
C	0.950	0.897	0.952	0.905	0.959	0.910
D	0.856	0.911	0.879	0.918	0.874	0.908
E-1	0.922	0.904	0.931	0.921	0.931	0.913
E-2	0.944	0.831	0.951	0.867	0.955	0.857

Cover. of CI: coverage proportion of 95% confidence interval; ICC: intraclass (intracentre) correlation

Table 8 Properties of point and 95% interval estimates calculated from Models A – E based on 1000 simulated datasets in scenario 9 – unbalanced, 46 centres, same centre composition as the COMPETE II trial.

SCENARIO 9					
Model	Mean effect (SD)	Ave. SE	MSE	Cover. of CI	Power
ICC = 0.125					
A	1.254 (0.236)	0.249	0.056	0.965	0.999
B	1.253 (0.240)	0.236	0.058	0.952	1
C	1.253 (0.237)	0.235	0.056	0.949	0.999
D	1.253 (0.237)	0.230	0.056	0.944	0.999
E-1	1.256 (0.405)	0.207	0.165	0.787	0.991
E-2	1.257 (0.270)	0.261	0.073	0.935	0.995

SD: empirical standard deviation; Ave. SE: average estimated SE; MSE: mean squared error; Cover. of CI: coverage proportion of 95% confidence interval; ICC: intraclass (intracentre) correlation

CHAPTER 3

Assessment and Implication of Prognostic Imbalance in Randomized Controlled Trials with a Binary Outcome – A Simulation Study

Rong Chu[1], Stephen D. Walter[1], Gordon Guyatt[1], P. J. Devereaux[1,2], Michael Walsh[1], Kristian Thorlund[1], Lehana Thabane[1,2,3,4]

Affiliations:

[1]Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada, L8N 3Z5

[2]Population Health Research Institute, Hamilton Health Sciences, 237 Barton Street East, Hamilton, Ontario, Canada, L8L 2X2

[3]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare Hamilton, 50 Charlton Avenue East, Hamilton, ON, Canada L8N 4A6

[4]The Centre for Evaluation of Medicines, St Joseph's Healthcare Hamilton, 105 Main Street East, Level P1, Hamilton, Ontario, Canada, L8N 1G6

Corresponding author:

Rong Chu
Email: chur@mcmaster.ca
Telephone: (905) 522 1155 ext. 34918
Fax: (905) 308 7212

Keywords:

Randomized controlled trial, baseline covariates, covariate adjustment, prognostic imbalance, sample size

Cited as: Chu R, Walter SD, Guyatt G, Devereaux PJ, Walsh M, Thorlund K, et al. (2012) Assessment and Implication of Prognostic Imbalance in Randomized Controlled Trials with a Binary Outcome – A Simulation Study. PLoS ONE 7(5): e36677. doi:10.1371/journal.pone.0036677

ABSTRACT

Background

Chance imbalance in baseline prognosis of a randomized controlled trial can lead to over or underestimation of treatment effects, particularly in trials with small sample sizes. Our study aimed to (1) evaluate the probability of imbalance in a binary prognostic factor (PF) between two treatment arms, (2) investigate the impact of prognostic imbalance on the estimation of a treatment effect, and (3) examine the effect of sample size (n) in relation to the first two objectives.

Methods

We simulated data from parallel-group trials evaluating a binary outcome by varying the risk of the outcome, effect of the treatment, power and prevalence of the PF, and n. Logistic regression models with and without adjustment for the PF were compared in terms of bias, standard error, coverage of confidence interval and statistical power.

Results

For a PF with a prevalence of 0.5, the probability of a difference in the frequency of the PF ≥ 5% reaches 0.42 with 125/arm. Ignoring a strong PF (relative risk =5) leads to underestimating the strength of a moderate treatment effect, and the underestimate is independent of n when n is > 50/arm. Adjusting for such PF increases statistical power. If the PF is weak (RR=2), adjustment makes little difference in statistical inference. Conditional on a 5% imbalance of a powerful PF, adjustment reduces the likelihood of

large bias. If an absolute measure of imbalance ≥ 5% is deemed important, including 1000 patients/arm provides sufficient protection against such an imbalance. Two thousand patients/arm may provide an adequate control against large random deviations in treatment effect estimation in the presence of a powerful PF.

Conclusions

The probability of prognostic imbalance in small trials can be substantial. Covariate adjustment improves estimation accuracy and statistical power, and hence should be performed when strong PFs are observed.

Introduction

Because randomization attempts to balance the distribution of known and unknown prognostic factors (PFs) between treatment groups, authorities view it as critical for ensuring unbiased assessment of treatment effects [1]. Despite randomization, imbalance in PFs as a result of chance (chance imbalance) may still arise, and with small to moderate sample sizes such imbalance may be substantial [2, 3]. Ignoring chance imbalance in key PFs between treatment groups may result in a biased estimate of the treatment effect, particular when a large between-group difference occurs in a powerful PF [4-7].

Control for unbalanced PFs is often achieved via statistical techniques such as regression analysis, sometimes in conjunction with other design features such as stratified randomization. Adjusting for balanced or marginally unbalanced PFs of high predictive value increases statistical power and reduces sample size requirements [8-13]. While including balanced baseline covariates in linear models does not change the estimate of treatment effect, omitting balanced covariates in logistic regression models may lead to underestimation of subject-specific treatment effects [14-16]. Although guidelines for RCTs recommend conducting both unadjusted and adjusted analyses [17-19], only a minority of trials report adjusted analyses [13, 20]. Moreover, although recommendations also suggest specifying key PFs in the protocol based on prior judgement, there is often insufficient prior knowledge to ascertain all important PFs before a trial commences [13, 21].

Sample size of RCTs plays a critical role in balancing known and unknown PFs between treatment groups. Although many clinical trials with a binary outcome employ power calculations to determine an adequate sample size, underpowered studies are common [22-24]. Among 519 PubMed-indexed RCTs published in December 2000, the median total sample size per trial was 52 (10th – 90th percentile: 12–310) considering all designs and 80 (10th – 90th percentile: 25–369) considering only parallel-group trials [25]. A more recent systematic review of 215 two-arm parallel group RCTS of superiority with a single primary outcome published in six high impact factor general medical journals between January 1, 2005 and December 31, 2006 indicates a larger median total trial size of 425 (interquartile range: 158-1041) [26]. Sample size calculations often assume a balance of prognosis between the treatment groups regardless of sample size, yet the distribution of the possible unobserved PFs can be difficult to examine using empirical data mainly because they are unobserved.

The current simulation study was designed to address three objectives: (1) to evaluate the probability of imbalance in a binary PF between two treatment groups in simple RCTs with standard randomization (without stratification, blocking or minimization) evaluating a binary outcome; (2) to investigate the impact of prognostic imbalance on the estimation of treatment effect; and (3) to examine the effect of sample size on the probability and impact of prognostic imbalance in RCTs.

Methods

Simulation framework

We considered parallel group RCTs with a binary outcome in which equal numbers of patients were randomized to the treatment and control groups. For simplicity, we confined our attention to only one baseline PF without stratification. Five trial design parameters were considered: the frequency of the outcome event in the control group; the effect of treatment on the outcome; the strength of the association between the PF and the outcome; the prevalence of the PF; and the sample size.

We explored two simulation settings. For setting #1, we did not impose any level of imbalance, but simply generated a binary PF (C=0, 1) independently from the treatment allocation (T=0, 1) for each simulated trial. We refer to this as the "unconditional setting". This setting allowed us to evaluate the cumulative probability of prognostic imbalance greater than or equal to some level, and address whether or not adjusting for a baseline PF that is subject to chance imbalance improves the accuracy, precision and efficiency of the estimation of treatment effects.

We refer to setting #2 as the "conditional setting" for which we imposed a particular level of imbalance in each simulated trial, specifically, 5% more patients in the control group having the PF than those in the treatment group. Although, over a large number of RCTs, the probability of repeated occurrence of imbalance approaches zero, the conditional setting allowed us to explore what would happen if there were a 5% imbalance in a particular trial. This provided a way to assess the magnitude of potential bias resulting from an imbalance if it was unobserved or omitted from the analysis. It also allowed us to study whether this potential bias could be controlled by increasing the sample size.

Setting #1: the unconditional setting

Each simulated dataset in the unconditional setting consisted of a binary indicator for treatment allocation (T = 0, 1), a binary baseline PF (C = 0, 1), and a binomial response variable (Y), indicating the number of patients who experience an outcome event (D = 0, 1) for each T-C categorization. We related the log odds of experiencing the outcome D =1 conditional on the allocated treatment and baseline prognosis through the following model:

Simulation model: $$\log\frac{\Pr(D=1|T,C)}{1-\Pr(D=1|T,C)}=\beta_0+\beta_1T+\beta_2C \quad \text{(Eq. 1)}$$,

where β_0 corresponds to the log odds of the outcome among patients without the PF in the control group, β_1 corresponds to the log odds ratio (OR) of having the outcome in the experimental treatment group relative to the control group conditional on baseline prognosis (i.e. the treatment effect), and β_2 corresponds to the log OR of the outcome among patients having the PF versus not conditional on treatment status.

We assumed equal numbers of patients being randomized to the experimental group (T=1) and control group (T=0), i.e. $n_1=n_0=n$. We sampled C independently of T from the binomial distribution Bin(n_i, λ), with prevalence λ being fixed at 14 values between 0.005 and 0.995, namely, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.995. We simulated 8 scenarios (Table 1) by varying each of three parameters to reflect typical features of a cardiovascular prevention trial: (a) risk of outcome event in the control group (a low risk of 0.05; and a moderate risk of 0.10), (b) treatment effect size in the absence of the PF (a moderate effect: relative risk [RR] of 0.75; and a zero effect: RR of 1), and (c) effect of the PF on the outcome in the control group (a strong effect: RR of 5; and a moderate effect: RR of 2). If the covariates are strongly predictive of the outcome, i.e. strong PFs, mild or moderate imbalance can result in a biased effect estimate [3, 27]. The potential impact of dissimilarity in such strong PFs between groups can plausibly be greater when the risk of event in the control group is low, because results of hypothesis testing may be more sensitive to the change in the numbers of outcome events in treatment groups when the outcomes are rare. For each scenario, we investigated six sample sizes and the 14 λ values listed above. Considering a clinical trial aiming to detect a moderate treatment effect (i.e. RR=0.75) and a moderate risk of the outcome in the control group (i.e. 0.10), a standard power calculation suggests a total of 4000 patients (2000 per group) is needed to yield type I and type II error rates of 5% and 20%, respectively. To assess the impact of sample size on prognostic imbalance, we also included ½, ¼ and 1/16 of this calculated sample size for each simulation scenario (corresponding to 1000, 500 and 125 patients per arm, respectively). We also considered two smaller sample sizes (25 and 50 patients per arm) because small trials occur frequently in medical publications [25]. We simulated 10,000 trials per prevalence per sample size per scenario.

Setting #2: the conditional setting

We also simulated 10,000 replicates for each combination of the prevalence and sample size per scenario as per Table 1 in the conditional setting. For each trial, 5% more patients had the PF in the control group than the treatment group. We fixed the overall proportion of the PF at each of the 11 values: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95, as the probability of observing a 5% imbalance is extremely low for $\lambda < 0.05$ or > 0.95. We conducted all simulations and analyses in ***R*** 2.12.1.

Analysis

Distribution of imbalance

For each scenario, we retained the simulated proportion of patients with the PF per arm, with continuity correction by adding 0.5 to each T-C categorization to handle sparse cells [28]. We quantified imbalance using two different measures: the absolute difference (D_1) and the standardized difference (D_2), as follows:

- $D_1 = \left| p_1^c - p_0^c \right|$, and
- $D_2 = \left| p_1^c - p_0^c \right| \Big/ \sqrt{0.5 p_1^c \left(1 - p_1^c\right) + 0.5 p_0^c \left(1 - p_0^c\right)}$,

where p^c_i = proportion of patients having the PF (C=1) (with continuity correction) at baseline in the treatment (i = 1) or control (i = 0) group. We decided to use the absolute difference, D_1, as the primary measure of imbalance, because it is more intuitive for clinicians. The standardized measure has been advocated for having better statistical properties and may be more appealing to the statistical audience [29, 30]. We assessed the probabilities of observing different levels of imbalance for each sample size: Pr ($D_i \geq d_l$), where d_l = 0.005, 0.01, 0.025, 0.05, 0.10, 0.15, 0.20.

Impact of prognostic imbalance on treatment effect estimation

We fit two logistic regression models to evaluate the effect of treatment with and without adjustment for the PF. The adjusted model was the same as the underlying simulation model (Eq. 1) and the unadjusted model took the form of Eq. 2.

Unadjusted Model: $$\log \frac{\Pr(D=1|T)}{1-\Pr(D=1|T)} = \alpha_0 + \alpha_1 T, \qquad \text{(Eq. 2)}$$

where α_0 represents the log odds of having the outcome among patients in the control group (with or without the PF), and α_1 represents the log OR of the outcome in the experimental treatment group relative to the control group regardless of baseline prognosis.

For each simulated RCT, we recorded the estimated regression coefficients, their associated estimated standard errors (SEs), 95% confidence intervals (CIs, based on Wald test), and fitted probabilities of the outcome for each T-C (for the adjusted model) or C (for the unadjusted model) categorization. For each scenario, bias of the estimated regression coefficient ($\hat{\alpha}_1$ or $\hat{\beta}_1$) relative to the true log OR (β_1), its empirical standard deviation (SD), and mean squared error (MSE) were recorded for each model. The empirical coverage of the 95% CI was computed as the proportion of CIs that contained the true effect; and power was calculated as the proportion of replications where the CI excluded the null.

Results

Distribution of imbalance

Figure 1 displays the cumulative probabilities of an imbalance using the absolute measure (D_1) with 25, 50, 125, 500, 1000, and 2000 patients per arm. For a fixed sample size, the probability of imbalance varied with the prevalence of the PF (λ): imbalance was more likely to occur when λ is close to 0.5, but probability diminished as λ approached 0 or 1. The probability of imbalance increased markedly as sample size decreased regardless of λ. For a PF with prevalence of 0.5, the probability of an imbalance $\geq$ 5% was about 0.02 with 1000 patients per arm, 0.1 with 500 patients per arm, and 0.42, 0.62 and 0.67 with 125, 50 and 25 patients per arm, respectively. When the prevalence of PF was 0.05, $\Pr(D_1 \geq 0.05)$ increased from ≤ 0.0001, 0.0004, 0.059, 0.24 and 0.29 as sample size decreased from 1000 to 25.

Figure S1 displays the cumulative probability of imbalance using the standardized measure (D_2). Because the absolute difference was scaled by the pooled SD to create the standardized measure, λ had little impact on the probability of D_2, except for the extreme values. Common for both imbalance measures, the chance of imbalance decreased with increasing sample size. However, the relationship between the probability of imbalance and the prevalence of the PF differed using different measures.

Impact of prognostic imbalance on treatment effect estimation

<u>Setup #1: the unconditional setting</u>

Scenario 1 corresponded to trials with a 10% risk of the outcome in the control group, a strong prognostic factor (RR=5), and a moderate treatment effect (RR=0.75, corresponding OR = 0.73) (Table 1). Figures 2 and 3 depict the bias and empirical SD of the point estimator of log OR, the coverage of the 95% CI, and the empirical statistical power for the adjusted and unadjusted models with 125 and 2000 patients per arm. When PF was omitted from the logistic regression, the estimated log OR was biased towards zero.

The magnitude of bias declined as λ approached 0 or 1, but varied little with sample size when each arm had 50 or more patients. The adjusted estimator $\hat{\beta}_1$ was unbiased conditional on baseline prognosis, and independent of λ and sample size, when there were over 50 patients per arm. With 25 patients per arm, estimated log ORs from both models tended to be biased towards zero for $\lambda \leq 0.1$; the adjusted estimates were slightly negatively biased for greater λ values. This was possibly due to the lack of outcome events to reliably estimate the treatment contrast (Figures S2 and S3).

Adjusting for the PF reduced precision of the point estimator, especially when the trial size was less than 500 per arm. The adjusted model was able to maintain the nominal coverage of the 95% CI for different trial sizes. In contrast, coverage of the unadjusted model was less than the nominal value for most λ values, when sample size exceeded 500 per arm; the decline was more drastic when λ was near 0.5. For a PF with prevalence of 0.5, the actual coverage of the unadjusted 95% CIs was 95%, 93.58%, 91.15%, and 88.12% with 25-125, 500, 1000 and 2000 patients per arm, respectively.

When λ decreased to 0.05, coverage of the unadjusted 95% CIs was roughly around the nominal value. Despite a slight loss of precision, the adjusted model had equal or greater statistical power across of the spectrum of the prevalence of PF. The gain in power was more marked when sample size was between 500 and 1000 per arm (Figures S4 and S5), probably due to the floor or ceiling effect associated with very small or large sample sizes, i.e. power from both models approached 0 or 100%, so the difference in power between models shrank accordingly.

For a PF with prevalence of 0.5, the loss of power of the unadjusted model relative to the adjusted model was 3.44%, 15.20%, 11.29%, 14.39%, 9.66%, and 2.61% with 25, 50, 125, 500, 1000 and 2000 patients per arm. The two models achieved similar power for a rare PF with $\lambda < 0.1$. For both models, the precision of point estimator and empirical power increased with the number of outcome events resulting from increasing sample size and λ.

As the relative risk of experiencing an outcome event for those with the PF versus those without in the control group reduced from 5 to 2 (scenario 2), bias associated with the unadjusted point estimator of log OR became negligible for all trial sizes (except for 25 per arm with $\lambda \leq 0.2$). The adjusted and unadjusted models were also similar in terms of precision, coverage of CI and statistical power (Figures S6, S7, S8, and S9).

When the treatment had no effect on the outcome of interest (scenarios 3 and 4), the adjusted and unadjusted models produced unbiased point estimate despite the predictive power of the PF. Adjusting for baseline PF was not necessary in this situation to remove bias, and in fact it led to a slight inflation of SD. Sample size had little impact on the comparative performance of the two models, and nominal coverage of CI was achieved for both models (Figure S10).

For scenarios 5-8, where there was 5% risk of the outcome in the control group, the results demonstrated patterns similar to those described above for the first four scenarios. Precision of the point estimates and statistical power were lower for both models in scenarios 5-8. The magnitude of bias of the unadjusted log OR estimator in scenario 5 was slightly less than those in scenario 1. Differences in statistical power between two models also decreased slightly with the risk of outcome event when a treatment difference truly existed.

Figure 4 and Figures S11 and S12 display distributions of the differences (D_{ORR}) between the estimator of OR reduction (ORR, defined as 1 - OR) and the true ORR, i.e. $\mathrm{D_{ORR}} = \mathrm{O\hat{R}R} - \mathrm{ORR}$, across the spectrum of λ, based on 10,000 trials in scenario 1, with

125 patients per arm. The vertical axis represents the proportion of trials associated with a difference greater than or equal to a certain value d_2, where $d_2 = 0$, 0.05, 0.10, 0.15, 0.2 or 0.25. While Figure 4 corresponds to the probability of deviations in either direction, $\Pr(|D_{ORR}| \le d_2)$, Figures S11 and S12 correspond specifically to underestimation, $\Pr(D_{ORR} \le -d_2)$ and overestimation, $\Pr(D_{ORR} \ge d_2)$, respectively. Figure 5 and Figures S13 and S14 present distributions of D_{ORR} for the same scenario with 2000 per arm. Tables 2 and 3 present the proportions of difference at selected λ values across all sample sizes in scenario 1.

Overall, the proportion of random deviations decreased when the sample size, λ and the size of the deviation increased. When λ = 0.05 in scenario 1, the probabilities of $D_{ORR} \ge 0.05$ (in either direction) from the true ORR was 0.87-0.88 and 0.52 with 125 and 2000 patients per arm, respectively, for both models (Table 2). In comparison, the probabilities of $D_{ORR} \ge 0.1$ dropped to 0.75-0.76 (125/arm) and 0.20 (2000/arm) at the same prevalence (Table 3). When the treatment effect was zero, the corresponding probabilities of a given deviation were higher (Tables 4 and 5). For instance, probabilities of $D_{ORR} \ge 0.1$ were 0.78-0.81 (125/arm) and 0.30-0.32 (2000/arm) when λ = 0.05 in scenario 3. The probabilities of $D_{ORR} \ge 0.1$ remained above 0.8 with 50 or less patients per arm in all scenarios, when λ was between 0.01 and 0.5.

In scenario 1, the distribution of the unadjusted ORR estimates was positively skewed, indicating a higher likelihood of underestimation than overestimation when PF was a strong predictor of the outcome and treatment was moderately efficacious.

Adjusting for PF made the distribution of the ORR estimator symmetric around the true effect, i.e. random fluctuations were equally likely in either direction. When the influence of the PF was moderate or the actual treatment effect was zero, adjusting for PF did not improve accuracy or precision of the estimate.

Setup #2: the conditional setting

For all 8 scenarios in the conditional setting, the adjusted model produced roughly unbiased estimates of the treatment effect and maintained nominal coverage of the 95% CI. The unadjusted model overestimated treatment effects, and the model performance was influenced by multiple factors including the treatment effect, the effect and prevalence of the PF, and the sample size.

Figures S15 and S16 display the performances of the adjusted and unadjusted models in scenario 1 under the conditional setting with 125 and 2000 patients per arm. Ignoring the fact that 5% more patients had this PF in the control arm led to substantial overestimation of treatment effect. Bias was comparatively larger when PF was rare: when λ ranged between 0.05 and 0.2, bias of the unadjusted estimate of log OR, $\hat{\alpha}_1$, relative to β_1 was between -0.18 and -0.09, with 125 per arm in scenario 1. Varying sample size led to little change in the magnitude of bias in scenario 1, though estimates were more variable with 125 or fewer patients per arm. Coverage of the unadjusted CI was greater than its nominal value with 125 or fewer per arm for most prevalence values between 0 and 1. The coverage reduced substantially as sample size went beyond 1000 per arm; when λ ≤ 0.2 or ≥ 0.8 coverage of the unadjusted CI dropped to 60%-90%. For a fixed sample size, the unadjusted estimate had slightly greater precision than the adjusted estimate; but the difference diminished as sample size increased.

With PF RR=2 in scenario 2, bias of the unadjusted point estimator decreased with sample size and varied little with λ. The average biases of $\hat{\alpha}_1$ over the 11 prevalence values investigated were -0.014, -0.055 and -0.050 with 25, 50 and 125 patients per arm respectively and reduced to -0.015, -0.007 and -0.004 when the sample size reached 500, 1000 and 2000 per arm. The corresponding biases of the adjusted log OR estimator, $\hat{\beta}_1$, were 0.024, -0.024, -0.012, -0.005, -0.002 and -0.001. Both models achieved similar coverage when sample sizes were greater than or equal to 50 per arm, and demonstrated comparable precision.

The unadjusted model had slightly greater power though this advantage decreased as sample size increased. When the treatment had no effect, performance of the adjusted and unadjusted models in scenarios 3 and 4 was similar to that in scenario 2. Omitting a stronger PF in analysis again led to a greater bias for a fixed sample size and bias again shrank as trial size enlarged. Findings similar to scenarios 1-4 were demonstrated when the risk of outcome events in the control group reduced from 0.1 to 0.05 (scenarios 5-8). A low event rate in each trial resulted in reduced precision and statistical power. Bias of the unadjusted log OR estimates decreased with sample size and was generally smaller than the counterpart in the previous scenarios.

Discussion

Our simulation results demonstrate that small sample size is associated with a high risk of imbalance in PFs in individual simple RCTs. The probabilities of an absolute imbalance ≥ 5% in a binary PF of prevalence 0.5 is 0.42, 0.62 and 0.67 with 125, 50 and 25 patients per arm. The probability of absolute imbalance decreases as sample size increases or prevalence of PF approaches 0 or 1.

Failing to adjust for a largely balanced strong PF (RR=5) in a logistic regression model leads to bias toward no treatment effect when the actual size of treatment effect is moderate (RR=0.75); this bias varies little with sample size greater than 50 patients per arm. Adjusting for such a PF reduces precision of the effect estimate but increases statistical power. The gain in power is comparatively larger when sample size is between 500 and 1000 per arm and prevalence is within 0.2 – 0.6, relative to other cases. When the PF is less powerful and a treatment difference exists, improvement in accuracy and efficiency associated with the adjustment for a largely balanced PF is less noticeable. When the treatment effect is zero, such covariate adjustment leads to minimal loss of precision. Overall the simulation results based on a single binary baseline PF suggest it is critical to adjust for important PFs in trials evaluating a binary outcome. If ignored, substantial bias due to confounding or non-collapsibility can emerge; bias would be more marked when PF has high predictive value and sample size is small to moderate.

It is challenging to establish a single rule for sample size requirement focused on the probability and impact of prognostic imbalance. Multiple factors influence the requirement.

Firstly, sample size should be sufficiently large that the probability of imbalance is restricted to a reasonably low value. The adequate sample size varies with the choice of imbalance measure, the size of imbalance that is deemed important, and the prevalence of the PF. For example, Figure 1 suggests that if an absolute measure of imbalance ≥0.05 is deemed important, 1000 patients per arm is a reasonable size.

Secondly, sample size should be sufficient to produce a reliable estimate of treatment effect. Although it has less impact on the magnitude of bias around the mean effect estimate in the unconditional setting, sample size does affect precision. While adjusting for PF removes systematic bias, estimates from an individual trial may still deviate from the true effect in either direction due to random sampling variation. Tables 2 and 4 suggest that probabilities of having an absolute deviation >= 0.05 (in either direction) from the true ORR are 0.87-0.93 and 0.52-0.62 for trials recruiting 125 and 2000 patients per arm, respectively. If trialists are willing to tolerate a slightly bigger deviation from the true ORR, for instance, no more than 0.1, the above probabilities decrease to 0.75-0.81 (125/arm) and 0.20-0.32 (2000/arm) for both models, and 2000 patients per arm then seems to be a reasonable sample size (Tables 3 and 5). As PF becomes less prevalent, larger trial sizes are required for purposes of precision. When randomization partially or completely fails, no statistical adjustment or increase in sample size can fully correct the resulting bias.

The current investigation on the likelihood of prognostic imbalance and its implications for sample size requirements is consistent with previous findings. A minimum of 100 patients per arm has been suggested to control the chance of imbalance of 20% or more in a single PF [31], and 1350 per arm may be needed to minimize the chance of a 5% imbalance [3]. Although Cui et al calculated the probabilities of a 20% imbalance in at least one out of *k* independent PFs (*k* = 2, 3, and 4) [31], situations involving multiple correlated PFs are worth further investigation.

Gail first demonstrated that omitting balanced baseline covariates in logistic regression asymptotically (i.e. for very large sample sizes) results in downward bias on the subject-specific treatment-outcome association [14]. This is also referred to as the non-collapsibility problem [16], because the odds ratio as the measure of association between the treatment and the binary outcome within each category of the baseline covariates (i.e. conditional or subject-specific association) is different from the association across all categories of the covariates (i.e. the marginal or average association).

In their simulation study [32], Negassa and Hanley showed that omitting an important balanced continuous or binary covariate in logistic regression model lowers both the coverage probability (that is, the proportion of the time that the CI contains the true value of interest in a set of hypothetical repetition of data collection and analysis procedure [33]) and study power in binary trials with moderate sample sizes (n=500 and 1000). These findings are complemented by a simulation study that explored the effect of imbalance in two continuous baseline covariates on power in a logistic regression framework when both variables were adjusted for in analyzing small trials (n = 50, 100 and 300) [12]. Others quantified the increase in statistical power resulting from covariate adjustment as a decrease in the sample size required in comparison to the unadjusted model [11].

It was not clear in the literature, however, how the interplay of chance imbalance, the risk of outcome and the prevalence of a binary PF affects treatment effect estimation in trials with a binary outcome. Our simulation study provided information on what constitute an adequate sample size to control against potential impact of prognostic imbalance. Our results based on trials subject to chance imbalance across six sample sizes in the unconditional setting are consistent with the previous findings.

When one is confident that all important PFs are distributed similarly between treatment groups in a binary trial, it is sensible to decide if the goal of a trial evaluating a binary outcome is to assess the marginal effect of treatment over patients with heterogeneous baseline prognosis, or to obtain a more individualized treatment effect estimate that is specific to a prognosis. These objectives can be achieved by using the unadjusted and adjusted logistic regression analyses. With a binary outcome, the two models produce mathematically different results in the presence of a non-zero treatment effect. Mismatch of the study objective, the statistical method, and interpretation of results can result in misleading messages. Due to the uncertainty around the existence or magnitude of the treatment effect and possibly different criteria to assess prognostic imbalance, we recommend reporting both the adjusted and unadjusted results in the manuscript.

The CPMP guideline recommends that including important PFs in the primary analysis can be justified only if their associations with the primary outcome are expected to be strong, based on previous evidence, and are specified a priori [18]. What constitutes adequate justification may be a matter of judgment. Our results demonstrate the value of adjustment, and suggest the merits of avoiding excessively stringent criteria when deciding whether prior evidence of prognostic power is adequate.

Our study has several limitations. First, we included only one binary baseline PF to illustrate the probability and impact of prognostic imbalance in RCTs evaluating a binary outcome. For continuous PFs, Ciolino and colleagues proposed a rank-sum ratio to measure the level of imbalance in addition to the commonly used mean values [12]. When multiple PFs are present at baseline, balancing distribution of the individual PFs and the overall prognosis needs to be assessed. Although the single binary PF considered in the current study can be conceptualized as a measure of the overall prognosis of a patient based on multiple PFs, for instance, in a propensity score framework [34], further investigation on the distribution and impact of multiple correlated PFs on effect estimation in RCTs is warranted.

Second, although our investigation was focused on prognostic balancing in individual RCTs, systematic reviews and meta-analyses face the same methodological challenges. The cumulative number of patients from individual RCTs and the between-study variation need to be considered to assess the impact of imbalance on obtaining an aggregated estimate of treatment effects. Future work is needed in these directions.

Our study provides useful new insights. The results can not only help to design clinical trials, but can also inform quality assessment of a body of evidence from RCTs. Our simulation findings provide insights on prognostic imbalance which pertains to both risk of bias and imprecision [35]. The current study was not designed to propose a single threshold value of sample size that can be readily employed to rate the quality of evidence with respect to precision. Rather it lends itself to guide selection of such threshold values over various combinations of trial parameters, a subjective process likely influenced by the tolerance of risk.

In summary, prognostic imbalance does not on average jeopardize internal validity of findings from RCTs, but if neglected, may lead to chance confounding and biased estimate of treatment effect in a single RCT. To produce an accurate estimate of the treatment-outcome relationship conditional on patients' baseline prognosis, balanced or unbalanced PFs with high predictive value should be adjusted for in the analysis. Covariate adjustment slightly reduces precision, but improves study efficiency, when PFs are largely balanced. Once chance imbalance in baseline prognosis is observed, covariate adjustment should be performed to remove chance confounding.

Acknowledgments

We thank the anonymous reviewers for their constructive comments that helped to strengthen the manuscript. Rong Chu holds a doctoral research award from the Canadian

Institutes of Health Research. This work was partially supported by a research grant from the Canadian Network and Centre for Trials Internationally (CANNeCTIN).

References

1. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, et al. (2000) Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. JAMA 284:1290-1296.

2. Senn S (1994) Testing for baseline balance in clinical trials. Stat Med 13:1715-1726.

3. Wang SJ, O'Neill RT, Hung HJ. (2010) Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. Clin Trials 7:525-536.

4. Brower RG, Lanken PN, MacIntyre N, Matthay MA, Morris A, et al. (2004) Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. N Engl J Med 351:327-336.

5. Meade MO, Cook DJ, Guyatt GH, Slutsky AS, Arabi YM, et al. (2008) Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive end-expiratory pressure for acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. JAMA 299:637-645.

6. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995) Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 333:1581-1587.

7. Demchuk AM, Tanne D, Hill MD, Kasner SE, Hanson S, et al. (2001) Predictors of good outcome after intravenous tPA for acute ischemic stroke. Neurology 57:474-480.

8. Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression models. International Statistical Review 58:227-240.

9. Lingsma H, Roozenbeek B, Steyerberg E, IMPACT investigators (2010) Covariate adjustment increases statistical power in randomized controlled trials. J Clin Epidemiol 63:1391; author reply 1392-1393.

10. Hauck WW, Anderson S, Marcus SM (1998) Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 19:249-256.

11. Hernandez AV, Steyerberg EW, Habbema JD (2004) Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol 57:454-460.

12. Ciolino J, Zhao W, Martin R, Palesch Y (2011) Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. Contemp Clin Trials 32:250-259.

13. Yu LM, Chan AW, Hopewell S, Deeks JJ, Altman DG (2010) Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. Trials 11:59.

14. Gail MH, Weiand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71:431-444.

15. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S (1991) A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol 44:77-81.

16. Greenland S, Robins JM, Pearl J (1999) Confounding and Collapsibility in Causal Inference. Statistical Science 14:29-46.

17. International Conference on Harmonisation E9 Expert Working Group (1999) ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Stat Med 18:1905-1942.

18. Committee for Proprietary Medicinal Products (CPMP) (2004) Committee for Proprietary Medicinal Products (CPMP): points to consider on adjustment for baseline covariates. Stat Med 23:701-709.

19. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, et al. (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134:663-694.

20. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB (2010) A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol 63:142-153.

21. Pocock SJ, Assmann SE, Enos LE, Kasten LE (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 21:2917-2930.

22. Keen HI, Pile K, Hill CL (2005) The prevalence of underpowered randomized clinical trials in rheumatology. J Rheumatol 32:2083-2088.

23. Moher D, Dulberg CS, Wells GA (1994) Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 272:122-1244.

24. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I (2000) Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 320:1308-1311.

25. Chan AW, Altman DG (2005) Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 365:1159-1162.

26. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P (2009) Reporting of sample size calculation in randomised controlled trials: Review. BMJ 2009;338:b1732.

27. Zelen M (1974) The randomization and stratification of patients to clinical trials. J Chronic Dis 27:365-375.

28. Walter SD (1985) Small sample estimation of log odds ratios from logistic regression and fourfold tables. Stat Med 4:437-444.

29. Flury BK, Riedwyl H (1986) Standard distance in univariate and multivariate analysis. The American Statistician 40:249-251.

30. Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39:33-38.

31. Cui L, Hung HM, Wang SJ, Tsong Y (2002) Issues related to subgroup analysis in clinical trials. J Biopharm Stat 12:347-358.

32. Negassa A, Hanley JA (2007) The effect of omitted covariates on confidence interval and study power in binary outcome analysis: a simulation study. Contemp Clin Trials 28:242-248.

33. Dodge Y (2003) The Oxford Dictionary of Statistical Terms. 6th ed. New York: Oxford University Press.

34. Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79:516-524.

35. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A (2011) GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 64:380-382.

Figure 1 Probability of imbalance using absolute measure (D_1) with different trial sizes. Lines correspond to Pr ($D_1 \geq d_1$), where $d_1 = 0.005$ (hollow circle), 0.01 (triangle), 0.025 (cross), 0.05 (X), 0.10 (diamond), 0.15 (inverted triangle), and 0.20 (filled circle), from the top to the bottom, respectively. Top left: 25/arm, top right: 50/arm, middle left: 125/arm, middle right: 500/arm, bottom left: 1000/arm, bottom right: 2000/arm.

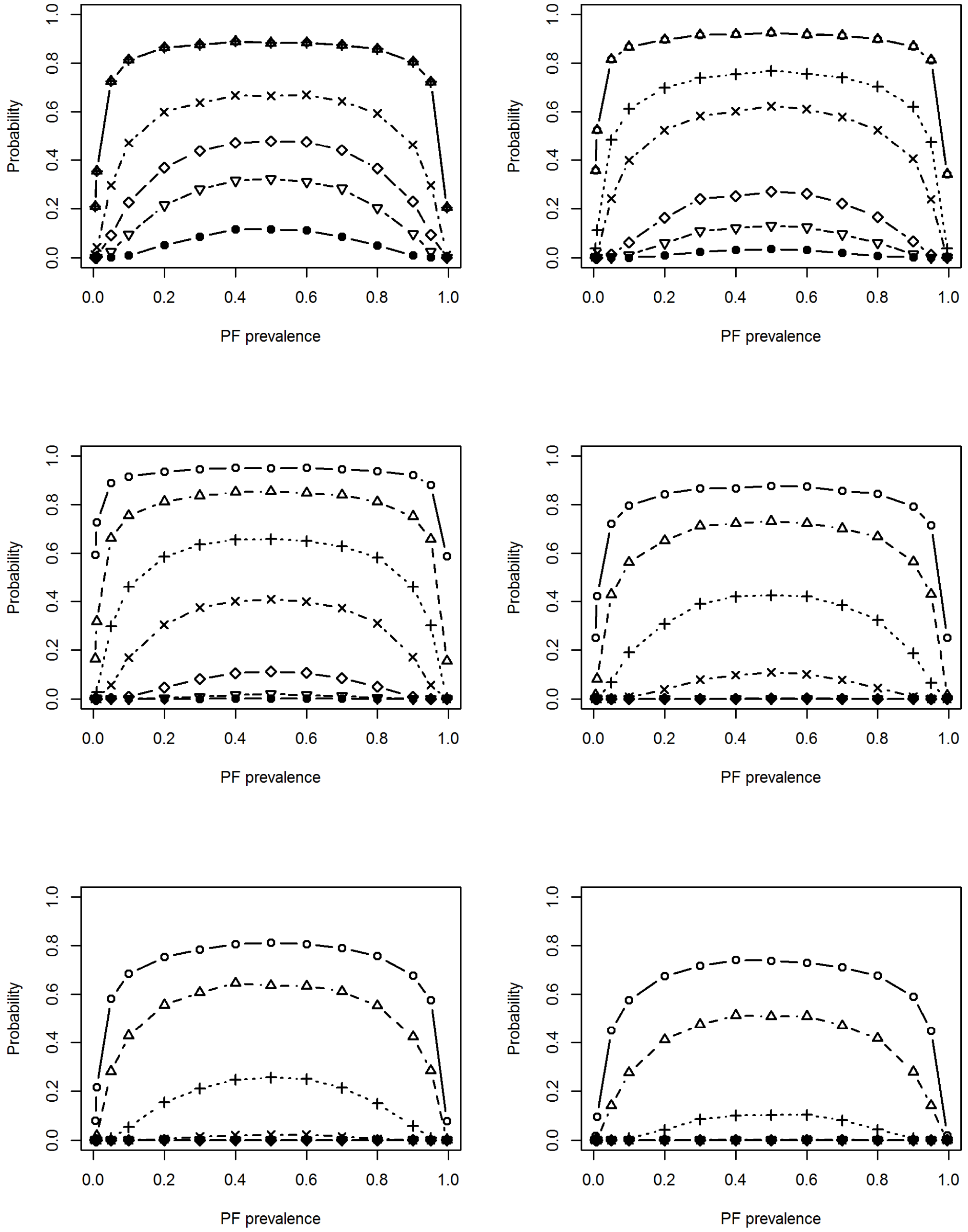


Figure 2 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 125 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

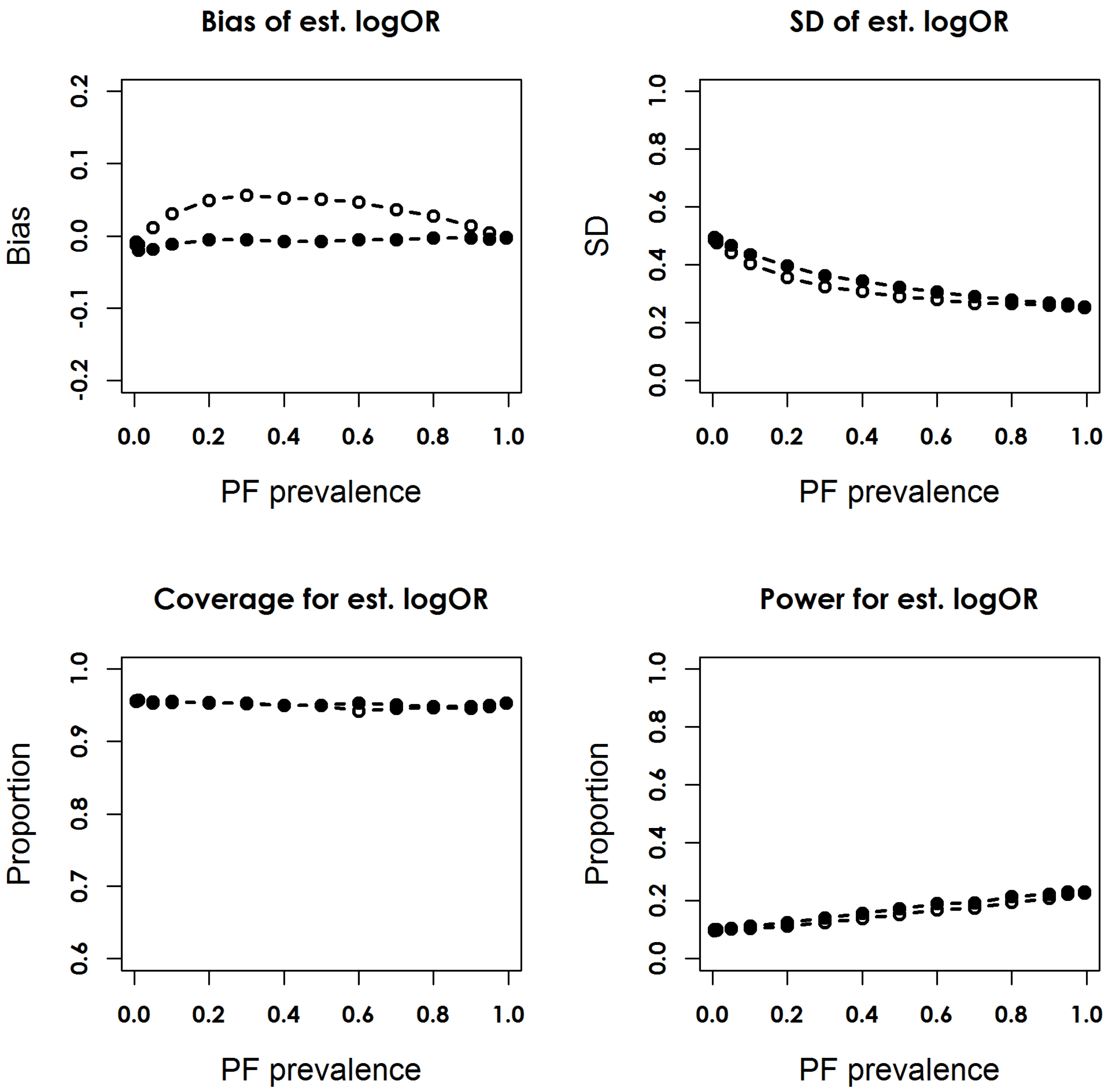


Figure 3 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 2000 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

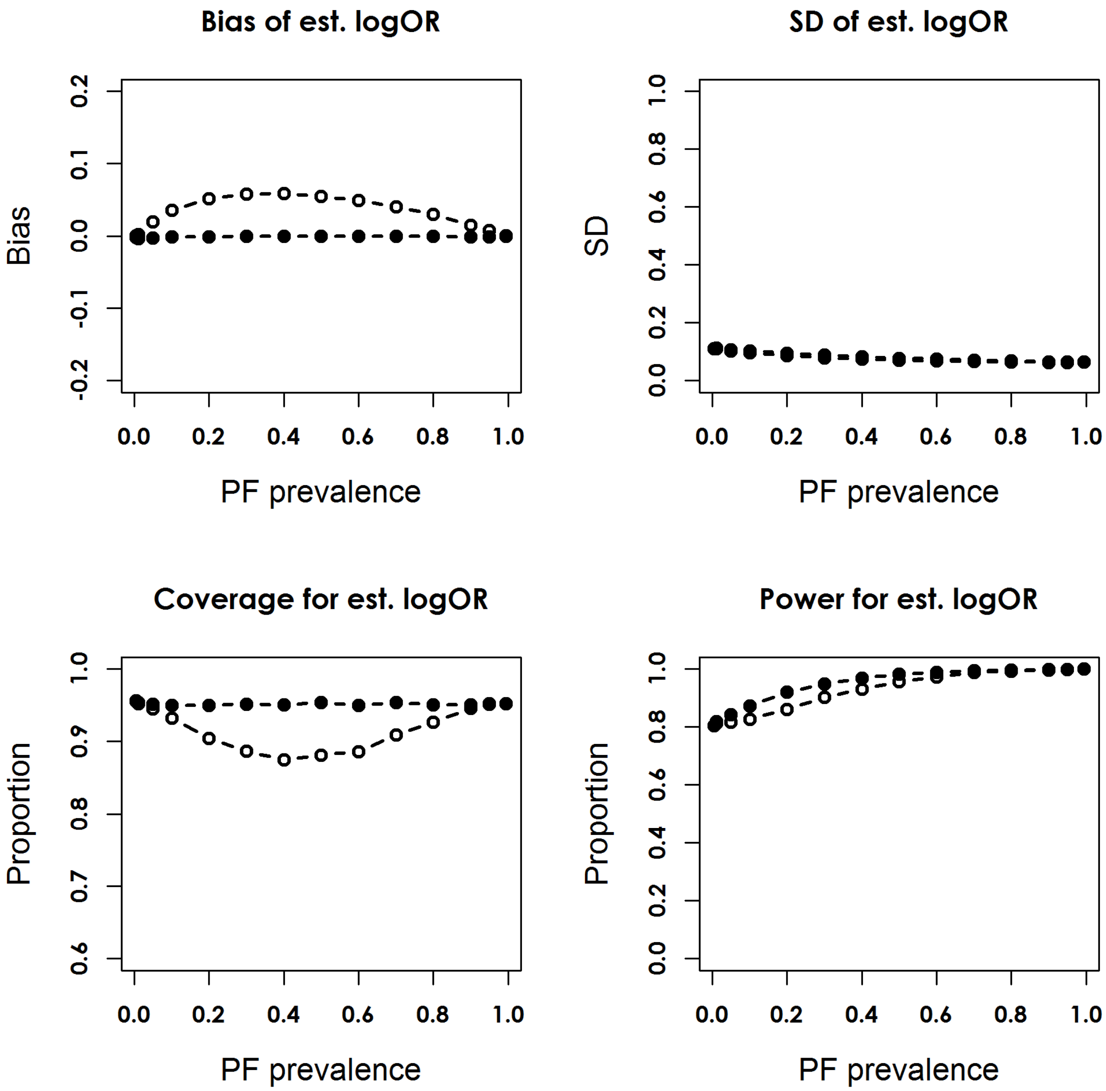


Figure 4 Probability of difference between the estimated and true ORR (deviation in either direction) in scenario 1, the unconditional setting, with 125 patients per arm. Within each graph, lines correspond to Pr ($|D_{ORR}| \geq d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

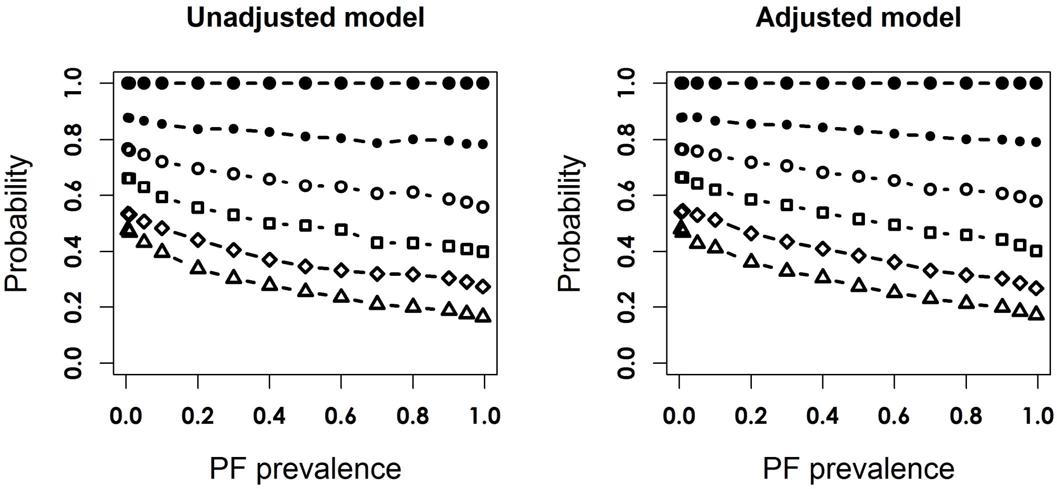


Figure 5 Probability of difference between the estimated and true ORR (deviation in either direction) in scenario 1, the unconditional setting, with 2000 patients per arm. Within each graph, lines correspond to Pr ($|D_{ORR}| \geq d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

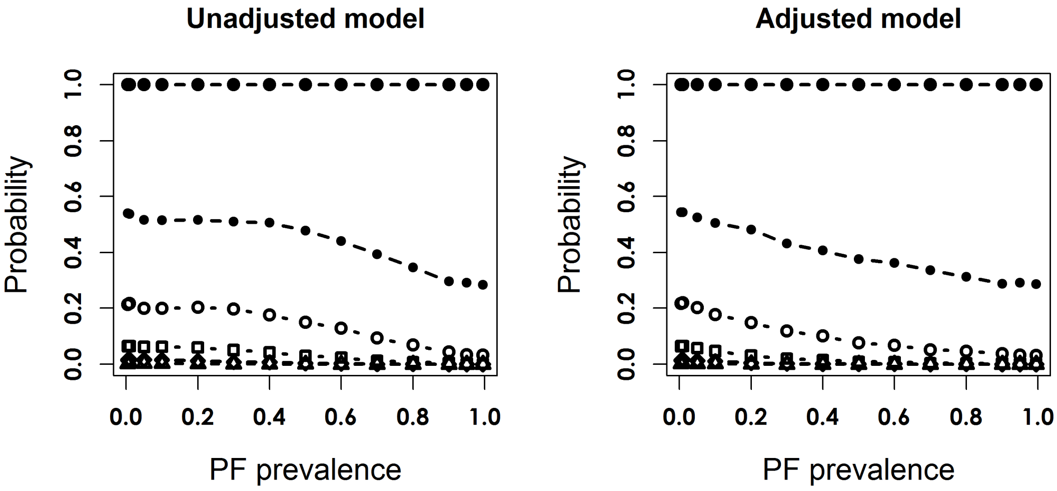


Table 1 Simulation scenarios for the unconditional and conditional settings

Scenario	**Effect of treatment in RR* (B1)**	**Effect of PF in RR† (B2)**	**Incidence of outcome (B0)**	**Prevalence of PF (C)**	**Sample size/arm**
1	0.75 (-0.315)	5 (2.197)	0.1 (-2.197)		
2	0.75 (-0.315)	2 (0.811)		0.005 – 0.995	(a) 25
3	1 (0)	5 (2.197)		(unconditional)	(b) 50
4	1 (0)	2 (0.811)			(c) 125
5	0.75 (-0.301)	5 (1.846)	0.05 (-2.944)	0.05 – 0.95	(d) 500
6	0.75 (-0.301)	2 (0.747)		(conditional)	(e) 1000
7	1 (0)	5 (1.846)			(f) 2000
8	1 (0)	2 (0.747)			

** Relative risk of having an outcome event for people receiving the experimental treatment (vs. control treatment) without the prognostic factor*

† Relative risk of having an outcome for people with vs. without the PF in the control group

SCN: Scenario

Table 2 Probability of difference between the estimated and true ORR ≥ 0.05 in the unconditional setting scenario 1

Difference from true ORR ≥ 0.05		**Unadjusted model**					**Adjusted model**				
	Sample size	**Prevalence of PF**					**Prevalence of PF**				
		0.01	**0.05**	**0.1**	**0.2**	**0.5**	**0.01**	**0.05**	**0.1**	**0.2**	**0.5**
Over-estimation	25	0.45	0.45	0.44	0.44	0.44	0.45	0.44	0.46	0.48	0.47
	50	0.46	0.44	0.43	0.42	0.39	0.47	0.47	0.47	0.47	0.45
	125	0.44	0.42	0.39	0.35	0.33	0.45	0.45	0.44	0.43	0.42
	500	0.37	0.33	0.29	0.24	0.20	0.38	0.38	0.37	0.36	0.33
	1000	0.32	0.27	0.21	0.16	0.11	0.33	0.33	0.31	0.29	0.26
	2000	0.25	0.19	0.14	0.08	0.04	0.27	0.26	0.25	0.23	0.18
Under-estimation	25	0.52	0.50	0.48	0.47	0.51	0.52	0.51	0.48	0.46	0.46
	50	0.47	0.49	0.49	0.47	0.49	0.46	0.46	0.46	0.44	0.45
	125	0.43	0.45	0.46	0.48	0.48	0.43	0.43	0.43	0.43	0.41
	500	0.39	0.41	0.44	0.46	0.47	0.39	0.37	0.37	0.36	0.33
	1000	0.34	0.37	0.42	0.46	0.46	0.33	0.32	0.32	0.31	0.28
	2000	0.28	0.33	0.38	0.44	0.44	0.27	0.26	0.26	0.25	0.20
Overall	25	0.96	0.94	0.92	0.91	0.95	0.97	0.95	0.94	0.94	0.93
	50	0.93	0.92	0.92	0.89	0.88	0.93	0.93	0.93	0.91	0.90
	125	0.88	0.87	0.86	0.84	0.81	0.88	0.88	0.87	0.86	0.83
	500	0.76	0.75	0.73	0.70	0.66	0.76	0.75	0.74	0.72	0.67
	1000	0.66	0.64	0.63	0.62	0.57	0.66	0.65	0.63	0.60	0.54
	2000	0.54	0.52	0.51	0.52	0.48	0.54	0.52	0.51	0.48	0.38

ORR: odds ratio reduction; PF: prognostic factor

Table 3 Probability of difference between the estimated and true ORR ≥ 0.10 in the unconditional setting scenario 1

Difference from true ORR ≥ 0.10		**Unadjusted model**					**Adjusted model**				
	Sample size	**Prevalence of PF**					**Prevalence of PF**				
		0.01	**0.05**	**0.1**	**0.2**	**0.5**	**0.01**	**0.05**	**0.1**	**0.2**	**0.5**
Over-estimation	25	0.37	0.37	0.38	0.40	0.38	0.39	0.41	0.43	0.44	0.44
	50	0.40	0.39	0.39	0.38	0.33	0.42	0.43	0.43	0.43	0.39
	125	0.38	0.35	0.32	0.28	0.24	0.39	0.38	0.37	0.35	0.33
	500	0.26	0.22	0.17	0.12	0.08	0.27	0.25	0.24	0.22	0.18
	1000	0.17	0.13	0.09	0.05	0.02	0.18	0.18	0.16	0.14	0.09
	2000	0.09	0.05	0.03	0.01	0.00	0.10	0.09	0.07	0.07	0.03
Under-estimation	25	0.51	0.49	0.47	0.43	0.42	0.52	0.49	0.46	0.43	0.42
	50	0.41	0.43	0.44	0.45	0.42	0.41	0.43	0.42	0.40	0.40
	125	0.39	0.40	0.41	0.42	0.39	0.38	0.38	0.38	0.37	0.34
	500	0.29	0.30	0.31	0.32	0.30	0.28	0.27	0.26	0.24	0.20
	1000	0.21	0.23	0.25	0.26	0.24	0.20	0.20	0.19	0.17	0.13
	2000	0.13	0.15	0.17	0.19	0.15	0.12	0.11	0.10	0.08	0.05
Overall	25	0.88	0.86	0.85	0.83	0.80	0.91	0.90	0.89	0.88	0.86
	50	0.81	0.81	0.83	0.82	0.75	0.83	0.85	0.85	0.83	0.79
	125	0.76	0.75	0.72	0.70	0.63	0.76	0.76	0.75	0.72	0.67
	500	0.55	0.52	0.49	0.45	0.38	0.55	0.52	0.50	0.47	0.39
	1000	0.38	0.36	0.34	0.32	0.26	0.38	0.37	0.35	0.30	0.22
	2000	0.22	0.20	0.20	0.20	0.15	0.22	0.20	0.18	0.15	0.08

ORR: odds ratio reduction; PF: prognostic factor

Table 4 Probability of difference between the estimated and true ORR ≥ 0.05 in the unconditional setting scenario 3

Difference from true ORR ≥ 0.05		**Unadjusted model**					**Adjusted model**				
	Sample size	**Prevalence of PF**					**Prevalence of PF**				
		0.01	**0.05**	**0.1**	**0.2**	**0.5**	**0.01**	**0.05**	**0.1**	**0.2**	**0.5**
Over-estimation	25	0.39	0.41	0.41	0.42	0.43	0.39	0.44	0.47	0.48	0.47
	50	0.44	0.44	0.44	0.45	0.46	0.44	0.47	0.47	0.47	0.46
	125	0.45	0.46	0.46	0.45	0.42	0.46	0.45	0.45	0.45	0.44
	500	0.41	0.41	0.39	0.39	0.36	0.33	0.32	0.31	0.30	0.27
	1000	0.36	0.35	0.35	0.33	0.31	0.37	0.36	0.36	0.35	0.32
	2000	0.31	0.30	0.29	0.27	0.23	0.31	0.30	0.30	0.29	0.26
Under-estimation	25	0.40	0.41	0.41	0.43	0.44	0.41	0.45	0.47	0.47	0.48
	50	0.43	0.43	0.45	0.45	0.45	0.43	0.46	0.47	0.47	0.46
	125	0.46	0.46	0.46	0.45	0.42	0.46	0.45	0.45	0.44	0.44
	500	0.40	0.40	0.39	0.38	0.35	0.41	0.40	0.40	0.39	0.37
	1000	0.37	0.36	0.35	0.34	0.31	0.37	0.37	0.35	0.35	0.33
	2000	0.32	0.31	0.30	0.28	0.24	0.32	0.32	0.31	0.29	0.26
Overall	25	0.79	0.81	0.82	0.85	0.87	0.80	0.89	0.93	0.95	0.95
	50	0.86	0.88	0.89	0.90	0.91	0.87	0.93	0.94	0.94	0.92
	125	0.92	0.93	0.92	0.91	0.83	0.92	0.90	0.90	0.89	0.88
	500	0.81	0.81	0.78	0.77	0.71	0.73	0.72	0.71	0.69	0.64
	1000	0.73	0.71	0.70	0.67	0.61	0.73	0.73	0.72	0.70	0.65
	2000	0.63	0.60	0.59	0.55	0.47	0.63	0.62	0.61	0.58	0.53

ORR: odds ratio reduction; PF: prognostic factor

Table 5 Probability of difference between the estimated and true ORR ≥ 0.10 in the unconditional setting scenario 3

Difference from true ORR ≥ 0.10		**Unadjusted model**					**Adjusted model**				
	Sample size	**Prevalence of PF**					**Prevalence of PF**				
		0.01	**0.05**	**0.1**	**0.2**	**0.5**	**0.01**	**0.05**	**0.1**	**0.2**	**0.5**
Over-estimation	25	0.39	0.41	0.41	0.42	0.43	0.39	0.42	0.45	0.46	0.44
	50	0.44	0.44	0.44	0.45	0.39	0.43	0.44	0.44	0.43	0.42
	125	0.38	0.39	0.39	0.38	0.36	0.39	0.40	0.39	0.39	0.37
	500	0.31	0.30	0.28	0.27	0.23	0.31	0.31	0.29	0.29	0.25
	1000	0.24	0.22	0.21	0.18	0.14	0.24	0.23	0.22	0.21	0.16
	2000	0.16	0.14	0.13	0.11	0.06	0.16	0.14	0.14	0.12	0.09
Under-estimation	25	0.40	0.41	0.41	0.43	0.44	0.40	0.43	0.45	0.45	0.46
	50	0.43	0.43	0.45	0.45	0.41	0.43	0.43	0.45	0.44	0.42
	125	0.41	0.39	0.39	0.39	0.37	0.42	0.41	0.40	0.39	0.38
	500	0.32	0.31	0.30	0.28	0.24	0.33	0.32	0.31	0.30	0.27
	1000	0.26	0.25	0.22	0.20	0.16	0.26	0.26	0.24	0.22	0.20
	2000	0.18	0.17	0.15	0.12	0.08	0.18	0.17	0.17	0.15	0.11
Overall	25	0.79	0.81	0.82	0.85	0.87	0.79	0.84	0.89	0.91	0.89
	50	0.86	0.88	0.89	0.89	0.80	0.86	0.88	0.88	0.87	0.84
	125	0.79	0.78	0.78	0.78	0.73	0.81	0.81	0.80	0.78	0.75
	500	0.63	0.61	0.58	0.54	0.47	0.64	0.63	0.60	0.59	0.52
	1000	0.49	0.47	0.43	0.38	0.30	0.50	0.49	0.47	0.43	0.36
	2000	0.34	0.30	0.27	0.23	0.15	0.34	0.32	0.30	0.27	0.20

ORR: odds ratio reduction; PF: prognostic factor

Supporting Information

Figure S1 Probability of imbalance using standardized measure (D_2) with different trial sizes. Lines correspond to Pr ($D_2 \geq d_1$), where d_1 = 0.005 (hollow circle), 0.01 (triangle), 0.025 (cross), 0.05 (X), 0.10 (diamond), 0.15 (inverted triangle), and 0.20 (filled circle), from the top to the bottom, respectively. Top left: 25/arm, top right: 50/arm, middle left: 125/arm, middle right: 500/arm, bottom left: 1000/arm, bottom right: 2000/arm.

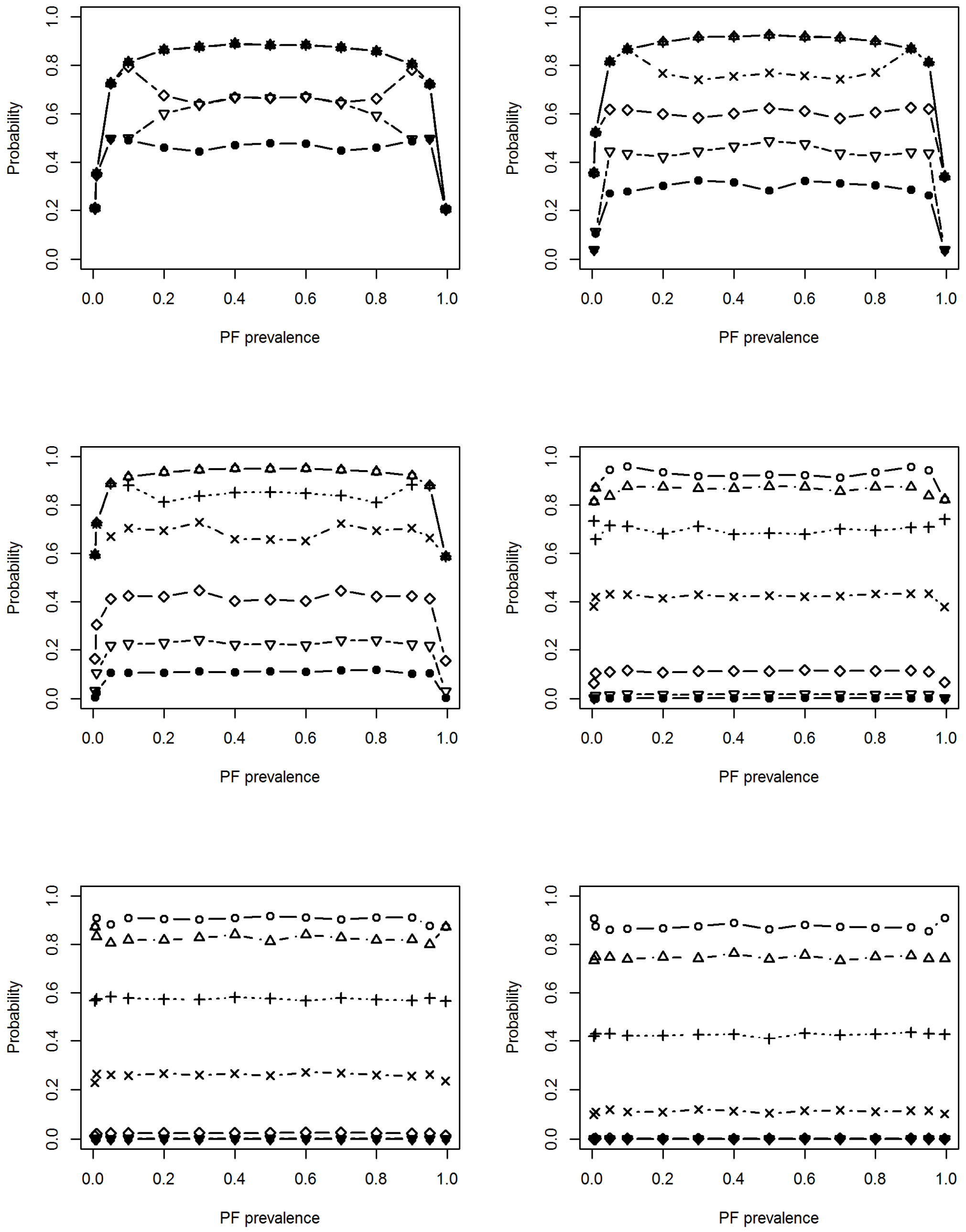


Figure S2 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 25 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

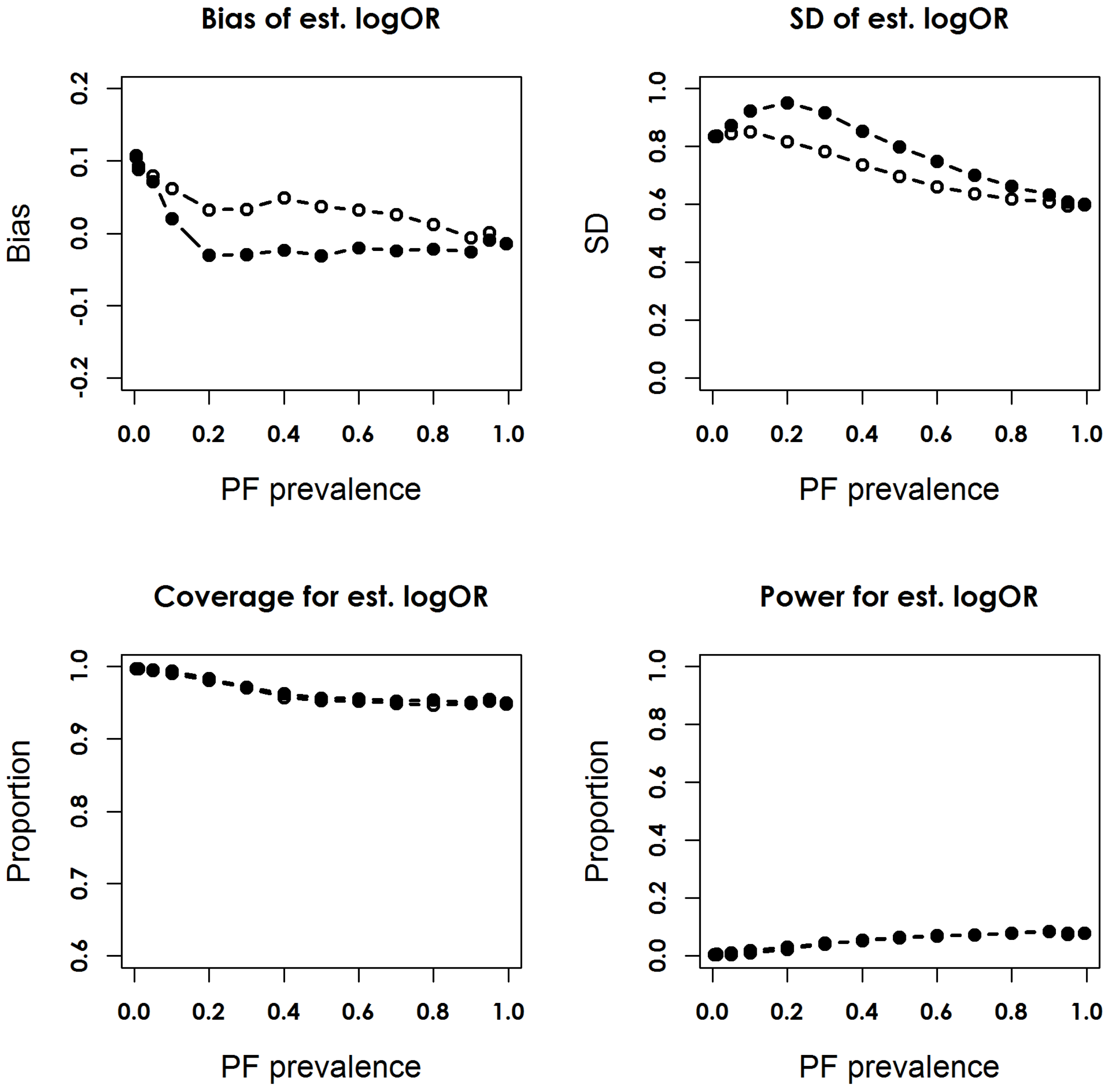


Figure S3 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 50 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

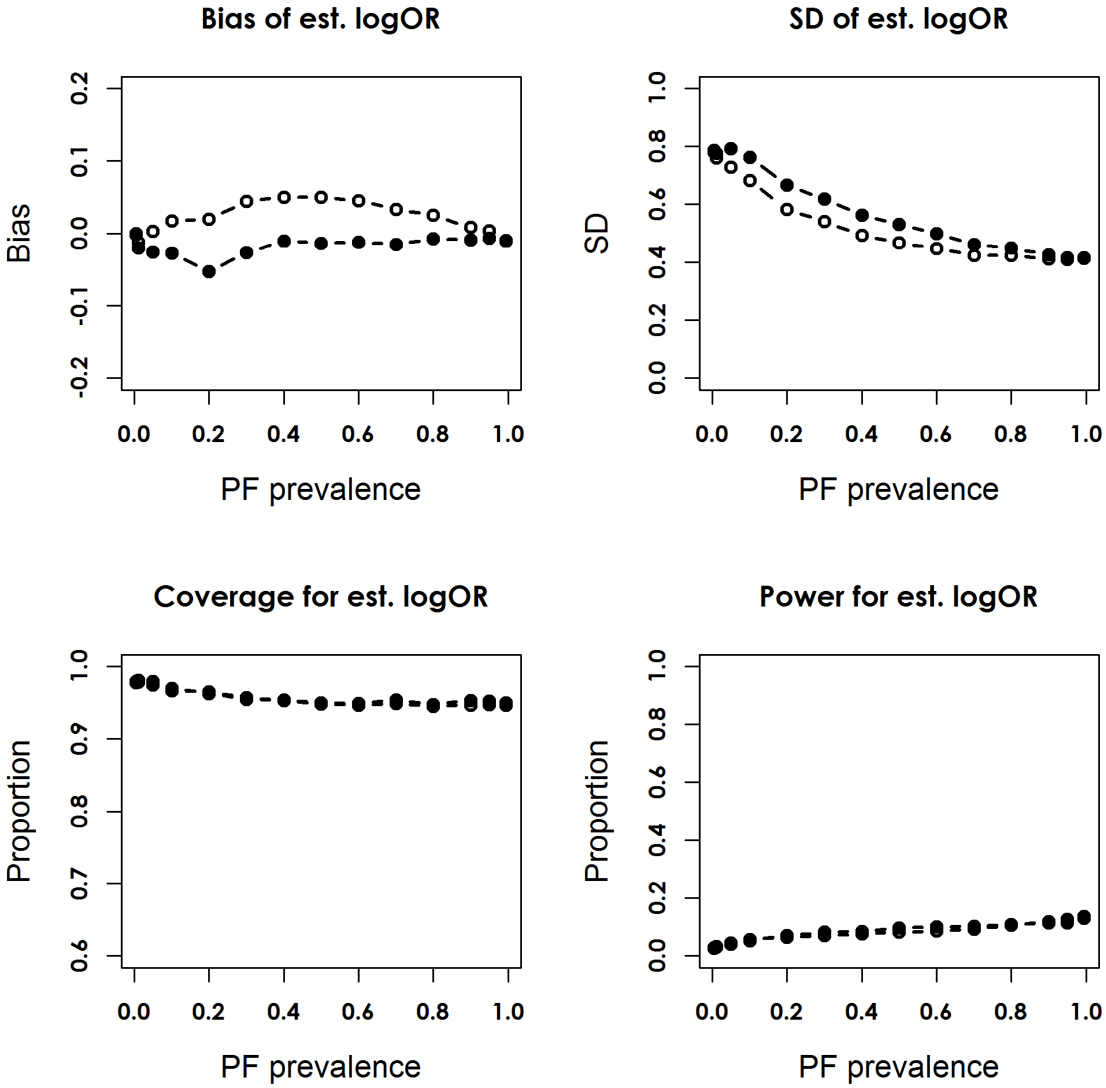


Figure S4 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 500 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

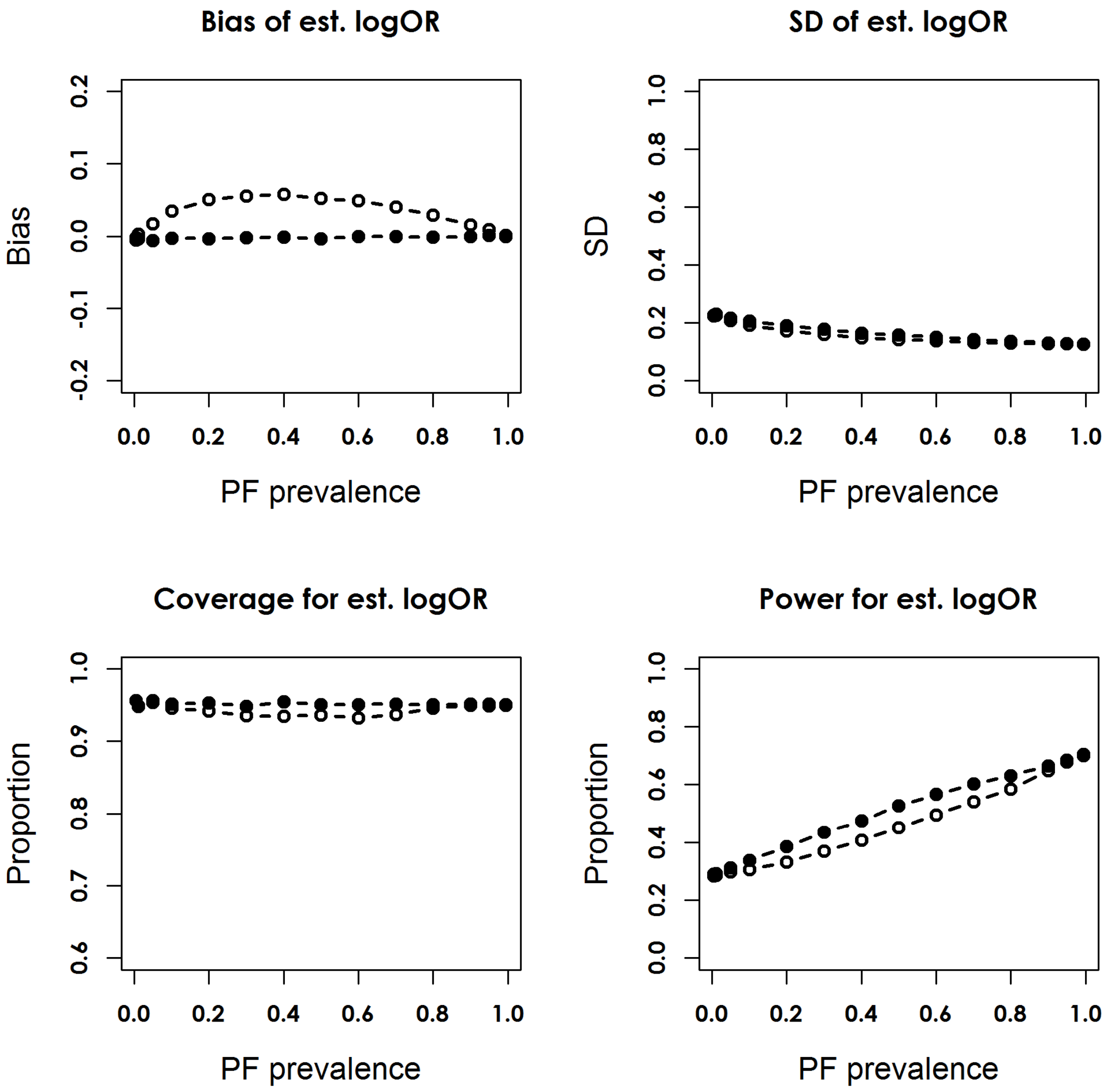


Figure S5 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the unconditional setting, with 1000 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

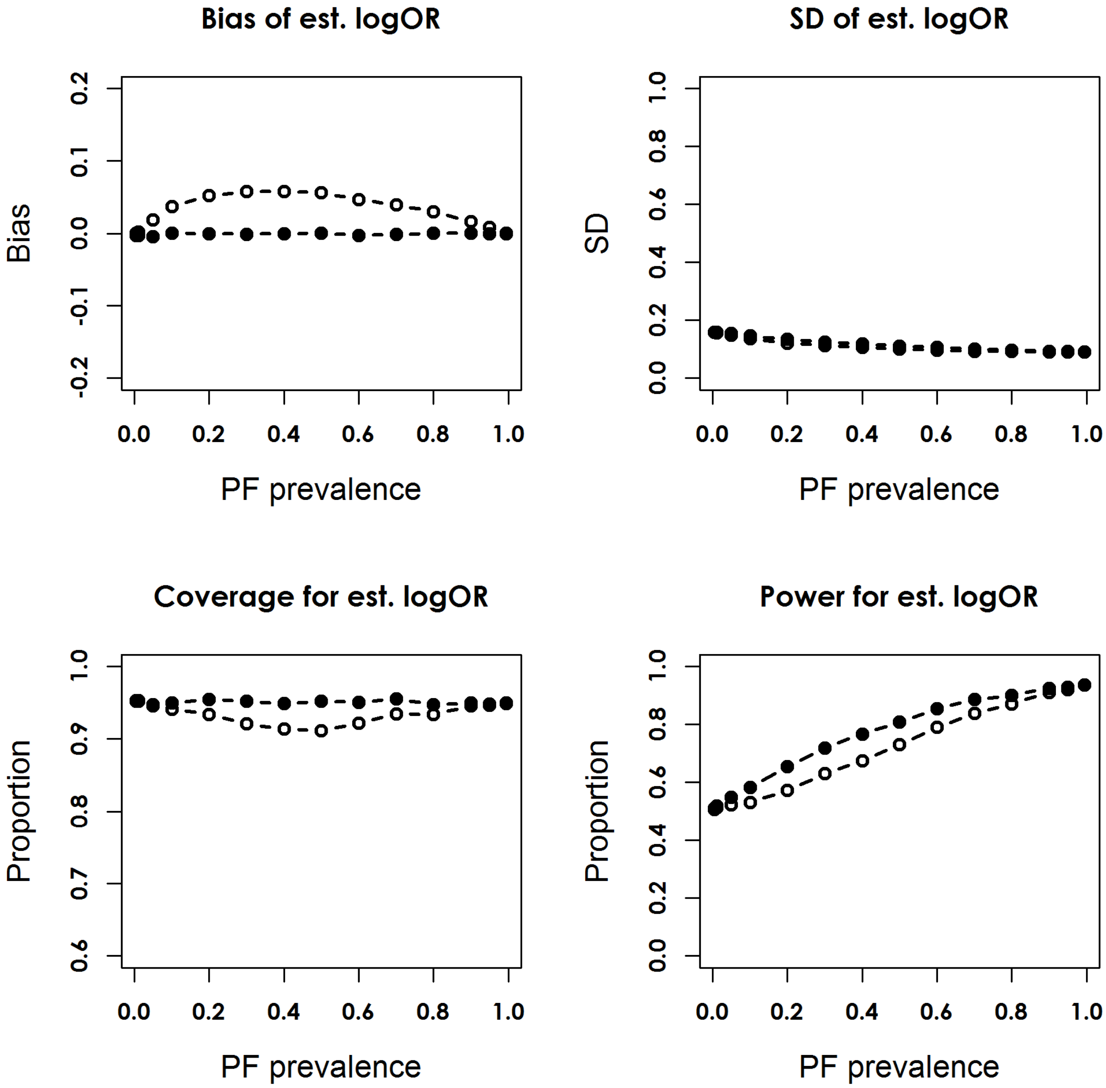


Figure S6 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 2, the unconditional setting, with 25 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by solid line with filled circles.

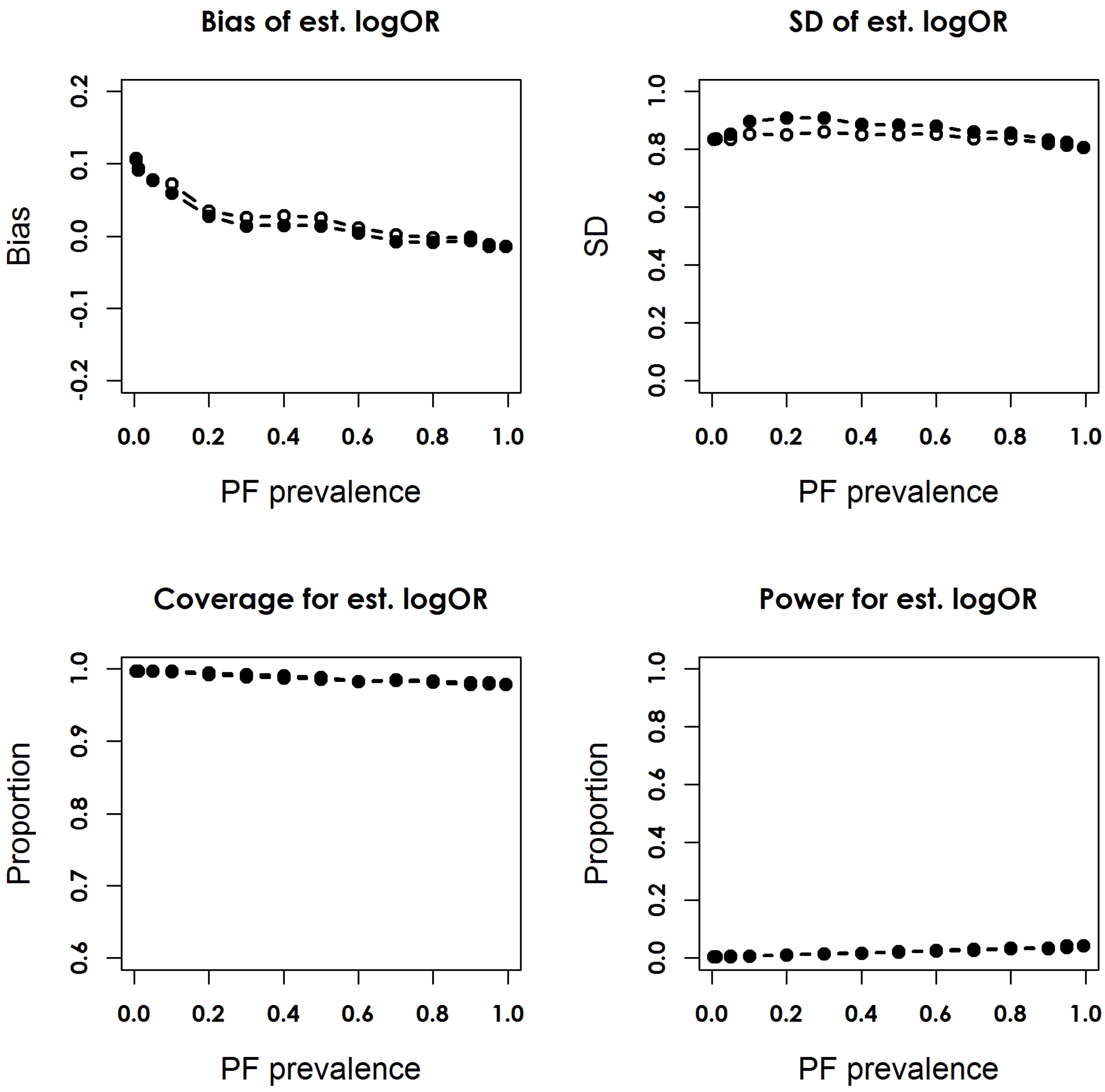


Figure S7 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 2, the unconditional setting, with 50 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

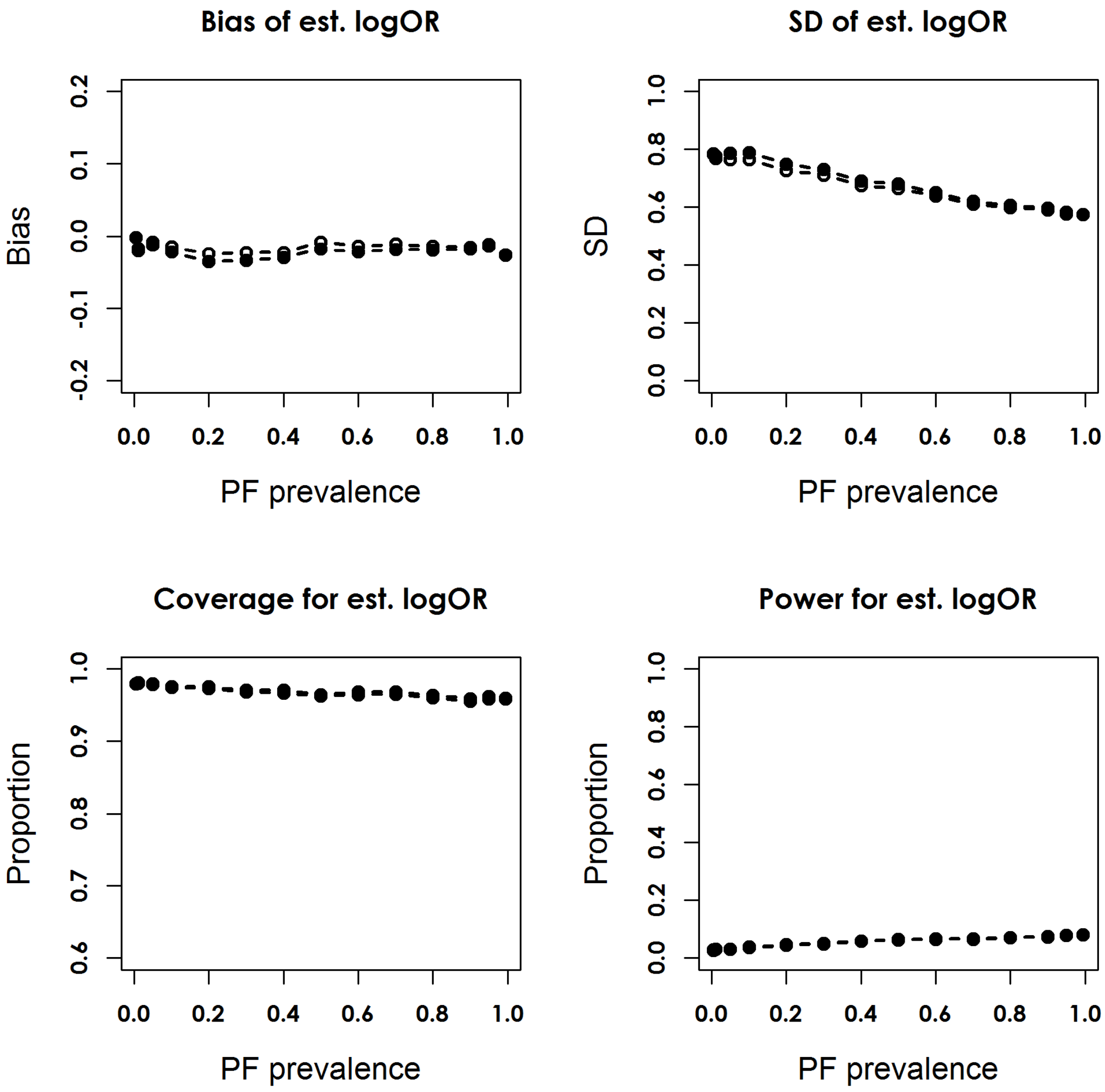


Figure S8 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 2, the unconditional setting, with 125 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by solid line with filled circles.

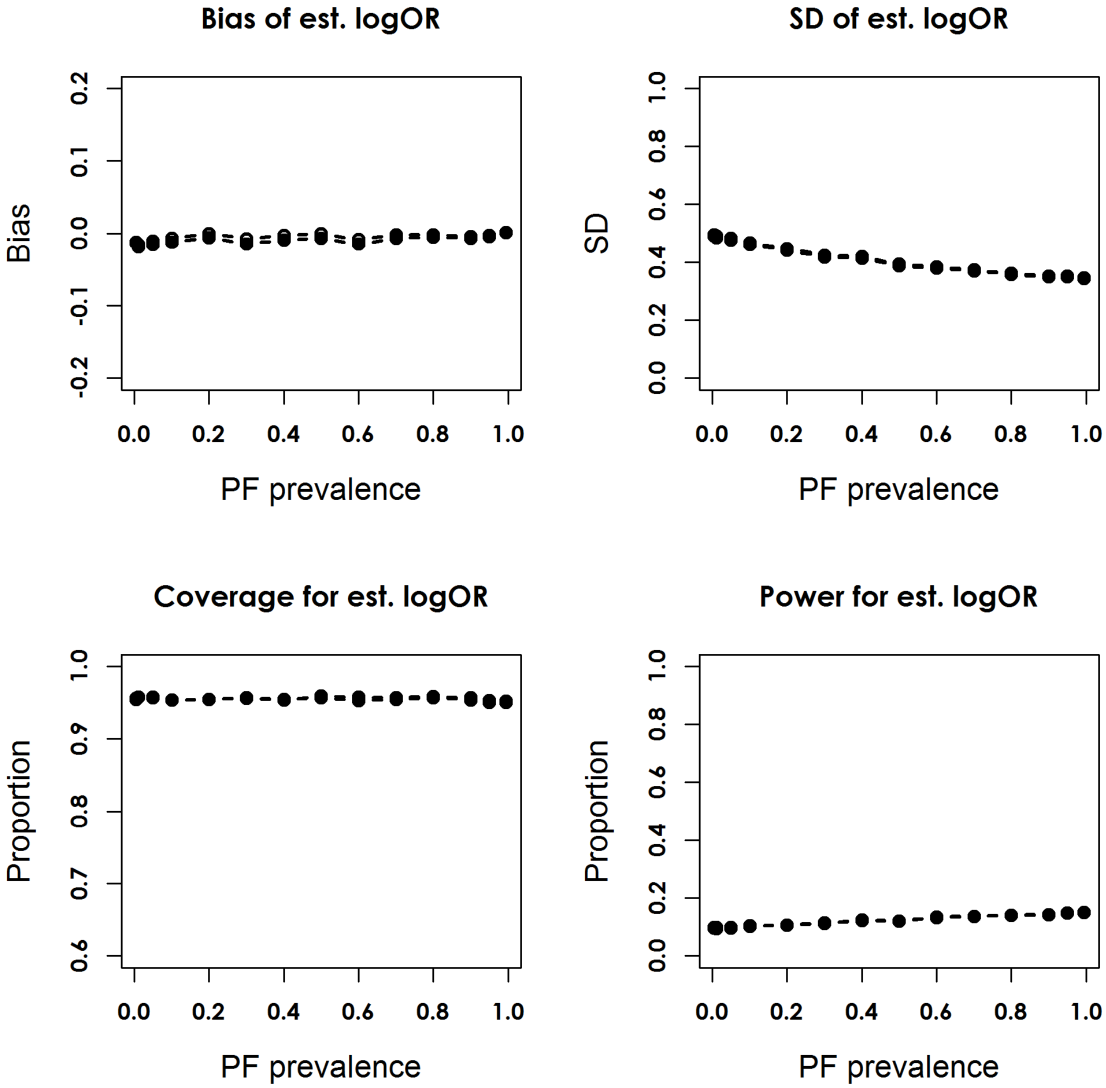


Figure S9 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 2, the unconditional setting, with 2000 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

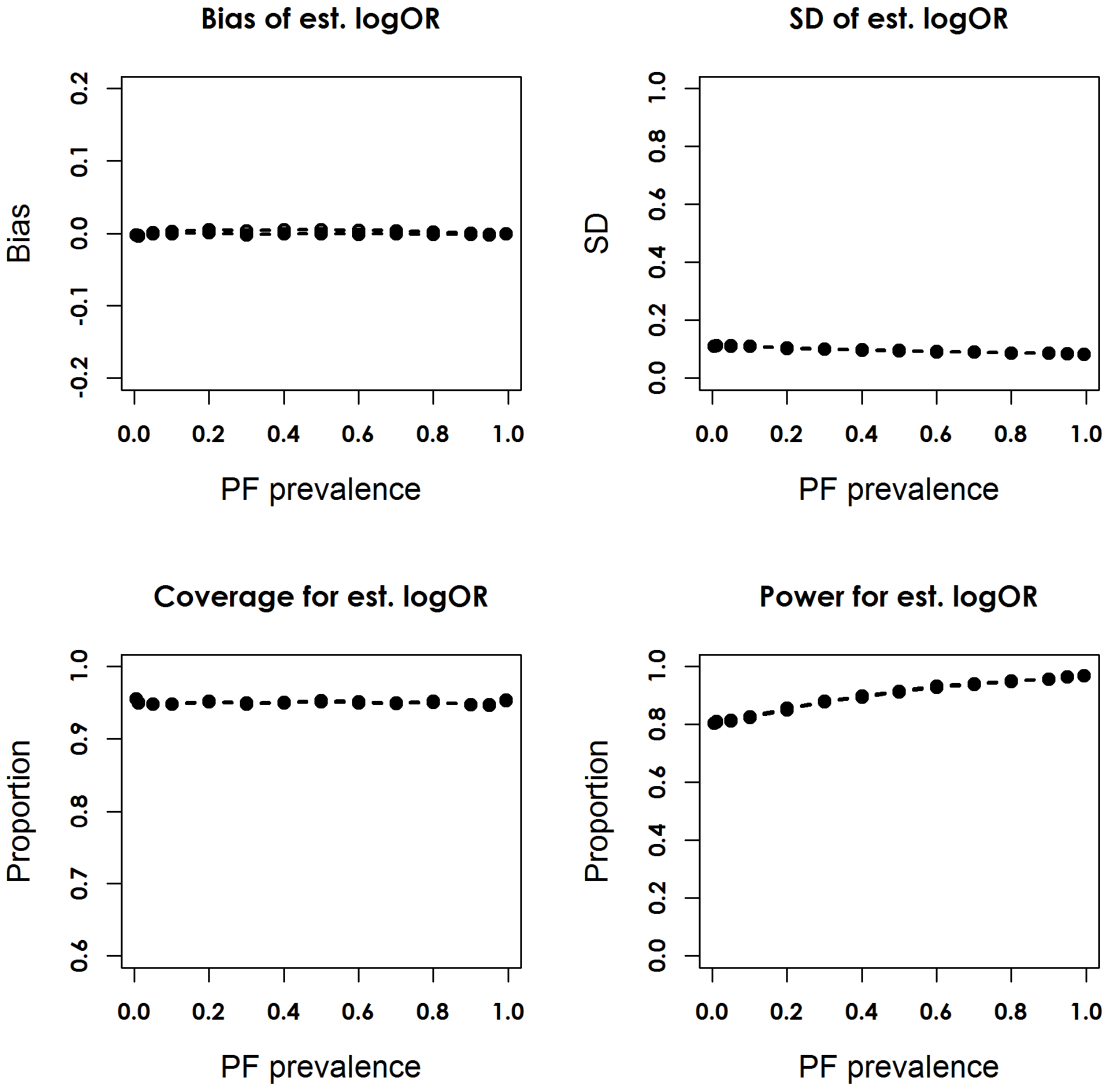


Figure S10 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 3, the unconditional setting, with 125 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

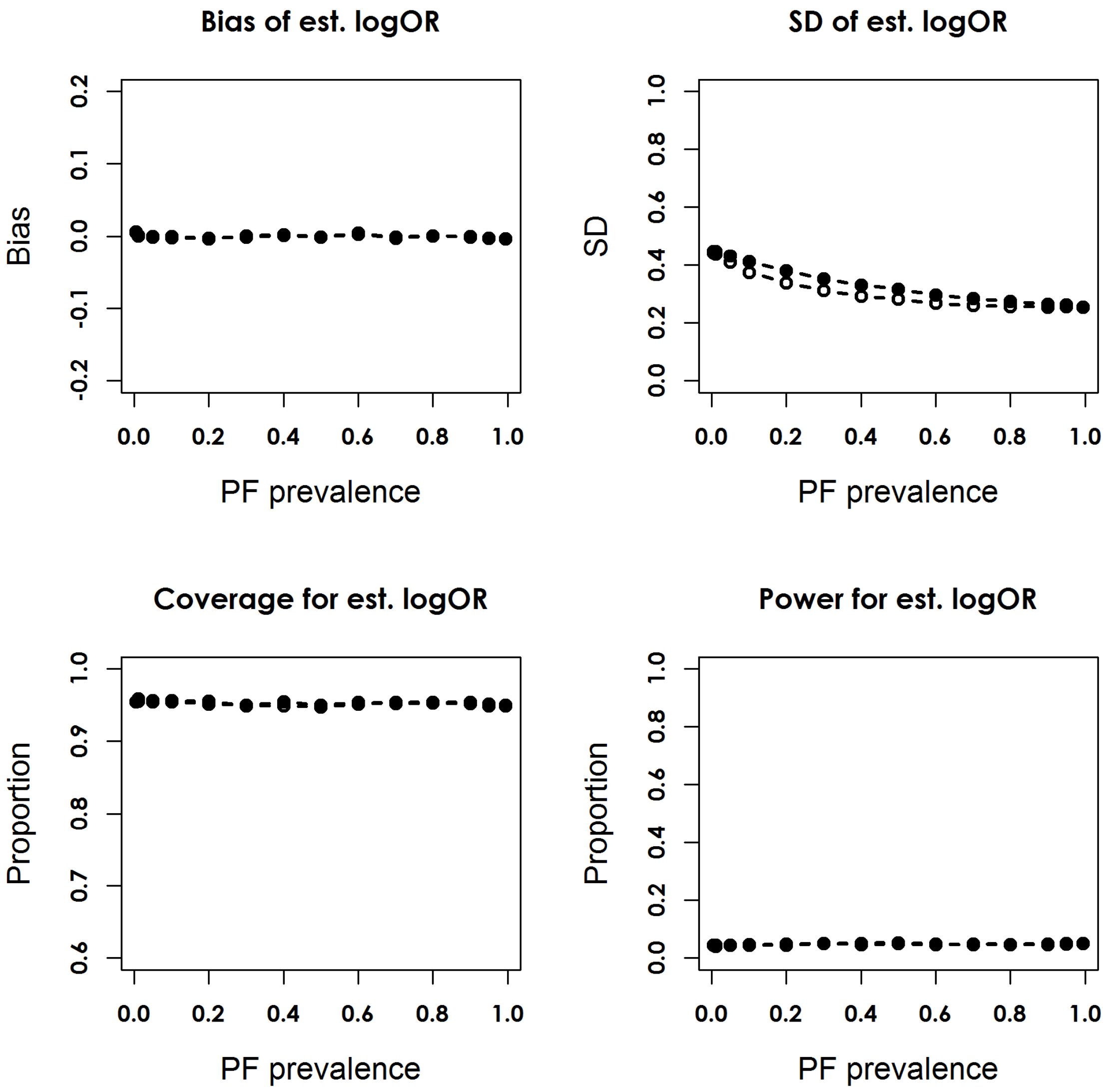


Figure S11 Probability of difference between the estimated and true ORR (underestimation) in scenario 1, the unconditional setting, with 125 patients per arm. Within each graph, lines correspond to Pr ($D_{ORR} \leq - d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

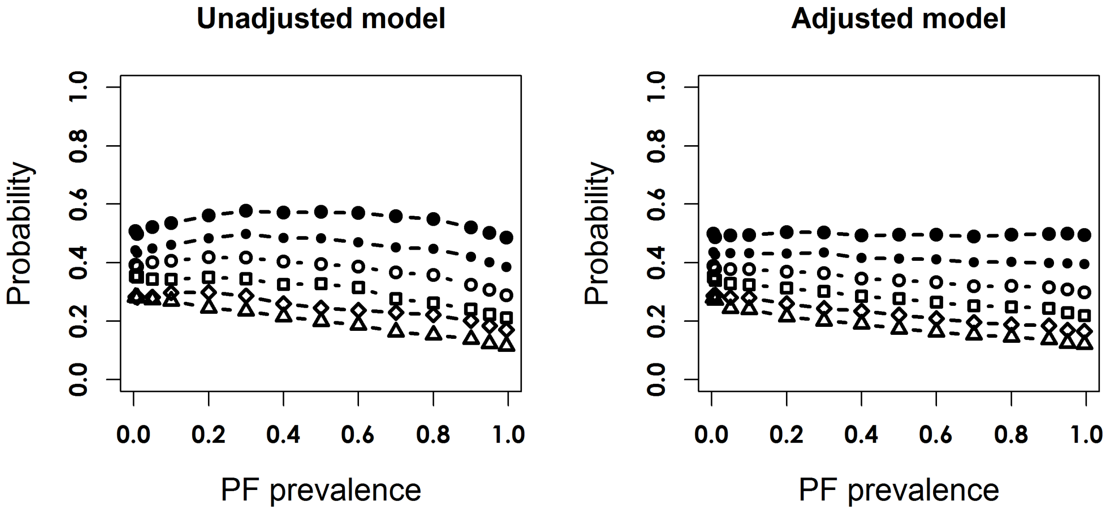


Figure S12 Probability of difference between the estimated and true ORR (overestimation) in scenario 1, the unconditional setting, with 125 patients per arm. Within each graph, lines correspond to Pr ($D_{ORR} \geq d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

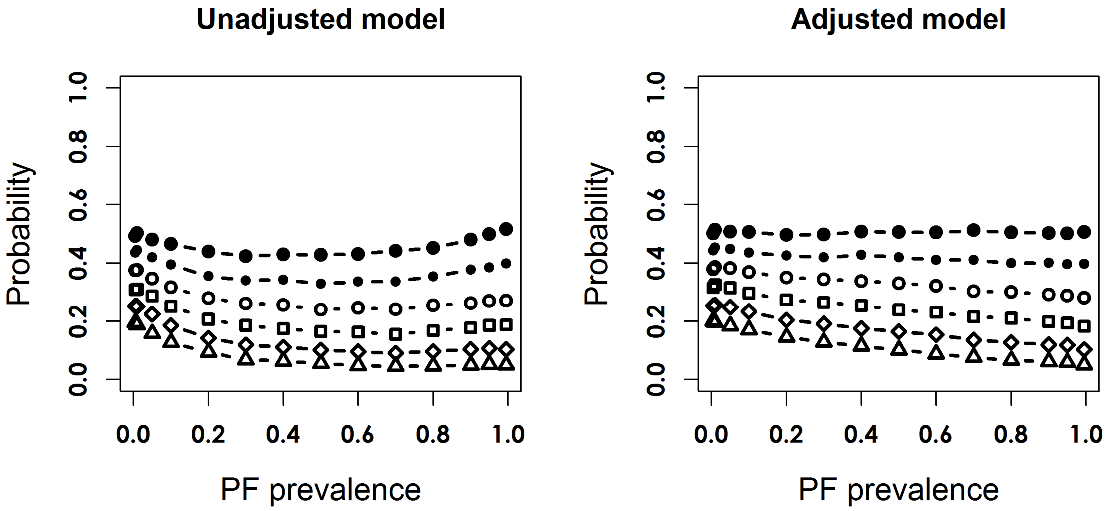


Figure S13 Probability of difference between the estimated and true ORR (underestimation) in scenario 1, the unconditional setting, with 2000 patients per arm. Within each graph, lines correspond to Pr ($D_{ORR} \leq - d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

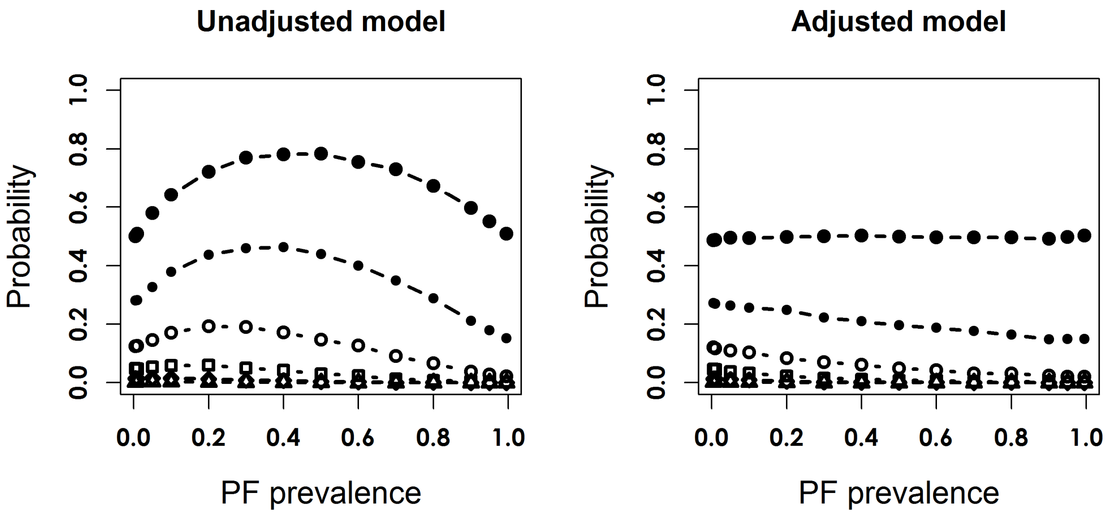


Figure S14 Probability of difference between the estimated and true ORR (overestimation) in scenario 1, the unconditional setting, with 2000 patients per arm. Within each graph, lines correspond to Pr ($D_{ORR} \geq d_2$), where $d_2 = 0$ (solid circle), 0.05 (bullet), 0.10 (little circle), 0.15 (square), 0.2 (diamond) and 0.25 (triangle), from top to bottom, respectively.

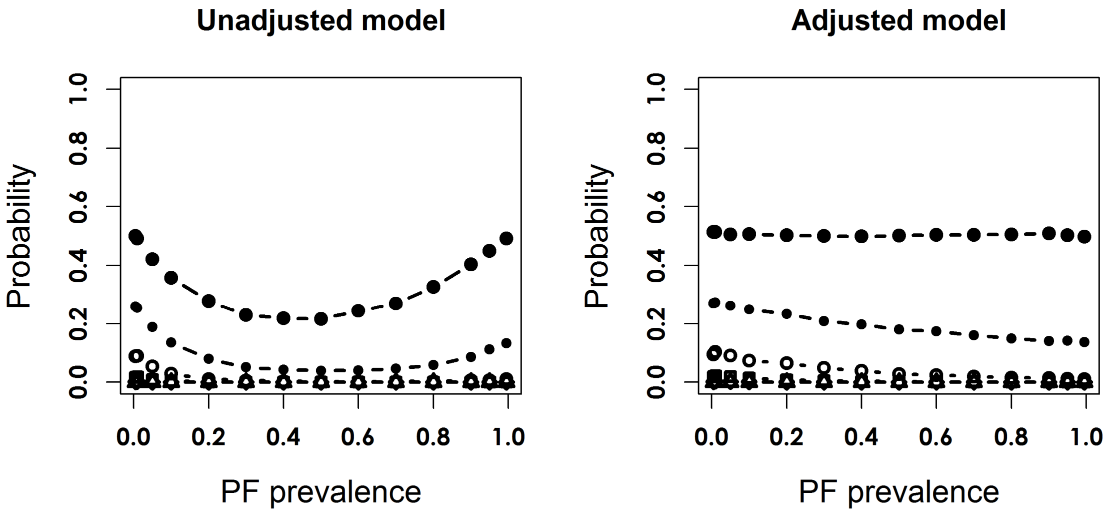


Figure S15 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the conditional setting, with 125 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.

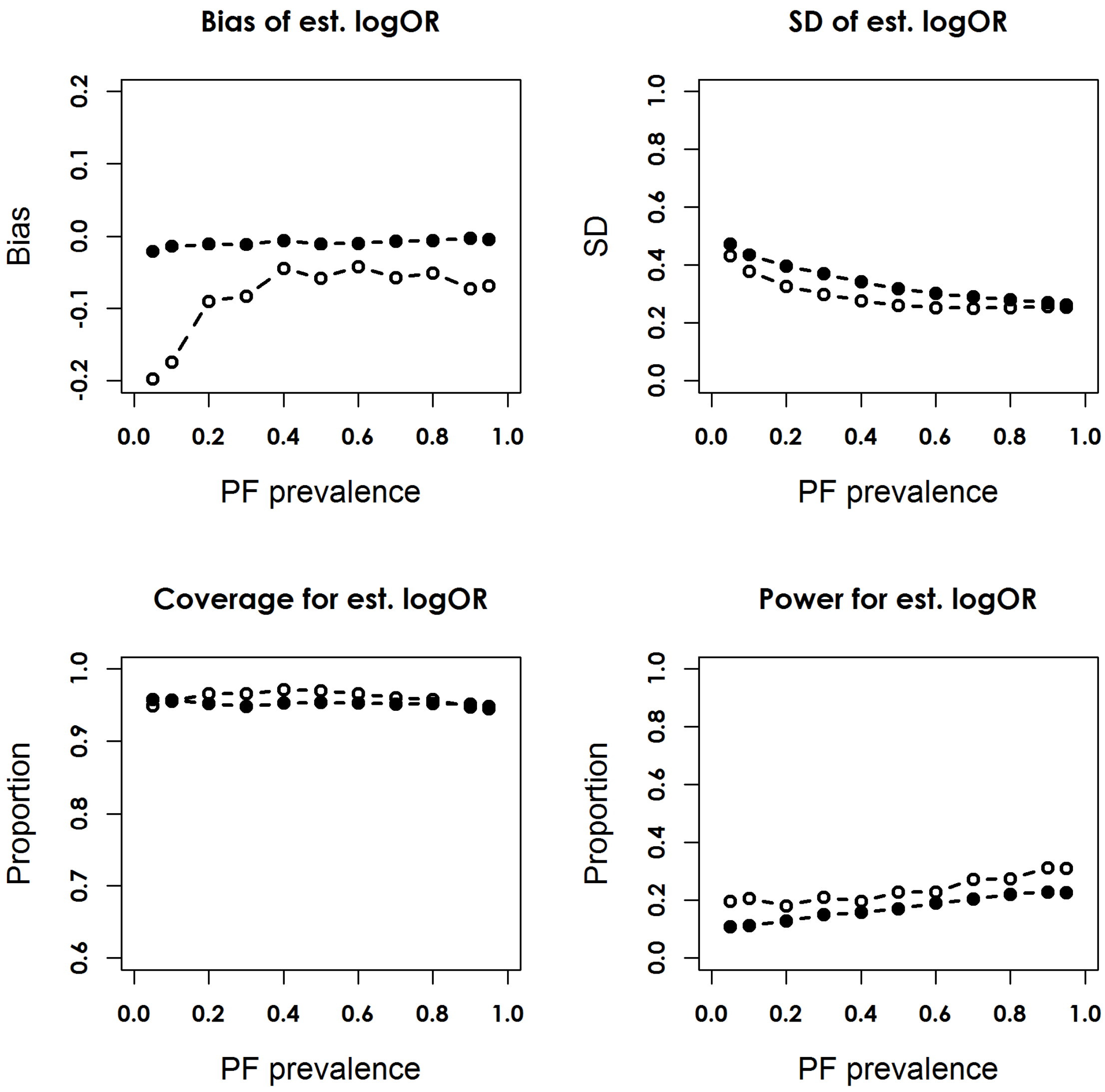
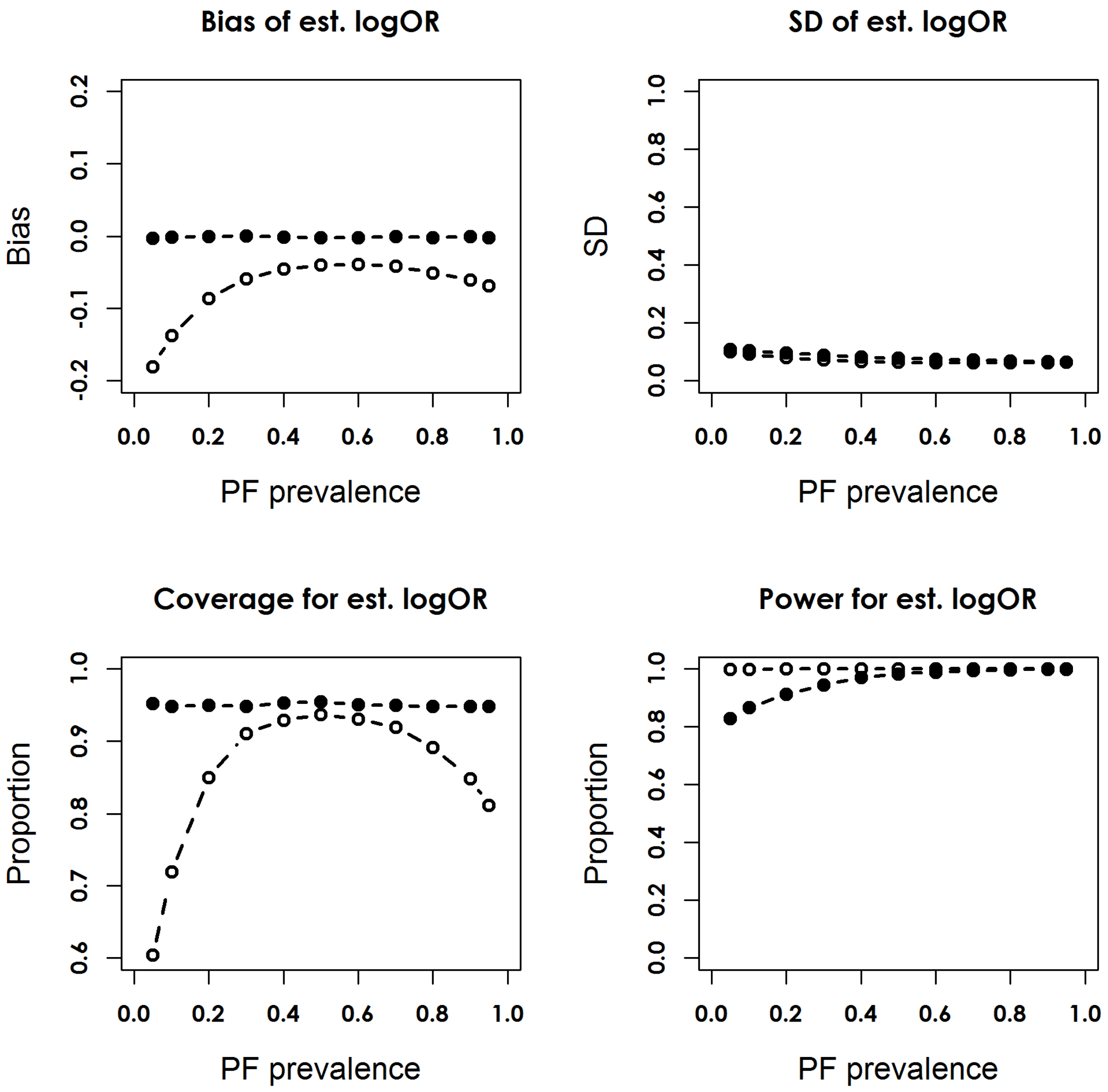


Figure S16 Bias, simulation standard deviation (SD), coverage proportion and statistical power for the unadjusted and adjusted logOR, in scenario 1, the conditional setting, with 2000 patients per arm. The unadjusted model is indicated by the dotted line with hollow circles, and the adjusted model is indicated by the solid line with filled circles.



CHAPTER 4

Impact of Tuberculosis on Mortality among HIV-Infected Patients Receiving Antiretroviral Therapy in Uganda

Rong Chu[1,2], Edward J Mills[1,3,4], Joseph Beyene[1], Eleanor Pullenayegum[1,2], Celestin Bakanda[5], Jean B. Nachega[6,7], P. J. Devereaux[1,8], Lehana Thabane[1,2]

Affiliation

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada

[2]Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton, Canada

[3]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada

[4]Faculty of Health Sciences, University of Ottawa, Ottawa, Canada

[5]The AIDS Support Organization (TASO), Kampala, Uganda

[6]Stellenbosch University, Department of Medicine and Centre for Infectious Diseases, Cape Town, South Africa

[7]Johns Hopkins Bloomberg School of Public Health, Departments of Epidemiology and International Health, Baltimore, MD, USA

[8]Population Health Research Institute, Hamilton Health Sciences, Hamilton, Canada

Corresponding author:

Lehana Thabane

Email: ThabanL@mcmaster.ca

Keywords: Antiretroviral therapy, HIV, tuberculosis, propensity score methods, Uganda, prospective cohort study

ABSTRACT

Background: Tuberculosis (TB) at the initiation of antiretroviral therapy (ART) is expected to importantly affect the likelihood of survival among HIV co-infected patients. Yet, the magnitude of increased mortality is poorly understood.

Methods: Using a prospective cohort study of 22,477 adult patients who initiated ART between August 2000 and June 2009 in Uganda, we assessed the effect of TB at the initiation of ART on all cause mortality using a Cox proportional hazard model on propensity score (PS) matched patients to control for potential confounding. Stratification and covariate adjustment for PS and the conventional Cox models were performed for sensitivity analysis.

Results: A total of 1,609 (7.52%) patients were diagnosed as having TB at the start of ART. TB patients were more likely to be male, have AIDS defining illnesses, belong to WHO disease stage III or IV, and have lower CD4 cell counts at baseline. The percentages of death were 10.47% (95% confidence interval [CI]: 9.01% – 11.93%) and 6.38% (95% CI: 6.05% – 6.71%) for patients with and without TB, respectively. The hazard ratio (HR) for mortality comparing TB and non-TB patients on 1,686 PS matched pairs (HR=1.37, 95% CI: 1.08 – 1.75) was less marked relative to the crude estimate (HR=1.74, 95% CI: 1.49 – 2.04). The other PS methods and the conventional Cox model produced similar results.

Conclusions: After controlling for the important confounding variables, HIV patients who had TB at the initiation of ART have an increase in the risk of overall mortality relative to the non-TB patients.

Introduction

The total number of people living with human immunodeficiency virus (HIV) reached 33.4 million (31.1 – 35.8 million) worldwide by the end of 2008 (1), among whom two thirds resided in Sub-Saharan Africa. It is estimated that one-third of HIV infected people are co-infected with mycobacterium tuberculosis, which could reactivate or get a patient re-infected with a new strain, leading to active tuberculosis (TB) disease. However, TB incidence rates vary importantly according to geography and patient's degree of immunosuppression as evidenced by CD4-T cell count. The incidence of active tuberculosis in HIV-infected patients with latent tuberculosis infection is about 10% per year as compared to 10% per lifetime for an HIV-uninfected individual (2). TB is a leading cause of HIV-related deaths and accounts for one half of the AIDS deaths worldwide (3). Recent trial data have shown that early (within 2 weeks) initiation of antiretroviral therapy during TB therapy can improve survival for patients with co-infection (4-6). Guidelines and policies on joint HIV/ TB interventions have been developed to promote synergies between TB and HIV/AIDS prevention and care activities (7-9), aimed at reducing morbidity and mortality in TB co-infected HIV patients. On the other hand, joint treatment containing ART and anti-TB drugs may be complicated by overlapping toxicity profiles, complex drug-drug interactions, and immune reconstitution inflammatory syndrome (10-13).

To date, the association between active TB and mortality in HIV-infected patients who receive ART is poorly understood, particularly in settings with a relatively lower TB prevalence, such as Eastern Africa (9,13,14) as compare to Southern Africa (14-16). An observational study from Tororo, Uganda, comprising 1,044 HIV patients showed that TB at the initiation or during follow-up of ART was associated with a 4.7 fold increase in cumulative mortality relative to those without TB, without accounting for baseline covariates (17). A study in South Africa, a higher incidence setting, found a lack of association between initiation of pulmonary TB treatment and risk of death when controlling for baseline covariates such as CD4 count and WHO stage IV (15). These results suggested that the association of TB disease and mortality at start of TB treatment was clearly confounded by degree of immunosuppression. Several studies in more developed countries suggest substantial effects of incident TB on AIDS-related mortality among HIV patients subject to moderate ART exposure (18,19). Small cohort numbers of co-infected patients and incomplete information on clinical outcomes have led to widespread misunderstanding about the impact of TB/HIV co-infection on clinical outcomes and the optimal timing of treatment of both diseases. A recent meta-analysis suggested little impact of TB on mortality in HIV patients exposed to ART, yet readers need to bear in mind the heterogeneity in the timing of TB diagnosis and initiation of ART when interpreting the results (20).

Our current study involves a large prospective cohort of HIV positive adult patients in Uganda, aiming to address two objectives: (1) to assess the impact of TB at ART initiation on overall survival among HIV patients during ART treatment by adjusting for potential confounders using propensity score (PS) methods (21-23), and (2) to explore the robustness of the study findings by comparing results from different PS methods (matching on PS, stratifying on PS, adjusting for PS as regression covariate) and the conventional regression model. The propensity score methods have been increasingly used to control for baseline confounding in observational studies (21-26). They are a powerful alternative to estimate the average exposure effects on the whole population or subpopulations using the observed datasets (27). In the PS methods, the vector of potential confounding variables reduces to a single score that reflects one's propensity of being exposed to an intervention or a risk factor. This provides relatively easy means to compare the distributions of individual potential confounders and the support of the PS between exposure groups (28). We hypothesized that having TB at the beginning of the treatment versus not was associated with increased overall mortality in HIV positive patients who receive ART within a 5-year follow-up. These results may help accurately plan TB and HIV/AIDS management activities in HIV patients with TB co-infection.

Methods

Setting, participants and data collection

Our prospective cohort contains HIV-infected patients initiating ART at ten service centres managed by the AIDS Support Organization (TASO) across different settings in Uganda since 2000. Patients received ART from experienced medical staff, at TASO clinics, outreach clinics in rural areas, and through community-based treatment programs. Details of treatment administration have been published previously (29). Non-nucleoside reverse transcriptase inhibitors form the primary initiation regimen. The current study included adult patients (aged 14 years or older at ART initiation) enrolled between

August 2000 and June 2009, as part of an ongoing observational study intended to evaluate the programmatic delivery of services. After ART initiation, patients were scheduled for clinic visits at least every 3 months. Patients' demographic, clinical, psychosocial and medication use data were collected by clinicians and field workers using standardized forms at ART initiation and each visit.

At the time of data collection, national guidelines in Uganda used three strategies for initiating treatment of TB in a co-infected patient: 1) patient with TB and CD4 cell count < 250 cells/mm^3 or a patient with extra-pulmonary TB or WHO stage IV disease: Start TB therapy first and when tolerated (usually within 2-6 weeks) then introduce ART; 2) Pulmonary TB and CD4 250-350 cells/mm^3: start TB therapy for two months then introduce ART, and; 3) pulmonary TB and CD4>350 cells/mm^3: defer ART, monitor clinically and also do CD4 cell counts regularly; re-evaluate the patient at eight weeks and the end of TB treatment (30).

Primary outcome, exposure and potential confounding variables

Our primary outcome was time from the initiation of ART to all-cause mortality. The primary exposure was active pulmonary TB as evidenced by sputum smear positive results at the initiation of ART, or suspected TB regardless of sputum results followed up by radiography. Fourteen variables containing clinical and demographic characteristics, and medicine history measured at the beginning of ART (baseline), were considered in the study and their role as potential confounders was investigated further. The 14 baseline covariates included gender, age, CD4 count, WHO clinical disease stage of HIV/AIDS, presence of AIDS defining illness, TASO service centre (10 sites), calendar year of ART initiation, education, marital status, partner sero-status, sexual activity, sexually transmitted infection (STI), history of pneumocystis jeroveci pneumonia, and toxoplasmosis.

Statistical analysis

We adopted PS methods to estimate the causal effect of TB at the initiation of ART on overall survival of the HIV patients since their initiation of ART and reported the crude and adjusted effect of TB using hazard ratio [HR]. We chose PS matching as the primary method of analysis because empirical evidence suggested it provided better control of confounding relative to the other PS methods (31), and the statistical diagnostic tools for assessing the balance of potential confounders between the matched pairs were readily available (28,32,33). Specifically, we fit a Cox proportional hazard (PH) model on 1-to-1 PS matched pairs to control for the observed potential confounding variables using a two-step procedure. First, we modeled the relationship between having active TB at baseline and the observed baseline covariates using a logistic regression model (“PS model”). We employed an iterative approach (34) to build the PS model by starting with the 14 observed baseline covariates; we then considered pairwise interactions, and polynomial terms of the continuous variables, if including them improved the predictive power the PS model and its ability to balance the distribution of the covariates between the PS matched pairs. We aimed to match each active TB patient to one non-TB patient on the logit of the

PS using calipers of width equal to 0.2 of the standard deviation of the logit of the estimated PS (31,35). Balance diagnostics of the baseline covariates were based on a series of numerical and graphical measures recommended in the literature, including the standardized difference, and for the continuous covariates (age and CD4 cell count), ratio of variance, 5-number summaries, QQ plot, nonparametric density plot, empirical cumulative distribution function and side-by-side boxplot. Second, we fit a stratified (on the matched pairs) Cox PH model to estimate the impact of baseline TB on survival while adjusting for the dependence of outcome within pairs (36). The appropriateness of the PH assumption was assessed using log-log survival curves.

We applied two other PS methods as supportive secondary analysis to assess the robustness or sensitivity of the PS matching analysis results, namely, the stratified Cox regression on 5 subclasses using quintiles of the estimated PS, and the Cox model adjusting for PS and its quadratic and cubic terms as covariates. The conventional Cox regression directly modeling the effect of TB on overall survival, adjusting for multiple baseline covariates in the same form as being included in the final PS model, was also performed as a sensitive analysis.

Complete information was recorded for the study sample on TB status at ART initiation and all but four baseline covariates, namely, patient age (8% missing), CD4 count (17.3% missing), sexual activity (17.3% missing) and WHO clinical stage of HIV/AIDS (34.4% missing). Excluding patients who had missing data on any of the covariates would lead to a loss of 48% subjects and 86% death events. Such exclusion could affect the stability of the PS matching method (due to low event rate) and jeopardize the internal validity and generalizability of the study. Therefore, we performed all analyses on five multiple imputation (MI) datasets (37). We used the Markov chain Monte Carlo method assuming a multivariate normality to create a monotone missing pattern, followed by separate imputations for the continuous (age, log CD4 count) and categorical (sexual activity and stage of disease) variables using linear and logistic regression models. A number of baseline characteristics (TB status, gender, TASO sites, AIDS status, education level, marital status, partner sero status, sexually transmitted infection, calendar year of ART initiation, and history of pneumocystis jeroveci pneumonia or toxoplasmosis) and follow-up variables (death, switch of treatment regimen, patient adherence, and AIDS status post ART initiation) were included in the imputation procedure. While the same covariates were used to achieve prognostic balance on the five imputed datasets, the propensity scores were estimated separately for each individual dataset based on which an estimated effect of baseline TB on mortality was obtained. We then calculated the overall estimate of TB effect reported as hazard ratio [HR] with a 95% confidence interval (CI), and the associated p-value, using Rubin's rule (37).

We conducted all statistical analyses in SAS version 9.2 (Cary, NC), R 2.14.0 (R Core Development Team) and Stata 10 (College Station, TX).

Ethics approval

University of British Columbia, University of Ottawa, and Mbale Regional Referral Hospital research ethics boards approved this study.

Results

Participant baseline characteristics

The study cohort consists of 22,477 HIV-infected adult patients aged 14 years or older who started ART during August 2000 to June 2009 (Supplementary File #1: Figure S1). A total of 1,690 amongst the 22,477 patients (7.52%, 95% CI: 7.17% -7.86%) in the primary study sample were diagnosed having active TB at the initiation of ART. Table 1 shows that TB patients tended to have a worse HIV-related prognosis at ART initiation. TB patients tended to be younger, more likely to be male, suffer more from AIDS defining illnesses, belong to WHO clinical stage of III or IV, have a lower CD4 cell count, and have a history of pneumocystis jeroveci pneumonia. The prevalence of baseline TB disease was also unbalanced among the participating TASO centres.

Association between TB at the initiation of ART and baseline covariates (PS model)

The final PS model included all 14 baseline covariates, their pairwise interactions with gender and baseline AIDS status, and additional quadratic and interaction terms for the continuous variables to improve model fit (Supplementary File #2: Final PS model). Overall, the PS model predicted baseline TB status well. We were able to match 1,686 among all 1,690 baseline TB patients to an equal number of non-TB patients. PS matching improved the similarity of the distributions of all baseline covariates between TB and non-TB patients and reduced the standardized differences to below 0.1. Table 2 and Figures 1 and 2 display the balancing distributions of the baseline covariates for one of the five imputed datasets. Consistent patterns were observed across the imputed datasets.

Association between TB and all-cause mortality

During a median of 21.5 months of follow-up, 1,503 (6.69%, 95% CI: 6.36%-7.01%) among the 22,477 HIV patients who received ART died from all causes. The percentages of death were 10.47% (95% CI: 9.01% – 11.93%) and 6.38% (95% CI: 6.05% – 6.71%) for patients with and without TB, respectively, in the original unmatched sample. The risk of mortality remained higher in the 1,686 PS matched pairs, with 176 (10.44%, 95% CI: 8.98%-11.90%) and 137 (8.13%, 95% CI: 8.13%-6.82%) deaths in the TB and non-TB groups. Compared with the crude difference in overall survival (HR=1.74, 95% CI: 1.49 – 2.04) in the original sample, the hazard ratio for all cause mortality comparing TB and non-TB patients on the 1,686 PS matched pairs (HR= 1.37, 95% CI: 1.08 – 1.75) was less marked, indicating TB was associated with a 37% increase in the risk of death over the course of the study (Table 3). Kaplan-Meier survival curves are displayed in Figure 3. No violation of the PH assumption was suggested by the log-log survival curves.

Sensitivity analyses

Stratifying on PS quintile and adjusting for PS as a covariate in Cox regression models yielded similar TB effects on survival (PS-stratified HR=1.36, 95% CI: 1.15-1.60; PS-adjusted HR=1.34, 95% CI: 1.14 -1.58) (Table 3). Adjusting for multiple baseline covariates simultaneously (as being entered in the PS model) in the conventional Cox model did not alter the association substantially (HR= 1.40, 95% CI: 1.19-1.65)

Discussion

This is, to our knowledge, the largest research cohort of HIV-infected patients receiving ART in a single African country that has evaluated TB outcomes in co-infected patients. The descriptive analysis showed that TB patients were more likely to be highly immunosuppressed and have a history of opportunistic infections (e.g. lower CD4 cell counts, more advanced WHO stage of HIV, and history of pneumocystis jiroveci pneumonia), and were more prevalent in certain geographical locations in Uganda. The results of the propensity score matching analysis showed that having TB at the initiation of ART led to a 37% increase in the risk of all cause mortality during the follow-up period between 2000 and 2009. Sensitivity analyses based on two other PS methods and the conventional Cox regression models showed similar results.

The prevalence of active TB at the initiation of ART of 7.52% (1,690/22,477) in the study is consistent with results from a smaller TASO cohort reported from Tororo, Uganda (17). This indicates that TB is less prevalent in Uganda compared with other African countries (14,15). An open cohort of 7,512 patients receiving ART in South Africa suggested that

15.9% of HIV patients were being treated for pulmonary TB at the time of ART initiation between 2004 and 2007. Our finding of a 37% increase in hazard of all cause mortality associated with having active TB at the initiation of ART after controlling for potential baseline confounding is both statistically significant and clinically relevant. This result is different from Westreich et al's findings on 7,512 HIV patients in South Africa, indicating pulmonary TB treatment at ART initiation was uncorrelated to the overall survival (HR=1.06, 95% CI: 0.75-1.49) after adjusting for multiple confounders (15). There are several potential explanations for these conflicting data. First, all TB patients in the South Africa study received ART soon after initiation of TB treatment if well tolerated, which was less the case in the current study. Second, different sets of potential confounders as a proxy of disease prognosis were used in the two studies. While the presence of AIDS defining illnesses, history of pneumocystis jiroveci pneumonia and partner information at ART initiation were collected in the current study, body mass index and presence of anemia were not gathered in our study, nevertheless, controlled for in the South Africa study. Third, our primary goal was to assess the effect of TB at ART initiation in TB patients, by comparing them to the non-TB patients who had similar baseline demographic and clinical characteristics, using PS matching method. The South Africa study on the other hand estimated the average effect of TB using inverse probability weighting, as if one shifted the entire population from having active TB to not having TB, or vice-versa, at ART initiation (31). Forth, in addition to having a lower prevalence of active TB at baseline, the current cohort had higher CD4 cell counts (TB patients: median = 111, inter-quantile range [IQR]: 45-193; non-TB patients: median

=139, IQR: 66-208) relative to the South Africa study (TB patients: median = 58, IQR: 22-116; non-TB patients: median = 94, IQR: 34-165). A potential interaction between the TB and CD4 count may also result in different effects of TB measured in the study population.

Guidelines on the treatment of TB co-infection with HIV have changed dramatically over the last few years and may not reflect within-country variability. For example, early guidelines feared drug interactions between TB medication and ART or negative impact on adherence of ART or TB due to pill burden and suggested early treatment for TB disease followed by stabilization with ART. However, more recent evidence from randomized trials (4-6) suggests that starting ART within 2 weeks of TB treatment to improve survival appears to be well tolerated, and has persuaded guideline makers to recommend the co-administration of treatment. In addition, recent data from South Africa found that although immune reconstitution inflammatory syndrome (IRIS) events were associated with slightly lower adherence rates, overall adherence to ART remained high in this TB-HIV co-infected population and that concerns about IRIS or drug-drug interactions should not deter clinicians from early ART initiation, but patients developing IRIS events should be monitored closely and potentially targeted for interventions to increase adherence(38). Other developments are related to the specific setting of co-infection. Uganda, for example, has a lower national prevalence of co-infection than Southern Africa and has not reported any cases of extremely drug resistant TB infections. South Africa, on the other hand, has had an epidemic of XDR co-infections that have resulted in quarantine of some patients and bi-directional recommendations about interacting with other individuals (39). National guidelines from Uganda do not match those of South Africa, as discussed earlier.

Limitations of the study emerge from three key sources. First, propensity score methods cannot balance unobserved confounders in estimating a causal relationship, which is an inherited issue faced by all observational cohort studies. Our results may be biased if some unobserved prognostic factors were unbalanced between TB and non-TB groups. More resources need to be dedicated to improve the completeness and accuracy of data collection in HIV patients in underdeveloped regions. Although statistical tools, for instance the instrumental variable analysis, has been developed to control for bias due to unobserved confounders (40), strong assumptions are required to ensure estimation accuracy, and such assumptions can be difficult to verify using empirical data. Second, the multiple imputation method employed to impute the missing values for baseline covariates requires the missing at random (MAR) assumption (i.e. the missingness can be explained by differences in the observed data). Although the current imputation procedure included a number of baseline characteristics and follow-up variables, we could not rule out the existence of additional variables that were highly associated with the missingness. The MAR assumption is unverifiable using data collected within a study. Third, our results are predominantly based on adult, HIV positive, African patients who initiated ART in Uganda, during 2004 to 2009, hence limiting their generalizability. The effect of the active TB on HIV patients in other age or ethnic groups and from different geographic regions needs to be further investigated.

Our study shows a 37% increase in the hazard f death among co-infected patients compared with those only infected with HIV. We used pulmonary TB as documented through sputum and radiologic confirmation for the diagnosis of TB as TASO, along with most AIDS Service Organizations in Africa lack the infrastructure to diagnose extra-pulmonary TB. Also, we recognize that the sensitivity of sputum smear for the diagnostic of pulmonary TB in HIV-infected patients is poor (less than 50%) and thus follow-up suspected patients with radiologic examinations regardless of sputum results. Also, there are certain populations that may be at increased risk of co-infection that may otherwise go unnoticed. Children and adolescents, for example, report high rates of co-infection, possibly due to close confinement of living quarters (e.g. sharing beds). However, this population is more likely to go undiagnosed for HIV infection, especially in the absence of parents, and may have challenges adhering to their medications or accessing treatment (41,42).

TASO patients are requested to identify a treatment supporter and discuss adherence with specific adherence counselors. However, they do not receive directly observed TB treatment, as is common in settings such as South Africa (43). The utility of directly observed treatment has been questioned in both HIV and TB separately (44,45), although the utility of this approach in co-infected patients has not been evaluated. It is possible that directly observed treatment has a role within this high-risk population as the first 3-

months of therapy represents the greatest likelihood of treatment success or failure (46). Innovative interventions to promote both adherence and retention of patients in treatment, especially in the early stages is potentially cost-effective in settings such as Uganda as they have lower rates of co-infections and thus would require fewer human resources.

Conclusions

In summary, our current study suggests a moderate increase in risk of death associated with having active TB disease at the initiation of ART in HIV positive patients receiving ART in sub-Saharan Africa. Our finding is complementary to and strengthens results of the recent SAPIT (4), CAMELIA (5), and STRIDE trials (6) that demonstrate significant improvement on survival when ART is initiated during TB therapy. The results also validate the WHO guidelines that urge a more aggressive approach to management of both TB and HIV (47).

Acknowledgements

RC received studentship from Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare Hamilton. RC holds holds a doctoral research award from the Canadian Institutes of Health Research. This study was supported in part by funds from the CANNeCTIN program.

RC participated in the conception and design of the study, data cleaning and analysis, results interpretation, and drafting and revision of the manuscript, and has read and approved the final manuscript.

EM participated in the conception and design of the study, data collection, results interpretation, and revision of the manuscript, and has read and approved the final manuscript.

JB participated in review of statistical methods, results interpretation, and revision of the manuscript, and has read and approved the final manuscript.

CN participated in data collection from TASO clinics and review of drafts, and has read and approved the final manuscript.

JBN participated in results interpretation and revision of the manuscript, and has read and approved the final manuscript.

LT participated in the conception and design of the study, results interpretation, and revision of the manuscript, and has read and approved the final manuscript.

References

(1) World Health Organization. Global summary of the HIV/AIDS epidemic. 2008; Available at: http://www.who.int.libaccess.lib.mcmaster.ca/hiv/data/2009_global_summary.gif. Accessed November/15, 2009.

(2) Selwyn PA, Hartel D, Lewis VA, Schoenbaum EE, Vermund SH, Klein RS, et al. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. N Engl J Med 1989 Mar 2;320(9):545-550.

(3) Centers for Disease Control and Prevention. TB and HIV/AIDS. CDC HIV/AIDS facts. 2008; Available at: http://www.cdc.gov/hiv/resources/factsheets/PDF/hivtb.pdf. Accessed November/15, 2009.

(4) Abdool Karim SS, Naidoo K, Grobler A, Padayatchi N, Baxter C, Gray A, et al. Timing of initiation of antiretroviral drugs during tuberculosis therapy. N Engl J Med 2010 Feb 25;362(8):697-706.

(5) Blanc FX, Sok T, Laureillard D, Borand L, Rekacewicz C, Nerrienet E, et al. Earlier versus later start of antiretroviral therapy in HIV-infected adults with tuberculosis. N Engl J Med 2011 Oct 20;365(16):1471-1481.

(6) Havlir DV, Kendall MA, Ive P, Kumwenda J, Swindells S, Qasba SS, et al. Timing of antiretroviral therapy for HIV-1 infection and tuberculosis. N Engl J Med 2011 Oct 20;365(16):1482-1491.

(7) World Health Organization. Joint HIV/Tuberculosis Interventions. Available at: http://www.who.int.libaccess.lib.mcmaster.ca/hiv/topics/tb/tuberculosis/en/. Accessed November/15, 2009.

(8) Centers for Disease Control and Prevention. Reported HIV Status of Tuberculosis Patients - United States, 1993-2005. Morbidity and Mortality Weekly Report 2007 October 26(56):1103-1106.

(9) Harries AD, Zachariah R, Corbett EL, Lawn SD, Santos-Filho ET, Chimzizi R, et al. The HIV-associated tuberculosis epidemic--when will we act? Lancet 2010 May 29;375(9729):1906-1919.

(10) Burman WJ, Jones BE. Treatment of HIV-related tuberculosis in the era of effective antiretroviral therapy. Am J Respir Crit Care Med 2001 Jul 1;164(1):7-12.

(11) Narita M, Ashkin D, Hollender ES, Pitchenik AE. Paradoxical worsening of tuberculosis following antiretroviral therapy in patients with AIDS. Am J Respir Crit Care Med 1998 Jul;158(1):157-161.

(12) McIlleron H, Meintjes G, Burman WJ, Maartens G. Complications of antiretroviral therapy in patients with tuberculosis: drug interactions, toxicity, and immune reconstitution inflammatory syndrome. J Infect Dis 2007 Aug 15;196 Suppl 1:S63-75.

(13) Gandhi NR, Nunn P, Dheda K, Schaaf HS, Zignol M, van Soolingen D, et al. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 2010 May 22;375(9728):1830-1843.

(14) Lonnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet 2010 May 22;375(9728):1814-1829.

(15) Westreich D, MacPhail P, Van Rie A, Malope-Kgokong B, Ive P, Rubel D, et al. Effect of pulmonary tuberculosis on mortality in patients receiving HAART. AIDS 2009 Mar 27;23(6):707-715.

(16) Lawn SD, Kranzer K, Edwards DJ, McNally M, Bekker LG, Wood R. Tuberculosis during the first year of antiretroviral therapy in a South African cohort using an intensive pretreatment screening strategy. AIDS 2010 Jun 1;24(9):1323-1328.

(17) Moore D, Liechty C, Ekwaru P, Were W, Mwima G, Solberg P, et al. Prevalence, incidence and mortality associated with tuberculosis in HIV-infected patients initiating antiretroviral therapy in rural Uganda. AIDS 2007 Mar 30;21(6):713-719.

(18) Lopez-Gatell H, Cole SR, Hessol NA, French AL, Greenblatt RM, Landesman S, et al. Effect of tuberculosis on the survival of women infected with human immunodeficiency virus. Am J Epidemiol 2007 May 15;165(10):1134-1142.

(19) Lopez-Gatell H, Cole SR, Margolick JB, Witt MD, Martinson J, Phair JP, et al. Effect of tuberculosis on the survival of HIV-infected men in a country with low tuberculosis incidence. AIDS 2008 Sep 12;22(14):1869-1873.

(20) Straetemans M, Bierrenbach AL, Nagelkerke N, Glaziou P, van der Werf MJ. The effect of tuberculosis on mortality in HIV positive people: a meta-analysis. PLoS One 2010 Dec 30;5(12):e15241.

(21) Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55.

(22) Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984;79:516-524.

(23) Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007 Jan 15;26(1):20-36.

(24) Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008 May 30;27(12):2037-2049.

(25) Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 2010 Dec;36(12):1993-2003.

(26) Tleyjeh IM, Kashour T, Zimmerman V, Steckelberg JM, Wilson WR, Baddour LM. The role of valve surgery in infective endocarditis management: a systematic review of observational studies that included propensity score analysis. Am Heart J 2008 Nov;156(5):901-909.

(27) Guo S, Fraser MW. Propensity Score Analysis: Statistical Methods and Applications. Thousand Oaks, CA: SAGE Publications, Inc.; 2010.

(28) Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009 Nov 10;28(25):3083-3107.

(29) Bakanda C, Birungi J, Nkoyooyo A, Featherstone A, Cooper CL, Hogg RS, et al. Cohort Profile: The TASO-CAN Cohort Collaboration. Int J Epidemiol 2011 Mar 7.

(30) Ministry of health. National Antiretroviral Treatment and Care Guidelines for Adults and Children. Kampala, Uganda: Earnest Publishers; 2008.

(31) Austin PC. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. Multivariate Behav Res 2011;46(1):119-151.

(32) Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers; 1988.

(33) Ho DE. Using propensity scores to help design observational studies: application to the tobacco litigation. Health sciences and outcomes research methodology 2001;2:169-188.

(34) Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2006 Jun 30;25(12):2084-2106.

(35) Coca-Perraillon M. Local and Global Optimal Propensity Score Matching. SAS Global Forum 2007.

(36) Cummings P, McKnight B, Greenland S. Matched cohort methods for injury research. Epidemiol Rev 2003;25:43-50.

(37) Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2002.

(38) Morroni C, Chaisson R, Goliath R, Efron A, Ram M, Maartens G, Nachega J. Influence of IRIS on ART Adherence in HIV+ Adults: South Africa.CROI 2012, Abstract #941.

(39) Gandhi NR, Moll A, Sturm AW, Pawinski R, Govender T, Lalloo U, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet 2006 Nov 4;368(9547):1575-1580.

(40) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007 Jan 17;297(3):278-285.

(41) Nachega JB, Hislop M, Nguyen H, Dowdy DW, Chaisson RE, Regensberg L, et al. Antiretroviral therapy adherence, virologic and immunologic outcomes in adolescents compared with adults in southern Africa. J Acquir Immune Defic Syndr 2009 May 1;51(1):65-71.

(42) Kiboneka A, Wangisi J, Nabiryo C, Tembe J, Kusemererwa S, Olupot-Olupot P, et al. Clinical and immunological outcomes of a national paediatric cohort receiving combination antiretroviral therapy in Uganda. AIDS 2008 Nov 30;22(18):2493-2499.

(43) Jack C, Lalloo U, Karim QA, Karim SA, El-Sadr W, Cassol S, et al. A pilot study of once-daily antiretroviral therapy integrated with tuberculosis directly observed therapy in a resource-limited setting. J Acquir Immune Defic Syndr 2004 Aug 1;36(4):929-934.

(44) Volmink J, Garner P. Directly observed therapy for treating tuberculosis. Cochrane Database Syst Rev 2007 Oct 17;(4)(4):CD003343.

(45) Ford N, Nachega JB, Engel ME, Mills EJ. Directly observed antiretroviral therapy: a systematic review and meta-analysis of randomised clinical trials. Lancet 2009 Dec 19;374(9707):2064-2071.

(46) Braitstein P, Brinkhof MW, Dabis F, Schechter M, Boulle A, Miotti P, et al. Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries. Lancet 2006 Mar 11;367(9513):817-824.

(47) World Health Organization. Treatment of tuberculosis: guidelines for national programmes. 3rd ed. Geneva, Switzerland: World Health Organization; 2003.

Figure 1 Distribution of age (in years) at ART initiation in TB and no-TB group in the original sample and PS matched pairs for a single MI dataset. From top to bottom: quantile-quantile plots, empirical cumulative distribution functions (circles represent TB and triangles represent no-TB), nonparametric density curves (solid line represents TB and dashed line represents no-TB), and side-by-side boxplots).

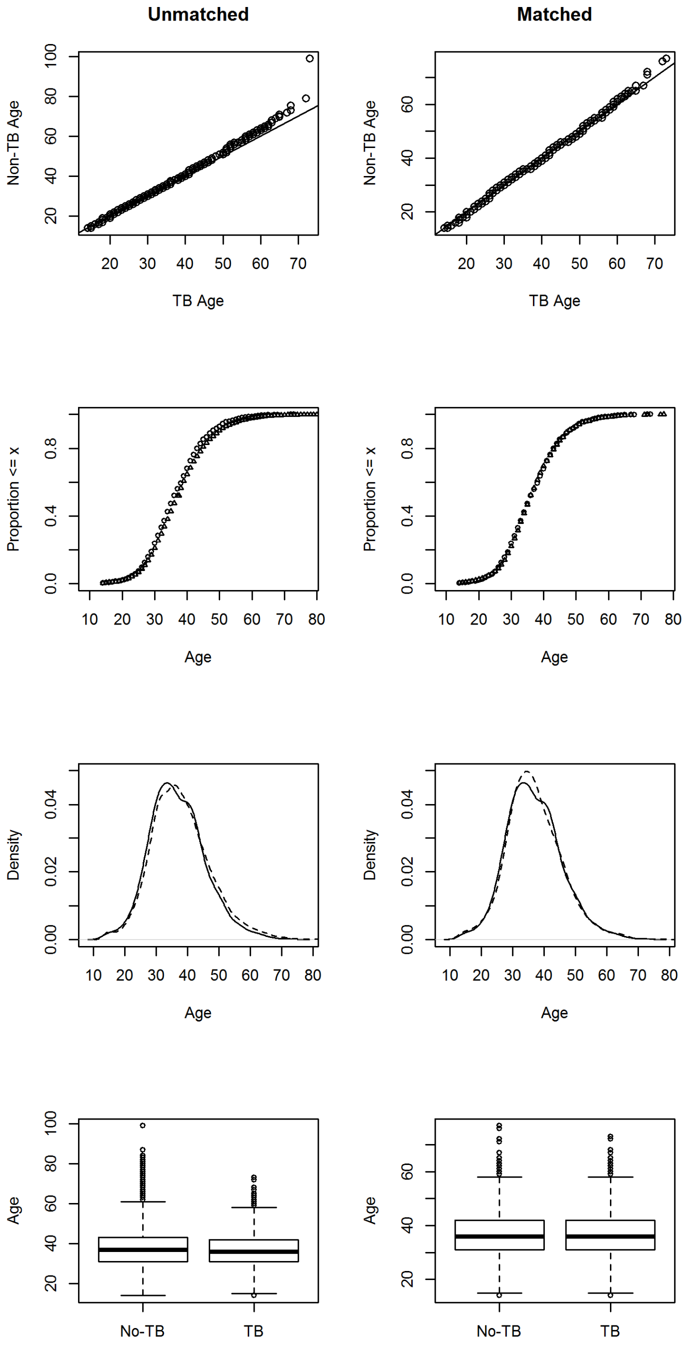


Figure 2 Distribution of CD4 cell counts at ART initiation in TB and no-TB group in the original sample and PS matched pairs for a single MI dataset. From top to bottom: quantile-quantile plots, empirical cumulative distribution functions (circles represent TB and triangles represent no-TB), nonparametric density curves (solid line represents TB and dashed line represents no-TB), and side-by-side boxplots.

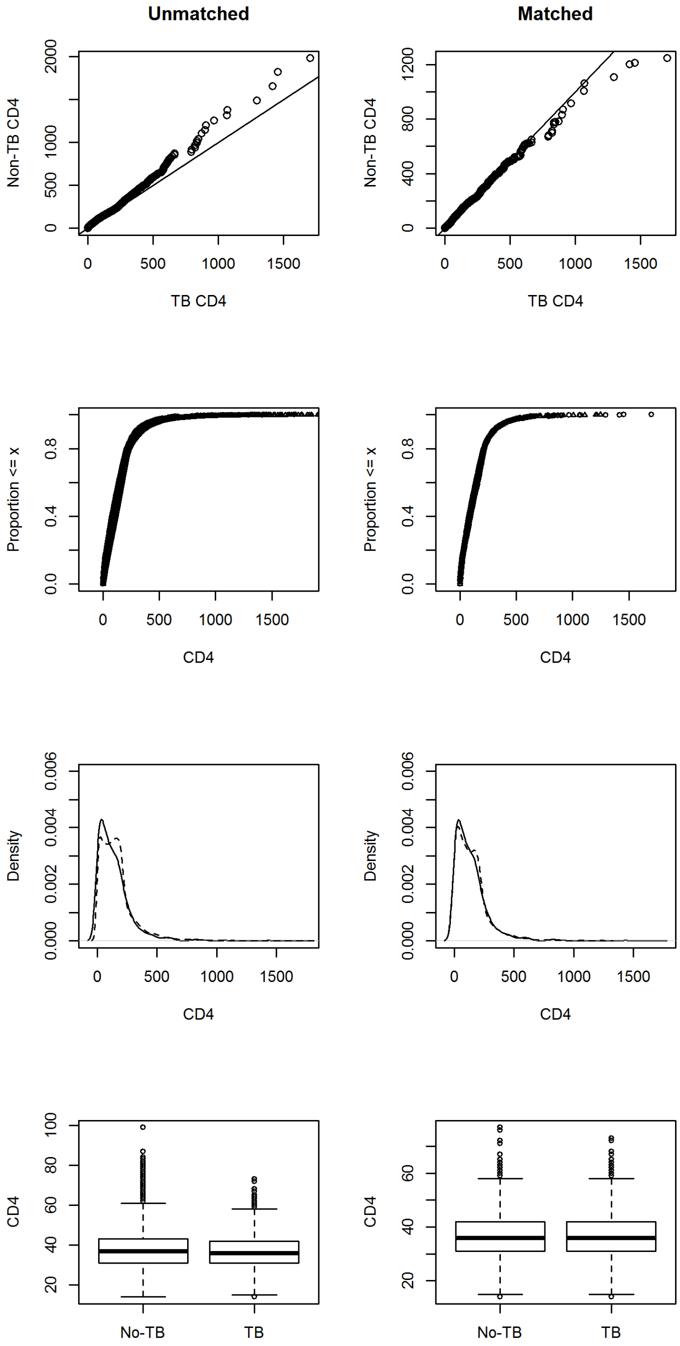


Figure 3 Kaplan–Meier survival curves by baseline TB status in the unmatched study sample (top: 1a) and in the propensity score matched pairs for a single MI dataset (bottom: 1b)

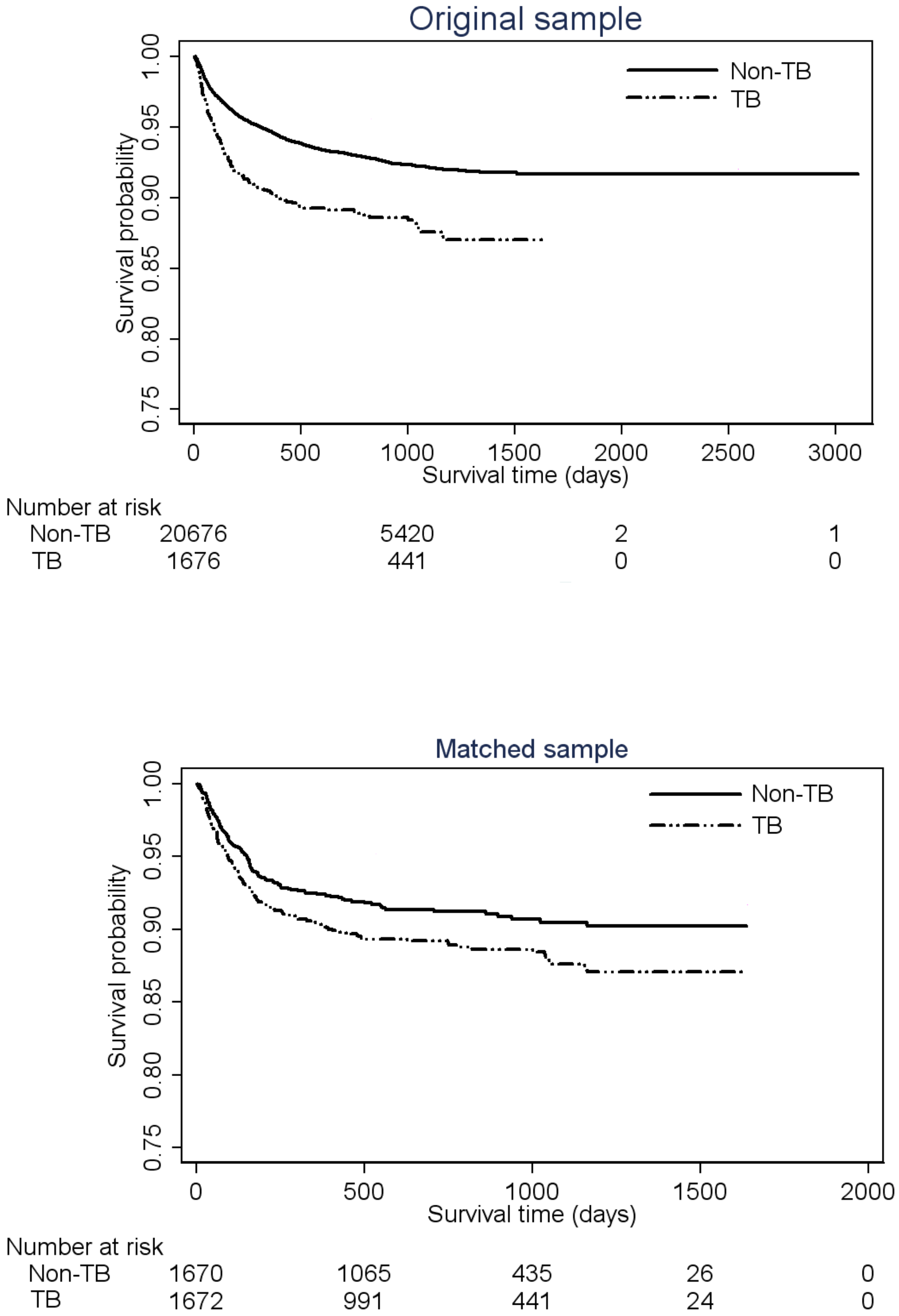


Table 1 Comparison of baseline and follow-up characteristics between HIV positive patients with and without co-infection of tuberculosis (TB) in the unmatched sample with imputed covariates (for a single MI dataset)

Characteristics	TB: yes (n=1690)	TB: no (n=20787)	P-value	Standardized difference (Ratio of variance for continuous covariate)
Demographics				
Age (year)				
Mean (SD)	36.87 (8.78)	37.93 (9.48)	<0.0001	-0.117 (0.86)
Min	14	14		
25th percentile	31	31		
50th percentile	36	37		
75th percentile	42	43		
Max	73	99		
Male: n (%)	670 (39.64)	6216 (29.90)	<0.0001	0.206
HIV prognosis				
CD4 count				
Mean (SD)	143.39	170.82	<0.0001	-0.170 (0.76)
Min	0	0		
25th percentile	45	66		
50th percentile	111	139		
75th percentile	193	208		
Max	1701	1983		
WHO stage:				
n (%)	43 (2.54)	992 (4.77)	<0.0001	-
1	441 (26.09)	11013 (52.98)		
2	999 (59.11)	7173 (34.51)		
3	207 (12.25)	1609 (7.74)		
4				
AIDS*: n (%)*	493 (29.17)	3881 (18.67)	<0.0001	0.248
History of drug use & illnesses: n (%)				
Pneumocystis jeroveci pneumonia	28 (1.66)	101 (0.49)	<0.0001	0.114
Toxoplasmosis	11 (0.65)	101 (0.49)	0.3542	0.022
Sexually transmitted infection	341 (20.17)	4353 (20.94)	0.4578	-0.019
Partner information: n (%)				

Married poly	125 (7.40)	1801 (8.66)	0.0734	-0.047
Partner sero positive	417 (24.67)	5801 (27.91)	0.0043	-0.073
Sexually active	1224 (72.43)	15272 (73.47)	0.3508	-0.023
Other characteristics: n (%)				
Site				
ENT	378 (22.37)	1950 (9.38)	<0.0001	-
GUL	161 (9.53)	1891 (9.10)		
JIN	276 (16.33)	2586 (12.44)		
MAS	130 (7.69)	2246 (10.80)		
MBL	144 (8.52)	2446 (11.77)		
MBR	116 (6.86)	2641 (12.71)		
MSD	61 (3.61)	1254 (6.03)		
MUL	211 (12.49)	2390 (11.50)		
SOR	29 (1.72)	1546 (7.44)		
TOR	184 (10.89)	1837 (8.84)		
Education (Higher institute)	58 (3.43)	623 (3.00)	0.3158	0.025
Start year since				
2000	0 (0)	2 (0.01)	<0.0001	-
0	23 (1.36)	946 (4.55)		
4	453 (26.80)	4532 (21.80)		
5	373 (22.07)	3627 (17.45)		
6	507 (30.00)	6715 (32.30)		
7	330 (19.53)	4943 (23.78)		
8	4 (0.24)	22 (0.11)		
9				
Follow-up				
Death: n (%)	177 (10.47)	1326 (6.38)	<0.0001	0.148
Median length of follow-up (days)	627.5	646	-	-

Note: For continuous variables, mean and standard deviation for each group, standardized mean difference and ratio of variances between TB and no-TB patients are reported. For categorical variables, frequency and percentage of the specified level, and standardized difference between TB and no-TB patients are reported, unless noted otherwise. MI: multiple imputation, SD: standard deviation, WHO: World Health Organization, HIV: human immunodeficiency virus, AIDS: acquired immune deficiency syndromes, ENT: Entebbe, JIN: Jinja, MAS: Masaka, MBL: Mbale, MBR: Mbarara, MUL: Mulago, TOR: Tororo, GUL: Gulu, SOR: Soroti, MSD: Masindi

Table 2 Comparison of baseline and follow-up characteristics between HIV positive patients with and without co-infection of tuberculosis (TB) in propensity score matched pairs with imputed covariates (for a single MI dataset)

Covariate	TB: yes (n= 1686)	TB: no (n= 1686)	Standardized difference (Ratio of variance for continuous covariate)
Demographics			
Age (year)			
Mean (SD)	36.89 (8.77)	36.93 (8.87)	-0.005 (0.978)
Min	14	14	
25th percentile	31	31	
50th percentile	36	36	
75th percentile	42	42	
Max	73	77	
Male: n (%)	668 (39.62)	673 (39.92)	-0.606
HIV prognosis			
CD4 count			
Mean (SD)	143.70	144.34	-0.004
Min	(149.98)	(139.78)	(1.15)
25th percentile	0	0	
50th percentile	45	44	
75th percentile	112	118	
Max	194	196	
	1701	1247	
WHO stage:			
n (%)	43 (2.55)	43 (2.55)	-
1	441 (26.16)	436 (25.86)	
2	996 (59.07)	983 (58.30)	
3	206 (12.22)	224 (13.29)	
4			
AIDS*: n (%)*	489 (29.00)	479 (28.41)	0.013
History of drug use & illnesses: n (%)			
Pneumocystis jeroveci pneumonia	25 (1.48)	33 (1.96)	-0.037
Toxoplasmosis	11 (0.65)	8 (0.47)	0.024
Sexually transmitted infection	340 (20.17)	339 (20.11)	0.001

Partner information: n (%)			
Married poly	125 (7.41)	136 (8.07)	-0.024
Partner sero positive	417 (24.73)	431 (25.56)	-0.019
Sexually active	1221 (72.42)	1223 (72.54)	-0.003
Other characteristics: n (%)			
Site			
ENT	374 (22.18)	378 (22.42)	-
GUL	161 (9.55)	176 (10.43)	
JIN	276 (16.37)	271 (16.07)	
MAS	130 (7.71)	125 (7.41)	
MBL	144 (8.54)	160 (9.49)	
MBR	116 (6.88)	121 (7.18)	
MSD	61 (3.62)	62 (3.68)	
MUL	211 (12.51)	203 (12.04)	
SOR	29 (1.72)	18 (1.07)	
TOR	184 (10.91)	172 (10.20)	
Education (Higher institute)	58 (3.44)	59 (3.50)	-0.003
Start year since			
2000	23 (1.36)	15 (0.89)	-
4	452 (26.81)	428 (25.39)	
5	372 (22.06)	360 (21.35)	
6	505 (29.95)	548 (32.50)	
7	330 (19.57)	330 (19.57)	
8	4 (0.24)	5 (0.30)	
9			
Follow-up			
Death: n (%)	176 (10.44)	137 (8.13)	0.080
Median length of follow-up	629	659	-

Note: For continuous variables, mean and standard deviation for each group, standardized mean difference and ratio of variances between TB and no-TB patients are reported. For categorical variables, frequency and percentage of the specified level, and standardized difference between TB and no-TB patients are reported, unless noted otherwise. MI: multiple imputation, SD: standard deviation, WHO: World Health Organization, HIV: human immunodeficiency virus, AIDS: acquired immune deficiency syndromes, ENT: Entebbe, JIN: Jinja, MAS: Masaka, MBL: Mbale, MBR: Mbarara, MUL: Mulago, TOR: Tororo, GUL: Gulu, SOR: Soroti, MSD: Masindi

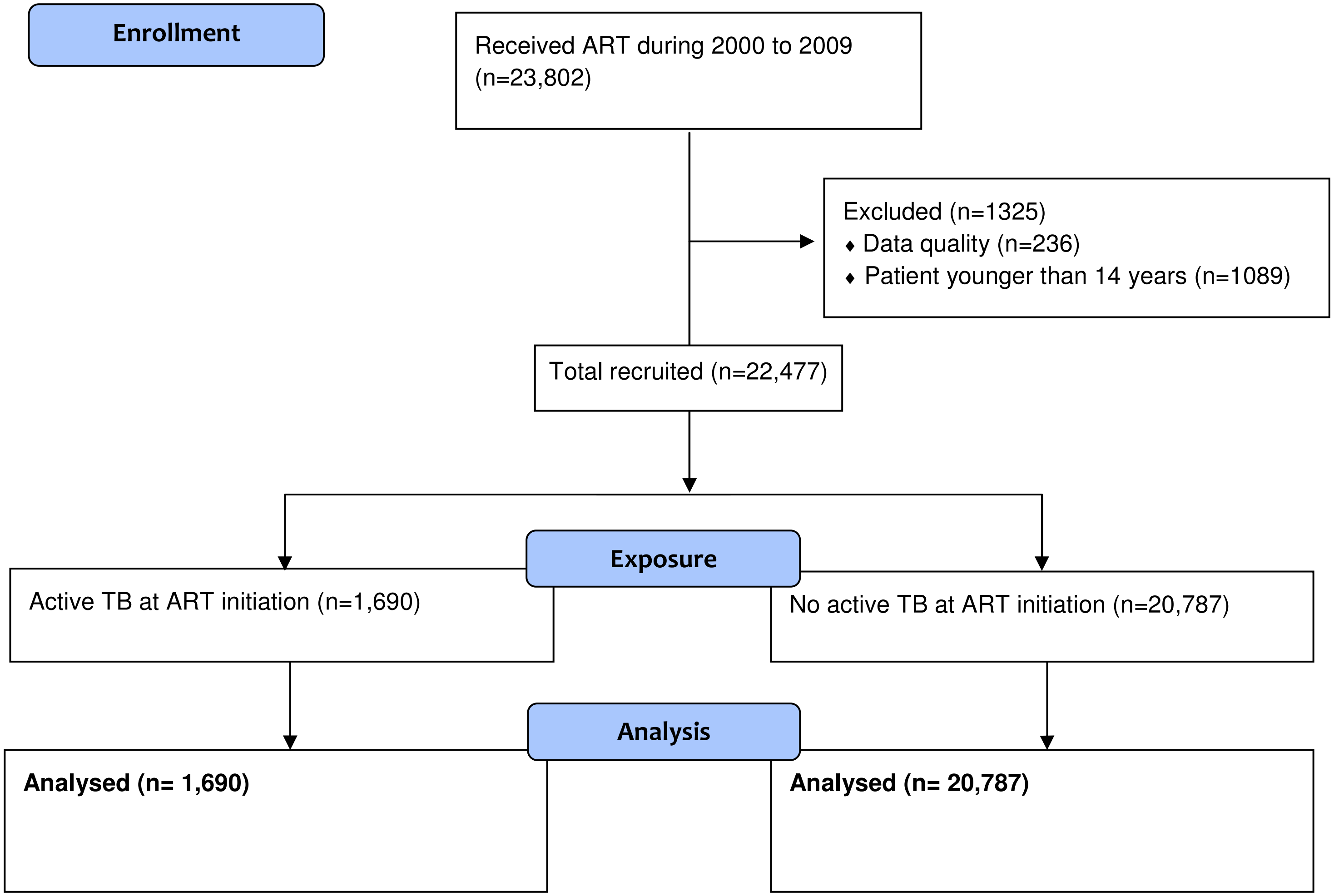
Table 3 Effect of TB at the initiation of ART on overall survival. Results for the PS or covariates adjusted Cox regression models were aggregated across 5 multiple imputed datasets using Rubin's rule.

Method	**HR**	**95% CI**	**P-value**
Primary analysis			
Matching on PS	1.37	1.08, 1.75	0.011
Sensitivity analyses			
Unadjusted Cox regression	1.74	1.49, 2.04	<0.001
Stratified on PS	1.36	1.15, 1.60	<0.001
PS as covariate (linear, quadratic and cubic terms)	1.34	1.14, 1.58	<0.001
Conventional Adjusted Cox regression	1.40	1.19, 1.65	<0.001

HR: hazard ratio; CI: confidence interval; PS: propensity score.

Supplementary files

Figure S1 Study flowchart


Enrollment
Received ART during 2000 to 2009 (n=23,802)
Excluded (n=1325)
♦ Data quality (n=236)
♦ Patient younger than 14 years (n=1089)
Total recruited (n=22,477)
Exposure
Active TB at ART initiation (n=1,690)
No active TB at ART initiation (n=20,787)
Analysis
Analysed (n= 1,690)
Analysed (n= 20,787)

Results and performance of the final propensity score (PS) model

Table S1 Results of the final PS model (logistic regression)

Parameter	**Log OR**	**SE**	**P-value**
Intercept	-10.1492	203.8	0.9603
Log CD4	-0.2434	0.1644	0.1388
Age	0.0542	0.0245	0.0271
Start year 2009	7.8238	203.8	0.9694
Start year 2008	6.9890	203.8	0.9726
Start year 2007	7.0638	203.8	0.9723
Start year 2006	7.2817	203.8	0.9715
Start year 2005	7.2678	203.8	0.9715
Start year 2004	5.9117	203.8	0.9769
Male	-1.8350	288.2	0.9949
Site TOR	-0.1025	0.1441	0.4770
Site SOR	-1.9095	0.3146	<0.0001
Site MUL	-0.4611	0.1417	0.0011
Site MSD	-1.0093	0.2234	<0.0001
Site MBR	-1.3222	0.1690	<0.0001
Site MBL	-0.8514	0.1572	<0.0001
Site MAS	-1.2095	0.1612	<0.0001
Site JIN	-0.3232	0.1350	0.0167
Site GUL	-0.4642	0.1459	0.0015
AIDS	1.2107	1.0281	0.2390
No higher institute education	-0.0769	0.2369	0.7455
Married mono	0.0785	0.1440	0.5858
Partner sero positive	-0.1566	0.1002	0.1181
Sexually transmitted infection	-0.0334	0.0892	0.7079
Pneumocystis jeroveci pneumonia	1.0783	0.3456	0.0018
Toxoplasmosis	0.4414	0.5596	0.4302
Sexually active	-0.1753	0.0858	0.0410
WHO stage 4	1.4352	0.9910	0.1476
WHO stage 3	1.7945	0.9313	0.0540
WHO stage 2	0.4028	0.9558	0.6734
Log CD4*Male	0.0236	0.0424	0.5777
Age*Male	-0.00735	0.00639	0.2498
Start year 2009*Male	4.3433	288.2	0.9880
Start year 2008*Male	3.2245	288.2	0.9911
Start year 2007*Male	3.0723	288.2	0.9915
Start year 2006*Male	3.0641	288.2	0.9915
Start year 2005*Male	3.0068	288.2	0.9917
Start year 2004*Male	3.0180	288.2	0.9916

Site TOR*Male	-0.5218	0.2065	0.0115
Site SOR*Male	0.0825	0.4017	0.8372
Site MUL*Male	-0.6340	0.2041	0.0019
Site MSD*Male	-0.4321	0.3002	0.1501
Site MBR*Male	-0.1690	0.2331	0.4686
Site MBL*Male	-0.1867	0.2194	0.3949
Site MAS*Male	-0.3524	0.2260	0.1190
Site JIN*Male	-0.4547	0.1880	0.0156
Site GUL*Male	-0.4697	0.2172	0.0306
AIDS*Male	0.1107	0.1377	0.4214
No higher institute education * Male	-0.0175	0.2958	0.9528
Married mono * Male	0.1036	0.2050	0.6134
Partner sero positive * Male	0.0284	0.1329	0.8308
Sexually transmitted infection * Male	0.0101	0.1562	0.9484
Pneumocystis jeroveci pneumonia * Male	0.3997	0.4930	0.4174
Toxoplasmosis*Male	-0.5613	0.7571	0.4584
Sexually active*Male	0.0111	0.1398	0.9367
WHO stage 4*Male	-0.5251	0.3667	0.1521
WHO stage 3*Male	-0.3691	0.3358	0.2717
WHO stage 2*Male	-0.5178	0.3439	0.1322
(Log CD4)^{2}	-0.00151	0.00774	0.8458
(Age)^{2}	-0.00046	0.000218	0.0330
WHO stage 4 * Log CD4	0.1695	0.1603	0.2903
WHO stage 3* Log CD4	0.1881	0.1545	0.2235
WHO stage 2* Log CD4	0.1932	0.1584	0.2227
WHO stage 4 * Age	-0.0241	0.0187	0.1976
WHO stage 3 * Age	-0.0320	0.0168	0.0573
WHO stage 2 * Age	-0.0299	0.0173	0.0843
Log CD4 * AIDS	-0.00706	0.0445	0.8739
Age * AIDS	0.00156	0.00712	0.8268
Start year 2009 * AIDS	-1.7342	1.7115	0.3109
Start year 2008 * AIDS	-0.6642	0.7857	0.3979
Start year 2007 * AIDS	-0.7006	0.7817	0.3701
Start year 2006 * AIDS	-0.9289	0.7852	0.2368
Start year 2005 * AIDS	-0.5780	0.7811	0.4593
Site TOR * AIDS	-0.6444	0.2459	0.0088
Site SOR * AIDS	-0.1947	0.4644	0.6750
Site MUL * AIDS	-0.0822	0.2005	0.6818
Site MSD * AIDS	-0.4462	0.3279	0.1735
Site MBR * AIDS	-0.0858	0.2415	0.7222
Site MBL * AIDS	-0.4701	0.2434	0.0535

Site MAS * AIDS	-0.4706	0.2540	0.0639
Site JIN * AIDS	-0.3984	0.1989	0.0452
Site GUL * AIDS	-0.2182	0.2683	0.4161
No higher institute education * AIDS	0.3807	0.3528	0.2805
Married mono * AIDS	-0.0368	0.2279	0.8719
Partner sero positive * AIDS	-0.1394	0.1504	0.3542
Sexually transmitted infection * AIDS	-0.0282	0.1446	0.8455
Pneumocystis jeroveci pneumonia * AIDS	0.0779	0.4706	0.8685
Toxoplasmosis * AIDS	-0.5476	0.7043	0.4369
Sexually active * AIDS	0.2564	0.1425	0.0719
WHO stage 4 * AIDS	-0.5620	0.4230	0.1840
WHO stage 3 * AIDS	-0.7115	0.4019	0.0767
WHO stage 2 * AIDS	-0.2125	0.4104	0.6046

Log OR: log odds ratio, SE: standard error, p-value: Wald Chi-squared p-value, WHO: World Health Organization, HIV: human immunodeficiency virus, AIDS: acquired immune deficiency syndromes, ENT: Entebbe, JIN: Jinja, MAS: Masaka, MBL: Mbale, MBR: Mbarara, MUL: Mulago, TOR: Tororo, GUL: Gulu, SOR: Soroti, MSD: Masindi

Table S2 Model Fit Statistics for the PS model

Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	11998.265	10913.351
SC	12006.285	11651.214
-2 Log L	11996.265	10729.351
Convergence criterion (GCONV=1E-8) satisfied.		

AIC: Akaike information criterion, SC: Schwarz Criterion, Log L: log likelihood

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**P-value**
Likelihood Ratio	1266.9134	91	<0.0001
Score	1363.4461	91	<0.0001
Wald	1080.6807	91	<0.0001

DF: degree of freedom

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**P-value**
10.9856	8	0.2025

DF: degree of freedom

CHAPTER 5

CONCLUSIONS

There are many clinical and methodological issues associated with baseline covariate adjustment in clinical trials. A subset of such challenges include: (1) analyzing correlated outcomes in multicentre randomized controlled trials (RCTs), (2) evaluating the probability and implication of baseline prognostic imbalance in RCTs, and (3) estimating the effect of tuberculosis (TB) on all-cause mortality in a large prospective cohort of HIV patients who received antiretroviral therapy (ART). We have studied these important topics in a manuscript-based thesis, with each chapter dedicated to investigating each of the issues. In this chapter, we summarize research findings, discuss their implications and limitations, and shed light for future investigation.

In Chapter 2, we conducted a simulation study to compare 6 statistical methods (4 patient-level models: ignoring centre effects, including centres as fixed effects, including centres as random effects, generalized estimating equation (GEE); and 2 centre-level models: fixed- and random-effects centre-level analysis) that are commonly used for estimating treatment effects on a continuous outcome in multicentre RCTs. We considered three designs with equal or varying numbers of patients per centre and a 1-to-1 randomization ratio, in the absence of treatment-by-centre interaction. We found that all 6 models yielded unbiased point estimates of the treatment effect over a wide spectrum of intra-class (or intra-centre) correlation (ICC) values when the numbers of patients randomized to the two treatment groups were equal or subject to chance imbalance. Ignoring centre effects or intra-centre correlation did not bias the estimation of treatment effect for even large ICC values. This is largely because when treatments are allocated in the same proportion in all centres (or subject to change imbalance only), centre has no association with the treatment allocation, hence adjusting for centre effect or not has little impact on the point estimate of the treatment - response relationship given that the response variable is a continuous. Yet, the models produced different standard errors (SEs) for the estimated treatment effects in several scenarios, which subsequently led to different confidence interval (CI) estimates. Our simulation study demonstrated the advantage of treating centres as random intercepts in the absence of treatment-by-centre interaction. Overall, the random-intercept model produced the most precise effect estimates among the methods investigated and attained nominal values for CI coverage and statistical power, under all simulated circumstances. The fixed-intercept model demonstrated extremely similar statistical properties (in terms of bias, precision, CI coverage and power) relative to the random-intercept model in balance design; this method was less efficient when the study was composed of many centres (20+) each recruiting a few patients. When the number of centres was below 40, the GEE method tended to slightly underestimate the SE, subsequently resulting in greater statistical power (i.e. the treatment effect estimate was more likely to be statistically significant with a smaller SE) and lower CI coverage. We also encountered non-convergence problem when running GEE algorithm when centre sizes were highly variable. When one failed to control for the potential centre effects in any form, as in a regular two-sample t test, the SE of the estimated treatment effect was severely overestimated for large ICC values greater than or equal to 0.2, leading to falsely high CI coverage and a decrease of statistical power. The centre-level models on average yielded larger SEs, lower coverage or statistical power relative to the patent-level models. Their incapability to adjust for patient-level covariates was a major drawback when prognostic imbalance was likely.

This project complements the ICH E9 guideline (1), by studying the impact of ICC on the evaluation of treatment effects, a practical challenge faced by many trialists yet inadequately discussed in the literature. The key implications on the analysis and design of multicentre RCTs include the following. First, the sample size needs to be increased by an inflation factor of 1/(1-ICC) to account for within centre clustering, if an initial sample size evaluation and intended primary analysis employ a two-sample t-test. It is important to obtain an approximate estimate of ICC for the primary outcomes either through literature review or by conducting a pilot study. Second, permutated block randomization should be conducted when feasible, so that the treatment allocation proportions across centres are approximately the same and the treatment contrast and centre effects can be estimated independently and efficiently. Third, random-intercept models can recover inter-centre information in an unbalanced design and produce an accurate and more precise estimate of the treatment effect than the t-test and fixed-intercept model, when a large number of participating centres enroll only a few patients. Forth, centre sizes for the majority of the centres should be sufficiently large to ensure reliable estimation of the within-centre variation, especially when the ICC value is unknown or possibly large. Although we observed similar results on simulations involving equal and varying centre sizes, similar centre sizes may help achieve convenience of the GEE approach when the value of ICC is below 0.1 or beyond 0.4.

A limitation of this project is preclusion of treatment-by-centre interaction in the simulations. Such interaction may exist in practice, due to differences in patient population or variability in patient care among the participating centres, yet can be difficult to detect because of the lack of statistical power. Methodological challenges include the development of a statistical measure for the overall treatment effect across trials sites, the focus of statistical inference (individual treatment effects for participating sites versus an aggregated effect across sites), statistical and computational performance of the analytic models in various circumstances (balanced or unbalanced design with equal or highly variable centre sizes and different ICC values). Future theatrical and simulation studies are needed.

In Chapter 3, our simulation study demonstrated that random error or chance plays an important role in the occurrence of prognostic imbalance and the estimation of treatment effects in individual RCTs. Simulation results showed that small sample size was associated with a high risk of imbalance in prognostic factors (PFs) in a particular trial. The probabilities of an absolute imbalance ≥ 5% in a binary PF of prevalence 0.5 were estimated 0.42, 0.62 and 0.67 with 125, 50 and 25 patients per treatment arm, respectively. The probability of absolute imbalance decreased when sample size increased or prevalence of PF approached 0 or 1. Our results based on a single binary baseline PF suggested that it is essential to adjust for important PFs in trials evaluating a binary outcome. Ignoring PFs of high predictive values in the analysis would lead to severe bias or loss of statistical power, due to non-collapsibility (2) or chance confounding. When the PF was less powerful or a treatment difference did not exist, improvement in accuracy and efficiency associated with the adjustment for the PF was less noticeable.

It was challenging to establish a single rule for sample size requirement based on the probability and impact of prognostic imbalance. Our study added to the current literature on what constitutes an adequate sample size to control against potential impact of prognostic imbalance, and demonstrated that the adequacy varies with multiple factors, including the choice of imbalance measure, the size of imbalance deemed important, one's tolerance of the random error around the estimated treatment effects, and the prevalence of the PF. The proposed tool for sample size requirement not only helps to design clinical trials, but is also useful to assess the quality of completed trials by evaluating the likelihood and impact of potential risk of chance confounding.

The probabilities of prognostic imbalance can also be calculated mathematically based on the difference between two independent and identically distributed binomial variables, each representing the number of patients who have the baseline prognostic factor in a treatment group. Computer program is likely to be needed to obtain numerical values of the mathematical expression given the large sample sizes being investigated in Chapter 3. The simulation technique employed in the study allows us to address multiple research objectives in a unified framework while maintaining validity of the scientific investigation. This project had the following limitations. First, only one binary baseline PF was considered in the simulation. The balancing distributions of multiple correlated PFs between treatment groups needs to be assessed and their impact on effect estimation in RCTs is warranted. Second, systematic reviews and meta-analyses face the similar methodological challenges on prognostic balancing. Future work is needed to assess the impact of imbalance on obtaining an aggregated estimate of treatment effects in meta-analyses, where the cumulative number of patients from individual RCTs and between-study variation need to be taken into consideration.

In Chapter 4, we assessed the relationship between having active tuberculosis (TB) at the initiation of antiretroviral therapy (ART) and overall survival among 22,477 adult HIV patients who received ART during August 2000 to June 2009, in Uganda, Africa. At the beginning of ART (baseline), 1,690 (7.52%) HIV patients were identified as having TB. At baseline, TB patients were more likely to be male, have AIDS defining illnesses, belong to WHO disease stage III or IV, and have lower CD4 cell counts, compared with the no-TB patients. To reduce bias and improve estimation stability and efficiency, we applied multiple imputation (MI) procedure to impute missing values for 4 important baseline covariates (age, CD4 cell count, WHO disease stage and sexual activity), and performed all statistical analyses on 5 MI datasets. The final propensity score model (PS) included 14 baseline covariates, their pairwise interactions with gender and baseline AIDS status, and additional quadratic and interaction terms for the continuous variables to improve model fit. We managed to match 1,686 among all 1,690 baseline TB patients to an equal number of no-TB patients. The similarity of the distributions of all baseline covariates between the TB matched pairs was improved substantially.

We estimated the hazard ratio for all-cause mortality to be 1.37 (95% CI: 1.08-1.75), comparing TB and no-TB patients on the 1,686 PS matched pairs with similar disease prognosis at baseline. This indicated having TB at the initiation of ART was associated with a 37% increase in the instantaneous risk of death, over a median of 21.5 months of follow-up. Stratification and covariate adjustment for the PS and controlling for multiple baseline covariates using Cox regression yielded similar results (PS-stratified HR=1.36, 95% CI: 1.15-1.60; PS-adjusted HR=1.34, 95% CI: 1.14 -1.58; adjusted Cox HR= 1.40, 95% CI: 1.19-1.65), all less marked than the crude estimate (HR=1.74, 95% CI: 1.49-2.04).

This project, to our knowledge, involved a large research cohort of HIV-infected patients receiving ART in a single African country that evaluated TB outcomes in co-infected patients. Results suggested that TB was less prevalent in Uganda compared with other

African countries (3,4). Our finding of the 37% increase in hazard of all-cause mortality associated with having active TB at the initiation of ART after controlling for baseline confounding is complementary to and strengthens results of the recent SAPIT, CAMELIA, and STRIDE trials (5-7) that demonstrated significant improvement on survival when ART was initiated during TB therapy. Our results validate the WHO guidelines that urge a more aggressive approach to management of both TB and HIV (8) .

In this project, we described a PS-based framework to estimating the causal effect of an exposure variable (that cannot be manipulated in the study) on a clinical outcome using an observational prospective cohort design. When the sample size or the number of outcome events is small relative to the proportion subjects being exposed, the conventional statistical adjustment approaches may not provide adequate control against important confounding. Stratifying on or adjusting for a large number of confounders during the estimation of the exposure-outcome relationship may create sparse data problem and subsequently affect model stability or efficiency. Non-overlapping supports of the confounders in the comparison groups can lead to invalid results. Comparability of the confounding variables between the exposed and unexposed groups post covariate adjustment can be difficult to examine. The PS methods are a powerful alternative to estimate the average exposure effects on the whole population or subpopulations with observed datasets. In such methods, the vector of potential confounding variables reduces to a single score that reflects one's propensity of being exposed. As illustrated in Chapter 4, this provides relatively easy means to compare the distributions of individual potential confounders and the support of the PS between exposure groups. Sensitivity analysis is a useful tool to examine the validity and variability of statistical inference when only empirical data are available. Results of the supportive analyses showed our study findings were robust to different modeling options.

Limitations of this project emerged from three key sources. First, our results in Chapter 4 are predominantly based on adult HIV patients who initiated ART during 2004 to 2009 in Uganda. Information of TB treatment was not available in the dataset, hence limiting the generalizability of the results. The effects of active TB on mortality for HIV patients with and without TB treatment, in different age or ethnic groups, and from different geographic regions warrant further investigation. Second, PS methods similar to most other analytical approaches, cannot balance unobserved confounders in estimating a causal relationship. Our results may be biased if some unobserved prognostic factors were unbalanced between TB and no-TB groups. More resources need to be dedicated to improve the completeness and accuracy of data collection in HIV patients in underdeveloped regions. Statistical tools such as the instrumental variable analysis can also be explored to handle bias due to unobserved confounders (9). Lastly, the multiple imputation (MI) method employed to impute the missing values for baseline covariates requires the missing at random (MAR) assumption (i.e. the missingness can be explained by differences in the observed data) (10). We could not rule out the existence of additional variables that were highly associated with the missingness. Because the MAR assumption is unverifiable using data collected within a study, simulation studies may be needed to study the performance of MI and the determinants of MI behaviour, when missingness is not at random. The more sophisticated adjustment methods can make an impact on clinical research, only if user friendly software is accessible for the applied biostatisticians and clinical researchers.

In summary, this PhD dissertation identified and investigated issues of covariate adjustment in late phase clinical trials. The three papers make contributions by exploring the phenomena of imbalance in baseline covariates and intracentre correlation in randomized and observational studies. The impact of each phenomenon on the estimation of the intervention effects has been carefully investigated. The performances of multiple routine statistical methods have been compared under various design parameters and using an observational dataset. Overall we showed that baseline covariates contain useful information and should be taken into consideration in the analysis and review of clinical trials. Interaction between baseline covariates and the intervention in individual studies meta-analyses warrants further investigation.

References

(1) International Conference on Harmonisation E9 Expert Working Group. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Stat Med 1999 Aug 15;18(15):1905-1942.

(2) Gail MH, Weiand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71(3):431-444.

(3) Westreich D, MacPhail P, Van Rie A, Malope-Kgokong B, Ive P, Rubel D, et al. Effect of pulmonary tuberculosis on mortality in patients receiving HAART. AIDS 2009 Mar 27;23(6):707-715.

(4) Lawn SD, Kranzer K, Edwards DJ, McNally M, Bekker LG, Wood R. Tuberculosis during the first year of antiretroviral therapy in a South African cohort using an intensive pretreatment screening strategy. AIDS 2010 Jun 1;24(9):1323-1328.

(5) Abdool Karim SS, Naidoo K, Grobler A, Padayatchi N, Baxter C, Gray A, et al. Timing of initiation of antiretroviral drugs during tuberculosis therapy. N Engl J Med 2010 Feb 25;362(8):697-706.

(6) Blanc FX, Sok T, Laureillard D, Borand L, Rekacewicz C, Nerrienet E, et al. Earlier versus later start of antiretroviral therapy in HIV-infected adults with tuberculosis. N Engl J Med 2011 Oct 20;365(16):1471-1481.

(7) Havlir DV, Kendall MA, Ive P, Kumwenda J, Swindells S, Qasba SS, et al. Timing of antiretroviral therapy for HIV-1 infection and tuberculosis. N Engl J Med 2011 Oct 20;365(16):1482-1491.

(8) World Health Organization. Treatment of tuberculosis: guidelines for national programmes. 3rd ed. Geneva, Switzerland: World Health Organization; 2003.

(9) Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive

cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007 Jan 17;297(3):278-285.

(10) Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2002.