THE USE OF SENSITIVITY ANALYSES IN HEALTH RESEARCH

VERIFICATION, COMPARISON AND EXPLORATION: THE USE OF SENSITIVITY ANALYSES IN HEALTH RESEARCH

By

JI (EMMY) CHENG, B.A (Honours), B.Sc. (Honours), M.Sc. (Statistics)

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

McMaster University DOCTOR OF PHILOSOPHY (2016) Hamilton, Ontario (Health Research Methodology ─ Biostatistics Specialization)

TITLE: Verification, Comparison and Exploration: The Use of Sensitivity Analyses In Health Research

AUTHOR: Ji (Emmy) Cheng, B.A. (University of International Business and Economics, Beijing, China), B.Sc. (McMaster University, Ontario, Canada), M.Sc. (McMaster University, Ontario, Canada)

SUPERVISOR: Dr. Lehana Thabane

NUMBER OF PAGES: xiv, 146

ABSTRACT

Background and Objectives:

I investigated the use of sensitivity analyses in assessing statistical results or analytical approaches in three different statistical issues: (1) accounting for within-subject correlations in analyzing discrete choice data, (2) handling both-armed zero-event studies in meta-analyses for rare event outcomes, and (3) incorporating external information using Bayesian approach to estimate rare-event rates.

Methods:

Project 1: I empirically compared ten statistical models in analyzing correlated data from a discrete choice survey to elicit patient preference for colorectal cancer screening. Logistic and probit models with random-effects, generalized estimating equations or robust standard errors were applied to binary, multinomial or bivariate outcomes.

Project 2: I investigated the impacts of including or excluding both-armed zero-event studies on pooled odds ratios for classical meta-analyses using simulated data. Five commonly used pooling methods: Peto, Mantel-Haenszel fixed/random effects and inverse variance fixed/random effects, were compared in terms of bias and precision.

Project 3: I explored the use of Bayesian approach to incorporate external information through priors to verify, enhance or modify the study evidence. Three study scenarios were derived from previous studies to estimate inhibitor rates for hemophilia A patients treated with rAHF-PFM: 1) a single cohort of previously treated patients, 2) individual patient data meta-analysis, and 3) an previously unexplored patient population with limited data.

Results and Conclusion:

Project 1: When within-subject correlations were substantial, the results from different statistical models were inconsistent.

Project 2: Including both-armed zero-event studies in meta-analyses increased biases for pooled odd ratios when true treatment effects existed.

Project 3: Through priors, Bayesian approaches effectively incorporated different types of information to strengthen or broaden research evidence.

Through this thesis I demonstrated that when analyzing complicated health research data, it was important to use sensitivity analyses to assess the robustness of analysis results or proper choice of statistical models.

PREFACE

This thesis is a “sandwich thesis”, which combines three individual projects prepared for publication in peer-reviewed journals. The following are contributions of J. Cheng in all the papers included in this dissertation: developing the research ideas and questions; developing analysis plans; designing the simulations and programming the codes; conducting all the statistical analysis; preparing all figures and tables; writing all of the manuscripts; submitting the manuscripts; and responding to reviewers’ comments. The work in this thesis was conducted between Fall 2010 and Winter 2015.

The work of the first paper has been published. The send and third papers have been submitted to peer-reviewed journals.

ACKNOWLEDGEMENTS

To all the special people who talked me through this long journey: Without you, life would be different.

First, I would like to express my deepest gratitude to my supervisor, Dr. Lehana Thabane. Your expertise, understanding, patience and passion towards students guided my feet and lighted my path.

My special thanks to my committee members, Drs. Eleanor Pullenayegum, John Marshall and Alfonso Iorio. Your support, guidance, encouragement and friendship made huge a difference in my life.

I am grateful to my co-authors, Drs. Deborah Marshall, Vadim Romanov and Maura Marcucci, who collaborated with me by providing data and expertise.

Last but not the least, I would like to thank my friends: Your everlasting love is the source of my joy and success.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 4

LIST OF FIGURES

Chapter 4

LIST OF APPENDICES

Page

Chapter 4

CHAPTER 1

INTRODUCTION

"Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."

-- David Sackett et al. BMJ. 1996[1]

Decision-making in patient care is an interactive process which integrates three components: clinical state, patient preference and research evidence [2]. As Sackett et al [1] note in the above quote, the use of the current best research evidence under the principle of evidence-based medicine (EBM) is involved in evidence generating, synthesizing, appraising and implementing. Throughout the entire research process, properly implementing statistical methods is crucial to ensure the use of appropriate study design, data analysis methods and reporting/interpretation of the results.

Statistical inference is the process of applying certain statistical procedures (models) to some collected data (sample) to generate a statistical property (evidence) which can be

generally applied to all unknown subjects (population) with similar characteristics[3–5]. All statistical models or analytical approaches are based on certain underlying statistical assumptions. Choosing appropriate statistical method(s) according to the research questions and distribution of data is important. Unlike simulated data, clinical research data are real world samples collected through health research projects under different types of designs which can be experimental, such as randomized controlled trials (RCTs), or non-experimental, also called observational data such as cross-sectional survey. When analyzing the collected study data, the data are assumed to follow certain distributions for certain statistical models, but the assumptions may not be perfectly true. This is why sometimes discrepancies between the results obtained from different statistical models or approaches can be found. Therefore, fully assessing the appropriateness of the choice of the statistical models or approaches is important to generate a reliable statistical inference which can later be confidently transferred to clinical evidence. However, choosing the most appropriate statistical model or analytical approach over its alternatives is not straightforward, particularly when the clinical data have a complex structure or represent a complicated clinical setting.

The assessment of the credibility of the statistical analysis results can be done through sensitivity analysis, an array of comparisons aiming to examine the consistency (robustness) and discrepancy (uncertainty) caused by various reasons such as model

choice or sample selection. In the Dictionary of Epidemiology[6], sensitivity analysis is defined as a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions. Regardless of the design or scope of the study, sensitivity is a useful tool in providing the statistical result in a comprehensive way, as has been discussed in guidelines or methodological papers for conducting observational study[7] , RCTs[8], knowledge synthesis[9]/meta-analysis[10] and health economic study[11]. In this thesis, I would discuss the use of sensitivity analysis on three unsolved issues regarding analyzing health research data.

The objectives of this thesis are: 1) addressing the challenges in analyzing health research data when no consensus of the statistical method is available; and 2) providing some resolutions on choosing statistical analysis approaches through the scope of sensitivity analyses. I compared alternative statistical models or analytical approaches for some unsolved or not fully investigated statistical problems by assessing the consistency and discrepancy of analysis results and their impacts on the implications. Three specific statistical issues examined are: 1) analyzing correlated discrete choice outcomes on eliciting patient preference; 2) dealing with both-armed zero-event studies in meta-analyses; and 3) the use of Bayesian statistical approach to incorporate external

information or existing literature with available data in generating estimates for rare diseases or events.

Issue 1: Within-subject correlation in discrete choice survey data

Discrete choice experiment (DCE) design originated in marketing research as a tool to differentiate consumer's choice among alternative products[12] and has been increasing used in the area of health economics and policy making to help researchers and policy makers to elicit patient and other stakeholder's preference for alternative healthcare programs or services[13–16]. In recent years, many researchers have dedicated their work to improve DCE methods by providing guidelines in health research[17,18]. However, compared to the attention given to the design aspects such as defining the key attributes, constructing the choice sets and administering the survey[19,20], there was not a lot of research addressing some analytical issues regarding the analysis of DCE data[14] with within-subject correlations, and the statistical methods used in analyzing clustered DCE data are inconsistent[16].

A typical DCE uses the factorial or partial factorial design principle to create a series (panel) of hypothetical choice scenarios (choice sets) to describe the attributes (or characteristics) and their associated levels for certain products or services[13]. The panel

of several choice sets is delivered through survey questionnaires, and each participant is asked multiple times to choose his/her preferred products or services over their alternatives. The fundamental assumption that DCE is based on is random utility theory (RUT) which assumes that any choice from any respondent is made by maximizing the utilities (or benefits)[21]. However, decision making is a complex process. The choices made by one person or group could be inter-related[12], particularly when the "opt-out" option is available[22]. Therefore, the within-group or with-subject correlations need to be accounted for in analyzing DCE data using proper statistical models[13,23,24].

Depending on the number of choice alternatives, basically all fixed or random effects statistical models for binary or multinomial outcomes can be used to analyze DCE data. A newly published review of DCE studies in the health economics field reported that the most used models in recent years (2009-2012) in analyzing DCE data were fixed-effect logit model (10%) and random-effects probit model (10%) for the designs with two choice alternatives and fixed-effect multinomial logit model (44%) for the designs with three or more choice alternatives[25]. Regarding adjusting potential correlations, health researchers paid more attention to two types of correlations that may occur in DCE data. One is the correlation among the choice alternatives, i.e. the violation of orthogonal design[23], which is dealt with typically by using probit models. Another is the so-called

preference heterogeneity, a type of correlations among certain groups of respondents, which is dealt with by using random-effects, nested or latent class models[15,26].

However, the within-subject correlation is largely ignored in the analysis of health research related DCEs. Although I found a few investigations of this problem that have been conducted in other research fields that used DCE designs such as marketing [27], and transportation engineering[28,29], the proposed solutions were mainly theoretical and thus difficult to adopt using common statistical analysis software.

Issue 2: Inconsistence in handling both-armed zero-event in meta-analysis

A systematic review (SR) synthesizes the available literature on a certain topic through a rigorous and systematic searching and selecting process using predefined inclusion and exclusion criteria, and meta-analysis (MA) quantitatively synthesize the evidence by pooling results of individual studies identified by the SR using statistical methods[30]. As a matter of principle, all studies with available data included in an SR need to be included in the MA. However, this principle cannot always be applied to the MA with binary outcomes when both-armed zero-event (BAZE) studies are among the identified studies[31].

BAZE study, also called zero-total event study, is a study that has no observed outcome event in both comparison groups, for example, treatment and control arms in a randomized control trial (RCT). Currently, there are no guidelines developed as to how to deal with BAZE studies in meta-analysis, and thus BAZE studies are handled inconsistently in the practice of conducting MA[32]. The analytical approaches may vary depending on the choice of the effect measure of statistical pooling methods and as well as the considerations or decisions of the researchers. Although several recently conducted simulation studies provided some statistical procedures to include BAZE studies[33–37], it remains unclear as to how including or excluding BAZE studies in or from MA may impact the accuracy.

Issue 3: How to incorporate external information to enhance, modify or compare the evidence presented in the observed data for rare event outcomes

Statistics is the essential tool to quantitatively summarize the evidence for the available study data. There are two main approaches in statistics: Frequentist and Bayesian. Unlike Frequentist which is also known as classical statistical approach, in which the hypothesis is tested based on the long-run frequency[38], Bayesian approach rooted on the Bayes' Theorem is a conditional probability which updates the current knowledge based on newly obtained data[39] and previous evidence. Although with its broadly educated base and easy-to-use software, classical statistics is dominant force in analyzing health

research data, Bayesian approach as the alternative has been increasingly used in many health research areas[40,41].

With the way of adaptively updating all available knowledge by incorporating the evidence from past (priors) to the current observations (data) to make prediction for the future (posterior estimates), Bayesian analyses naturally simulate how human brains process information and make decisions[42,43]. With the ability of combining the external information or historical events, Bayesian approach is more appealing when studies are conducted to investigate rare diseases or events[44]. An example to show the methodological or statistical challenges of using the classical approach to analyze rare event data is estimating inhibitor rate of the patients under hemophilia A treatment with the Factor VIII or IX replaced products. Hemophilia A is a rare blood disorder which occurs in 8 of 100,000 males in North America. An inhibitor which is an antibody to the product used to treat or prevent bleeding episodes, is considered to be a serious complication affecting 1-6% of hemophilia A patients[45]. A large systematic review (2013) summarized the rates of developing inhibitor among the previously treated hemophilia A patients: 43 inhibitors were reported in 4323 patients across 33 cohorts[46]. Due to the small sample size and extremely low event rate, the estimates of inhibitor rates from most individual studies presented huge uncertainties with unreasonably wide 95%

confidence intervals (CIs): 9 estimates reported the lower bounds of 95% CI as 0%; the widest 95% CI was between 0.6% and 23.5%.

With the challenges presented in analyzing rare event data, more researchers have turned to the Bayesian approach for solutions, in particular for assessing the robustness of the results by using different priors that incorporate relevant information from different sources. However, conducting Bayesian analyses can sometimes be complicated by programming Bayesian codes, properly choosing priors, setting up the likelihood function and interpreting the results. Therefore, to efficiently promote the use of the Bayesian approach among health researchers, more technical supports with examples need to be provided.

Summary of Chapters

In this sandwich thesis, the issues described above were investigated through three independent but inter-connected projects under the general topic of sensitivity analyses. The papers dedicated to these projects were separated in the next three chapters starting with Chapter 2.

In Chapter 2, I empirically compared the commonly used statistical models with the ability of adjusting within-subject correlate in analyzing DCE data. The data used in the project were collected through a survey conducted in Hamilton, Ontario, Canada in 2002. The aim of this survey was to elicit participant preference for colorectal cancer (CRC) screening tests. A two-staged DCE design with the opt-out option was used to investigate how six important attributes (process, pain, preparation, specificity, sensitivity and cost) which defined the four CRC screening tests could impact participants' choice of one test over its alternatives and their willingness to undertake the test. The choices made by the participants were organized in three ways: binary, multinomial and bivariate-binary outcomes. Six statistical models for analyzing clustered binary data were applied, which included logistic and probit regression with cluster-robust standard error (SE), random-effects logistic and probit models, and logistic and probit models using generalized estimating equation (GEE) approaches. For the multinomial outcomes, I fitted three models: multinomial logistic/probit models with clustered robust SE and random-effects multinomial logistic model. The bivariate probit model with clustered-robust SE was used to analyze bivariate-binary outcome which treated the choices in two stages as two correlated binary outcomes. The rank of relative importance of attributes and the magnitude of $\boldsymbol{\beta}$ were used to assess the model's robustness.

Chapter 3 is a simulation study in evaluating the impact of including or excluding bother-armed zero-event studies in meta-analysis of RCTs using rare event outcomes. The values of simulation parameters were chosen based on a review paper which summarized the characteristics of MAs in the Cochrane Database of Systematic Reviews. Some 2500 datasets were generated for a series of scenarios which represented the different settings of treatment effect, control arm event rate, number of patients of each individual trial and between study variance. I investigated five pooling methods using odds ratio (OR) as the effect measure for classical meta-analyses, namely Peto, Mantel-Haenszel (M-H) method with fixed-effects and random-effects model, and inverse variance (IV) method with fixed-effects and random-effects model. The above methods were applied to each simulated dataset using the approaches of including and excluding BAZE studies. With the focus of the potential bias of the treatment effect introduced when trials with both zero-event arms were included or excluded in the MAs, I assessed the performance of the above methods using percentage bias, root mean square error (RMSE), length of 95% confidence interval (CI), and coverage.

Chapter 4 is a methodological paper to explore the merits of using Bayesian approaches to generate evidence for complex clinical settings. This paper also serves as a tutorial for clinicians who are interested in this topic. I aim to illustrate how to adopt the Bayesian approach to analyze the current available data while incorporating the external

information for rare event rates. After introducing the concepts behind Bayesian inference, step by step, I showed the process of choosing non-informative and informative priors, comparing the results to thresholds and evaluating the impact of sample size in three study scenarios based on published papers which collectively investigated the inhibitor rate of hemophilia A patients treated with rAHF-PFM (ADVATE): 1) analyzing the inhibitor rate (a rare adverse event) in a single cohort of previously treated patients (PTPs)[46]; 2) meta-analyzing inhibitor rate by pooling a set of studies with individual patient level data[47]; and 3) generating evidence of inhibitor rate using very limited data for a previously unexplored patient population[48]. The individual patient level data used in this project were from PASS (Post-Authorization Safety Surveillance) studies provided by Baxter Healthcare, Global Affairs (Westlake Village, California, USA).

Chapter 5 summarized the findings of Chapter 2 to Chapter 4, and discusses the implications of the findings and the limitations. I hope to use this thesis to raise awareness among researchers regarding the importance of assessing the robustness of statistical analysis results through a range of sensitivity analyses by sharing our experience using real examples. The individual papers also provide some solutions or suggestions for certain statistical and methodological issues in health research field.

Reference

1 Sackett D, Rosenberg W, Muir Gray J, *et al.* Evidence based medicine : what it is and what it isn ' t. *Br Med J* 1996;**312**:71–2.http://www.ebscohost.com.au

2 Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *Vox Sang* 2002;**83 Suppl 1**:383–6. doi:10.1136/ebm.7.2.36

3 De Muth JE. Overview of biostatistics used in clinical research. *Am J Heal Pharm* 2009;**66**:70–81. doi:10.2146/ajhp070006

4 Kier KL, Ph D, Sc M. Biostatistical Applications in Epidemiology. 2011.

5 Upton G, Cook I. *Oxford Dictionary of Statistics*. 2nd ed. Oxford University Press 2008. doi:DOI: 10.1093/acref/9780199541454.001.0001

6 Parta M. A Dictionary of Epidemiology. *Int J Epidemiol* 2008;**15**:277. doi:10.1093/ije/15.2.277

7 Sauer B, VanderWeele TJ. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. *Agency Healthc Res Qual* 2013;:177–84.

8 Thabane L, Mbuagbaw L, Zhang S, *et al.* A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol* 2013;**13**:92. doi:10.1186/1471-2288-13-92

9 Grimshaw J. A knowledge synthesis chapter. *CIHR* 2010;:1–56.http://www.cihr-irsc.gc.ca/e/41382.html

10 Cochrance Group. 9_7_sensitivity_analyses @ handbook.cochrane.org. http://handbook.cochrane.org/chapter_9/9_7_sensitivity_analyses.htm

11 Walker D, Fox-Rushby J. Allowing for uncertainty in economic evaluations: qualitative sensitivity analysis. *Health Policy Plan* 2001;**16**:435–43. doi:10.1093/heapol/16.4.435

12 Louviere JJ, Hensher DA, Swait JD. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press 2000.

13 Ryan M, Bate a, Eastmond CJ, *et al.* Use of discrete choice experiments to elicit preferences. *Qual Health Care* 2001;**10 Suppl 1**:i55–60. doi:10.1136/qhc.0100055..

14 Viney R, Lancsar E, Louviere J. Discrete choice experiments to measure consumer preferences for health and healthcare. *Expert Rev Pharmacoecon Outcomes Res* 2002;**2**:319–26. doi:10.1586/14737167.2.4.319

15 Clark MD, Determann D, Petrou S, *et al.* Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics* 2014;**32**:883–902. doi:10.1007/s40273-014-0170-x

16 Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. *BMC Health Serv Res* 2014;**14**:367. doi:10.1186/1472-6963-14-367

17 Johnson FR, Lancsar E, Marshall D, *et al.* Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value Heal* 2013;**16**:3–13. doi:10.1016/j.jval.2012.08.2223

18 World Health Organization (WHO). How to Conduct a Discrete Choice Experiment for Health Workforce Recruitment and Retention in Remote and Rural Areas: A User Guide with Case Studies. Published Online First: 2012.http://www.who.int/hrh/resources/DCE_UserGuide_WEB.pdf?ua=1

19 Mangham LJ, Hanson K, McPake B. How to do (or not to do)...Designing a discrete choice experiment for application in a low-income country. *Health Policy Plan* 2009;**24**:151–8. doi:10.1093/heapol/czn047

20 Abiiro GA, Leppert G, Mbera GB, *et al.* Developing attributes and attribute-levels for a discrete choice experiment on micro health insurance in rural Malawi. *BMC Health Serv Res* 2014;**14**:235. doi:10.1186/1472-6963-14-235

21 Louviere JJ, Flynn TN, Carson RT. Discrete choice experiments are not conjoint analysis. *J Choice Model* 2010;**3**:57–72. doi:10.1016/S1755-5345(13)70014-9

22 Campbell D. Identification and analysis of discontinuous preferences in discrete choice experiments. *Eur Assoc Environ Resour Econ Annu Conf Gothenburg, Sweden* 2008;:25–8.http://www.webmeets.com/files/papers/EAERE/2008/685/eaere_discon_pref_DannyCampbell.pdf

23 Carson RT, Louviere JJ, Anderson D a., *et al.* Experimental analysis of choice. *Mark Lett* 1994;**5**:351–67. doi:10.1007/BF00999210

24 McFadden D, Train K. Mixed MNL models for discrete response. *J Appl Econom* 2000;**15**:447–70. doi:10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1

25 Davey J, Turner RM, Clarke MJ, *et al.* Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-

sectional, descriptive analysis. *BMC Med Res Methodol* 2011;**11**:160. doi:10.1186/1471-2288-11-160

26 Ryan M, Gerard K, Amaya-Amaya N. *Using Discrete Choice Experiments to Value Health and Health Care*. Dordrecht, The Netherlands: : Springer 2008.

27 Montopoli G, Anderson DA. THE ANALYSIS OF DISCRETE CHOICE EXPERIMENTS WITH CORRELATED ERROR STRUCTURE. 2001;**30**:615–26.

28 Cho HJ, Kim KS. Analysis of heteroscedasticity and correlation of repeated observations in Stated Preference (SP) data. *KSCE J Civ Eng* 2002;**6**:161–9. doi:10.1007/BF02829133

29 Cantillo V, Ortuzar JD, Williams H. Modeling discrete choices in the presence of inertia and serial correlation. *Transp Sci* 2007;**41**:195–205. doi:10.1287/trsc.1060.0178

30 Cochrance Group. Cochrane handbook: Meta-analysis of dichotomous outcomes. http://handbook.cochrane.org/chapter_9/9_4_4_meta_analysis_of_dichotomous_outcomes.htm

31 Rucker G, Schwarder G, Carpenter J, *et al.* Why add anything to nothing? The arcsine differnce as a measure of treatment effect in meta-analysis with zero cells. *Stat Med* 2009;**28**:721–38. doi:10.1002/sim.3511

32 Warren FC, Abrams KR, Golder S, *et al.* Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC Med Res Methodol* 2012;**12**:64. doi:10.1186/1471-2288-12-64

33 Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. *Biom J* 2015;**00**:1–16. doi:10.1002/bimj.201400184

34 Spittal MJ, Pirkis J, Gurrin LC. Meta-analysis of incidence rate data in the presence of zero events. *BMC Med Res Methodol* 2015;**15**:42. doi:10.1186/s12874-015-0031-0

35 Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Stat Med* 2010;**29**:2078–89. doi:10.1002/sim.3964

36 Liu D, Liu RY, Xie M. Exact Meta-Analysis Approach for Discrete Data and its Application to 2 x 2 Tables With Rare Events. *J Am Stat Assoc* 2014;**109**:1450–65. doi:10.1080/01621459.2014.946318

37 Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Stat Med* 2015;**34**:1097–116. doi:10.1002/sim.6383

38 Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philos Trans R Soc London* 1937;**236**:333–80. doi:10.1098/rsta.1937.0005

39 O'Hagan L, O'Hagan A, Luce BR, *et al.* A Primer On Bayesian Statistics In Health Economics And Outcomes Research. 2003;:76. doi:10.4249/scholarpedia.5230

40 Etzioni RD, B KJ. Bayesian Statistical Methods in Public Health and Medicine. *Annu Rev Public Health* 1995;**16**:23–34.

41 Sander GD, Inou L, Samsa G, *et al.* Use of Bayesian Techniques in Randomized Clinical Trials: A CMS Case Study. 2009.

42 Kadane JB. Bayesian methods for health-related decision making. *Stat Med* 2005;**24**:563–7. doi:10.1002/sim.2036

43 Gurrin LC, Kurinczuk JJ, Burton PR. Bayesian statistics in medical research: An intuitive alternative to conventional data analysis. *J Eval Clin Pract* 2000;**6**:193–204. doi:10.1046/j.1365-2753.2000.00216.x

44 FDA. Guidance for the use of Bayesian statistics in medical device clinical trials. *Guid Ind FDA Staff* 2010;:1–50.papers2://publication/uuid/2A24E02F-7D9A-4010-AB7A-A0982B757816

45 DiMichele. Inhibitors in haemophilia: A primer. *Haemophilia* 2000;**6**:38–40. doi:10.1046/j.1365-2516.2000.00045.x

46 Xi M, Makris M, Marcucci M, *et al.* Inhibitor development in previously treated hemophilia A patients: a systematic review, meta-analysis, and meta-regression. *J Thromb Haemost* 2013;**11**:1655–62. doi:10.1111/jth.12335

47 Oldenburg J, Goudemand J, Valentino L, et al. Postauthorization safety surveillance of ADVATE [antihaemophilic factor (recombinant), plasma/albumin-free method] demonstrates efficacy, safety and low-risk for immunogenicity in routine clinical practice. *Haemophilia*. 2010;16(6):866-877. doi:10.1111/j.1365-2516.2010.02332.x.

48 Marcucci M, Cheng J, Oldenburg J, et al. Meta-analysis of Post Authorization Safety Studies: worldwide postmarking surveillance of hemophilia A patients treated with antihemophilic factor recombinant plasma/albumin-free method rAHF-PFM. *J Thromb Haemost*.

2013;11(Suppl 2):1075.

http://onlinelibrary.wiley.com/doi/10.1111/jth.12443/pdf.

49 Romanov V, Marcucci M, Cheng J, Thabane L, Iorio A. Evaluation of Safety and Effectiveness of factor VIII treatment in Hemophilia A patients with low titer inhibitors or a personal history of inhibitor. *Thromb Haemost*. 2015;113(3):Inpress.

CHAPTER 2

AN EMPIRICAL COMPARISON OF METHODS FOR ANALYZING CORRELATED DATA FROM A DISCRETE CHOICE SURVEY TO ELICIT PATIENT PREFERENCE FOR COLORECTAL CANCER SCREENING

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

An empirical comparison of methods for analyzing correlated data from a discrete choice survey to elicit patient preference for colorectal cancer screening

Ji Cheng[1,2], Eleanor Pullenayegum[1,2], Deborah A Marshall[3], John K Marshall[4] and Lehana Thabane[1,2,5*]

Abstract

Background: A discrete choice experiment (DCE) is a preference survey which asks participants to make a choice among product portfolios comparing the key product characteristics by performing several choice tasks. Analyzing DCE data needs to account for within-participant correlation because choices from the same participant are likely to be similar. In this study, we empirically compared some commonly-used statistical methods for analyzing DCE data while accounting for within-participant correlation based on a survey of patient preference for colorectal cancer (CRC) screening tests conducted in Hamilton, Ontario, Canada in 2002.

Methods: A two-stage DCE design was used to investigate the impact of six attributes on participants' preferences for CRC screening test and willingness to undertake the test. We compared six models for clustered binary outcomes (logistic and probit regressions using cluster-robust standard error (SE), random-effects and generalized estimating equation approaches) and three models for clustered nominal outcomes (multinomial logistic and probit regressions with cluster robust SE and random effects multinomial logistic model). We also fitted a bivariate probit model with cluster-robust SE treating the choices from two stages as two correlated binary outcomes. The rank of relative importance between attributes and the estimates of β coefficient within attributes were used to assess the model robustness.

Results: In total 468 participants with each completing 10 choices were analyzed. Similar results were reported for the rank of relative importance and β coefficients across models for stage one data on evaluating participants' preferences for the test. The six attributes ranked from high to low as follows: cost, specificity, process, sensitivity, preparation and pain. However, the results differed across models for stage-two data on evaluating participants' willingness to undertake the tests. Little within-patient correlation (ICC ≈ 0) was found in stage-one data, but substantial within-patient correlation existed (ICC = 0.659) in stage-two data.

Conclusions: When small clustering effect presented in DCE data, results remained robust across statistical models. However, results varied when larger clustering effect presented. Therefore, it is important to assess the robustness of the estimates via sensitivity analysis using different models for analyzing clustered data from DCE studies.

Keywords: Discrete choice experiment, Intra-class correlation, Statistical model, Patient preference

* Correspondence: thabanl@mcmaster.ca
[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada
Full list of author information is available at the end of the article

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Background

With increased emphasis on the role of patients in healthcare decision making, discrete choice experimental (DCE) designs are more often used to elicit patient preferences among proposed health services programs [1,2]. DCE is an attribute-based design drawn from Lancaster's economic theory of consumer behaviour [3] and the statistical principles of the design of experiments [4]. This method measures consumer preference according to McFadden's random utility (benefit) maximisation (RUM) framework amongst a choice set which contains two or more alternatives of products or goods varying along several characteristics (attributes) of interest. In the early 1980s, Louviere, Hensher and Woodworth [5,6] introduced DCE into marketing research, and since then DCE has been rapidly adopted by researchers in other areas such as transportation, environment and social science. Its applications in health research emerged in the early 1990s, and it has been increasingly used to evaluate patient preferences for currently available and newly-proposed health services or programs in health economics and policy-making related topics. For example, in the health economics related research area, 34 published studies used DCE design in the period from 1990 to 2000, and 114 DCE design studies were published in the period from 2001 to 2008 [7].

In the short history of using DCE in health research, there were several reviews [7-9], and debates about methodological and design issues, challenges and future development [10-12]. In generating a DCE study, three major formats of the choice design have frequently been used: i) a forced choice between two alternatives, ii) a choice among three or more alternatives with an opt-out option, and iii) a two-staged choice process which forces participants to choose one of the alternatives and then an opt-out choice is provided to allow participants to say no to all proposed products [13]. Despite the rapid developments in design aspects [12,14], less attention was paid to the statistical analysis and model selection issues. Lancaster and Louviere [15] and Ryan and et al. [13] discussed several statistical models used for DCE including multinomial logistic model (MNL), multinomial probit model (MNP), and mixed logit model (MIXL). However, these studies did not provide detailed comparisons amongst competing models, or a clear indication of how to best deal with model selection issues. Another aspect related to the analysis of DCE data is adjustment for clustering effects. For example, in the DCE survey, it is common to ask participants to respond to several choice tasks in one survey. Each choice task has the same format but different attribute combinations. Naturally the choices made by same person would be expected to be more similar than the choices of other persons, leading to the within-patient correlation of responses. This within-subject correlation caused by the clustering effects or repeated observations needs to be accounted for in the analysis [16]. It is often measured using the intra-class correlation coefficient (ICC) where ICC = 0 indicates no intra-person correlation and ICC = 1 indicates perfect intra-person correlation. In this paper, we empirically compared some commonly-used statistical models which also account for the clustering effects in DCE analysis. We assessed the robustness (consistency and discrepancy) of the models on ranking of the relative importance between the attributes and the estimates of the β coefficients within each level of the attributes.

The data we used were taken from the preference survey on colorectal cancer (CRC) screening tests conducted in Hamilton, Ontario, Canada in 2002 [17]. This project used a two-level choice design. Thus, the data structure allowed us to investigate the statistical models for analyzing binary, nominal and bivariate outcomes for DCE data.

Methods

Overview of the CRC screening project

The Canadian Cancer Society reported in 2011 that CRC is the fourth most commonly diagnosed cancer and the second leading cause of cancer death in Canada [18]. According to the same report, the estimates of new cases of CRC and CRC related death in 2011 were 22,200 (50 per 100,000 person) and 8,900 (20 per 100,000 persons) in 2011. Although CRC has a high incidence rate, patients have a better chance of successful treatment if diagnosis can be made earlier. Although a population-based CRC screening program is highly recommended for people over 50 years of age [19,20], the uptake rate in North America is only about 50% [21]. Therefore, better understanding of patient preferences for screening tests may be the key to the successful implementation and uptake of CRC screening programs. This survey was the first conducted in Canada to evaluate patient preferences for various CRC screening tests to identify the key attributes and levels that may influence CRC screening test uptake.

Traditional CRC screening modalities such as fecal occult blood testing (FOBT), flexible sigmoidoscopy (SIG), colonoscopy (COL) and double-contrast barium enema (DCBE) vary on their process, accuracy, comfort and cost [22]. In this survey, five important attributes of features of the screening tests were identified through review of the literature, consultation with clinical specialists and patient focus groups. They were: process (4 levels), pain (2 levels), preparation (3 levels), specificity (3 levels) and sensitivity (3 levels). In addition, cost (4 levels) was included due to its potential influence on the uptake (Table 1). To reduce the

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 1 Attributes and Levels Used in the Stated Preference Survey

Attributes	Attribute description as presented to patients	Levels	Level description as presented to patients
Process	How is it done?	Stool	You place 2 stool samples onto special cards for 3 consecutive days and return them to your doctor
		Scope	A flexible tube with a small camera at the tip is inserted into your rectum and through your colon
		CT	You lie on a special table while a machine moves around you and takes x-ray pictures (like a CAT scan)
		Enema and X-ray	Air and a white liquid are injected into your colon through a rectal tube. x-ray pictures are taken as the liquid moves through your colon*
Pain	Is there pain or discomfort?	None	You feel no pain during the test
		Mild	You may feel mild pain or discomfort during the test*
Preparation	What do you do to prepare?	None	No preparation required
		Diet	You must alter your diet for 5 days by avoiding some specific foods and over-the-counter medications
		Enema/lax	Before the test you must take laxatives or enemas which cause diarrhea to clean your colon*
Specificity	Is it accurate if you DO NOT have cancer?	100%	If you DO NOT have cancer, the test result will never say you may have cancer. No other test is needed.
		80%	If you DO NOT have cancer, the test result will say you may have cancer 2 out of 10 times. You then need to have a different test done
		50%	If you DO NOT have cancer, the test result will say you may have cancer 5 out of 10 times. You then need to have a different test done*
Sensitivity	Is it accurate if you DO	90%	If you DO have cancer, the test will miss it 1 out of 10 times
		70%	If you DO have cancer, the test will miss it 3 out of 10 times
		40%	If you DO have cancer, the test will miss it 6 out of 10 times*
Cost	How much would you pay?	$10	$10*
		$50	$50
		$250	$250
		$500	$500

*Reference level for attribute

burden on respondents for making their choices on 864 ($4 \times 2 \times 3 \times 3 \times 3 \times 4$) unique combination from full factorial design, we used a fractional factorial design. In this design, 40 choice tasks were divided into four blocks to create a subset of 10 choice tasks of the attribute combinations for each survey participant to evaluate. The original design was developed using the SAS Optex procedure and optimized several measures of efficiency: 1) level balance; 2) orthogonality; and 3) D-efficiency [17,23]. This design ensured the ability of estimating the main effects of the attributes while minimizing the number of combinations. No prior information on the ranking of attributes from the literature was available at the time of the design of the study. The survey used the pair-wise binary two-stage response design [24] with the choice between two choice sets of the attributes at different levels as the first step and the addition of an opt-out option as the second step (Table 2). This design maximized the information gained through the questionnaire to understand patient preferences on the CRC screening tests and the factors affecting the uptake rate. However, the analysis presented challenges. First, the answers were likely to cluster within subjects because each subject made two sequential choices for ten choice tasks. Therefore, a statistical model adjusting for within-subject correlation for repeated measurements was needed. Second, in the original paper, the analysis was done using the bivariate probit model, but the analysis could be approached using different methods: treating the responses at the two stages as independent responses, as sequential and correlated bivariate responses, or as a single response with three levels (Test A, Test B or No screening).

Outcomes

According to the unique data structure of the two-stage design, we conducted three analytic approaches. 1) Analyze the two-staged sequential choices of each choice task separately, i.e. binary outcomes: a) subject preferences on the screening modalities which only included patient responses at the first stage, and b) subject willingness to participate in the screening program which only included subjects' responses at the second stage. 2) Treat the two-staged data as paralleled three-choice options including Test A, Test B and "opt-out", i.e.

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 2 A sample question

Features	Test A	Test B
How is it done?	You place 2 stool samples onto special cards for 3 consecutive days and return them to your doctor	A flexible tube with a small camera at the tip is inserted into your rectum and through your colon
Is there pain or discomfort?	You feel no pain during the test	You may feel mild pain or discomfort during the test
What do you do to prepare?	You must alter your diet for 5 days by avoiding some specific foods and over-the-counter medications	Before the test you must take laxatives or enemas which cause diarrhea to clean your colon
Is it accurate if you DO NOT have cancer?	If you DO NOT have cancer, the test result will say you may have cancer 5 out of 10 times. You then need to have a different test done	Same as for Test A
Is it accurate if you DO have cancer?	If you DO have cancer, the test will miss it 3 out of 10 times	Same as for Test A
How much would you pay?	$50	$250
Which test would you prefer (please mark one box only)	Prefer A	Prefer B
Suppose you now have the option of no screening. What would you prefer now? (please mark one box only)	I would still prefer the test chosen above	
	I would prefer no screening	

nominal data. 3) Treat the two-staged data as two correlated binary choice sets, i.e. bivariate outcomes. Figure 1 presents the data structure of the original design and these three analysis approaches.

Random utility theory

As mentioned above, the DCE design is generally based on random utility theory [25] which expresses the utility (benefit) U_{in} of an alternative i in a choice set C_n

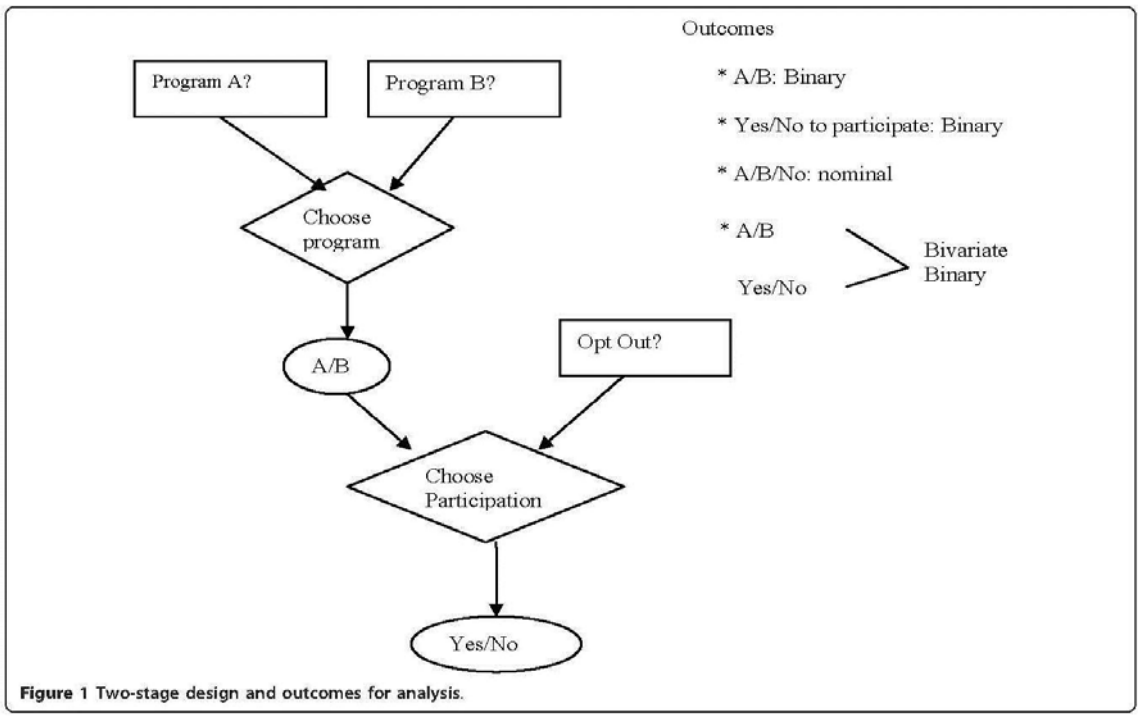


Figure 1 Two-stage design and outcomes for analysis.

(perceived by individual n) as two parts: 1) an explainable component specified as a function of the attributes of the alternatives $V(X_{in}, \beta)$; and 2) an unexplainable component (random variation) ε_{in}.

$$U_{in} = V(X_{in}, \beta) + \varepsilon_{in}$$

The individual n will choose alternative i over other alternatives if and only if this alternative gives the maximized utility. The relationship of the utility function and the observed k attributes of the alternatives can be assumed under a linear-in-parameter function.

$$V_{in} = \alpha_i + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

According to the assumption of the distribution of the error term ε_{in}, the models specification of DCE data can be varied.

Statistical methods

The statistical models discussed in this paper were organized according to the type of outcomes: i) logistic and probit models for binary outcomes, ii) multinomial logistic and probit models for nominal outcomes, and iii) bivariate probit model for bivariate binary outcomes. We provide some details on how the different statistical techniques account for the within-cluster correlation in analyzing clustered DCE data.

For the binary type of outcomes, we examined six statistical models which have the capacity to account for the within-patients correlations [26,27], including logistic regression with clustered robust standard error, random-effects logistic regression, logistic model using generalized estimating equations (GEE), probit regression with clustered robust standard error, random-effects probit regression, and probit regression using generalized estimating equation (GEE) model. Below are some brief descriptions of the methods.

Standard logistic regression and standard probit regression

Both standard logistic and probit regressions assume that the observations are independent. However in our dataset, each subject completed ten choice tasks, i.e. each subject had ten observations (choice tasks) which formed a cluster or can be considered repeated measurement. Normally, the observations in the same clusters are more similar (correlated) comparing to the observations out of the cluster. Therefore, adjusting the correlation within the cluster is necessary. We used three methods to adjust the within-cluster correlation.

Clustered robust standard error

In this method the independence assumptions are relaxed among all observations, but it is assumed that the observations across clusters are independent. The total variance is empirically estimated using Huber-White (also called Sandwich) standard error [28]. This method takes only the intra-class correlation into account, but the degrees of freedom are still based on the number of observations, not the number of clusters [29]. Therefore, this method only adjusts the standard error related to the confidence interval, but the point estimates are left unchanged.

Random-effects method

In this method, the total variance has two components: between-cluster variance and within-cluster variance. We assume that, at the cluster level, data follow a normal distribution with mean zero and between-cluster variance τ^2; and that within each cluster, data vary according to some within-cluster variance [30]. This method takes two types of variance into account when estimating the total variance and the degrees of freedom are calculated based on the number of clusters [31]. Therefore, the point estimates and their corresponding variances are adjusted for intra-cluster correlation. For the covariance structure, we assumed equal variances for the random effects and a common pairwise covariance [32]. This structure corresponds to the exchangeable correlation structure specified for GEE method, which we describe below. The key difference between the random-effects method and other methods discussed here is that the random-effects method estimates the parameters for each subject within cluster or clusters sharing the same random effects. Therefore, the random effect is also often called subject specific effect [33].

GEE method

This method allows a working correlation matrix to be specified to adjust the within-cluster correlation. We assumed that there was no ordering effect among the observation in each cluster, allowing us to use an exchangeable correlation matrix [34]. As in the random-effects method, the degrees of freedom are based on the number of clusters, which in turn adjusts the estimate of the confidence interval [35]. Unlike the random-effects method, the GEE approach estimates the regression parameters averaging over the clusters (so-called population average model) [36].

For the nominal type of outcomes, we used three statistical models [37]: multinomial logistic model with clustered robust standard error, random-effects multinomial logistic model, and multinomial probit model with clustered robust standard error. We also fitted a bivariate probit model in which the choices from two stages were treated as two binary outcomes [38].

Multinomial logistic model

McFadden's conditional logit model (CLM), also called multinomial logistic (MNL) model, was the pioneer and most commonly used model in the early DCE studies [39]. The key assumption of this model is that the error terms ε_{in} are independent and identically distributed

(IID) [13], which leads to the independence of irrelevant alternatives (IIA) property [40]. Another assumption for this model is that the error term has an extreme value distribution with mean 0 and variance $\pi^2/6$ [37]. To take the intra-class correlation into account, the clustered robust SE was used.

Random-effects multinomial logistic model

Similar to the random-effects models used for analyzing binary outcomes, this model takes two levels of variance, between-cluster variance and within-cluster variance, into account for clustered or longitudinal nominal responses [41,42].

Multinomial probit model

Multinomial probit model (MNP) (heteroscedastic models) is considered to be one of the most robust, flexible and general models in DCE, especially when the correlation (heteroscedasticity) between alternatives is presented [43]. The model is assumed to have a normally distributed error term. The benefit of using MNP model is that the IIA assumption which is the strict requirement for MNL model can be somehow relaxed [37]. The main concern in using this model is that its maximization involves Monte Carlo simulation but not the analytical maximization which could lead to a computational burden. Again, the clustered robust SE was used to incorporate the intra-class correlation.

Bivariate probit model

In this model, we assume that the choices between two stages (stage 1: choice between screening test; stage 2: choice between participation and opt-out) are not independent. It says that subject choice as to whether or not to participate in the screening program was conditional on subject preference for the screening modalities [44]. By fitting this model, two types of correlation can be taken into account: the correlation between the outcomes from stage 1 and stage 2, incorporated through the bivariate nature of the model itself, and the intra-class correlation, incorporated through use of the cluster robust SE.

To assess the necessity of accounting for the intra-class correlation for analyzing clustered correlated DCE data, we also presented the results from the above models using simple standard error (SE)—which does not take clustering into account. They are the standard logistic, probit, multinomial logistic, multinomial probit and bivariate probit models.

We compared results from the above models on the following criteria: rank on the relative importance of the attributes, and magnitude, direction and significance of the estimates of the β coefficient within each level of the attributes, which were obtained by regressing preference onto the difference in attributes between the two choices. The ranking criterion was measured by the percent change between the log-likelihood value of the full model and the value after removing one specific attribute from the model [45]. To evaluate the significance of the estimate of the β coefficients within each attribute, the criterion for statistical significance was set at alpha = 0.05. All statistical models were conducted using STATA 10.2 (College Station TX) and the figures were plotted using PASW Statistics 19 (SPSS: An IBM Company).

Results

A random sample of 1,170 patients was selected from a roster of 9,959 patients aged 40-60 years from the Hamilton Primary Care Network. After excluding the patients who did not pass the inclusion criteria, questionnaires were mailed to 1,049 patients. Of these, 547 were returned and 485 had complete data. Among the patients with complete data, we excluded 17 patients who did not pass the rationale test, which were two warm-up choice tasks. For these warm-up tasks, one alternative was dominant over another possessing all favourable attribute levels and the respondents who did not choose the dominant alternative were considered to have failed the rational test. Finally, we analyzed the data for 468 patients (Figure 2) from four blocks with the block size of 105, 124, 120 and 119 respectively.

The mean age of the subjects was 50.8 years (standard deviation, 5.95 years), which was similar to the recommended age to start CRC screening [46]. Of the 468 included subjects, about 48% were female, 12% had family history of CRC and two patients (0.2%) had been diagnosed with CRC. The detailed demographic characteristics are presented in Table 3.

For the two-point outcomes (binary), the rank of the attributes on the choice of Test A and Test B was consistent across models. From most important to least important, they ranked as follows: cost, specificity, process, sensitivity, preparation and pain (Figure 3). With the exception of the random-effects logistic and probit models, the ranking (from most important to least important) of the six attributes for assessing participation or opt-out (stage-two), was as follows: cost, sensitivity, preparation, process, specificity and pain. The ranking from random-effects models was: cost, sensitivity, process, specificity, preparation and pain (Figure 4). For the three-point outcomes (nominal and bivariate) in which the choices of Test A, Test B and opt-out were estimated simultaneously, the attributes were ranked consistently: cost, sensitivity, specificity, process, preparation and pain (Figure 5). Comparing to the models using simple SE, using clustered robust SE to incorporate intra-class correlation did have any effects on calculating the relative importance of attributes.

When looking at how certain levels of each attribute affected the choice between Test A and Test B (stage-

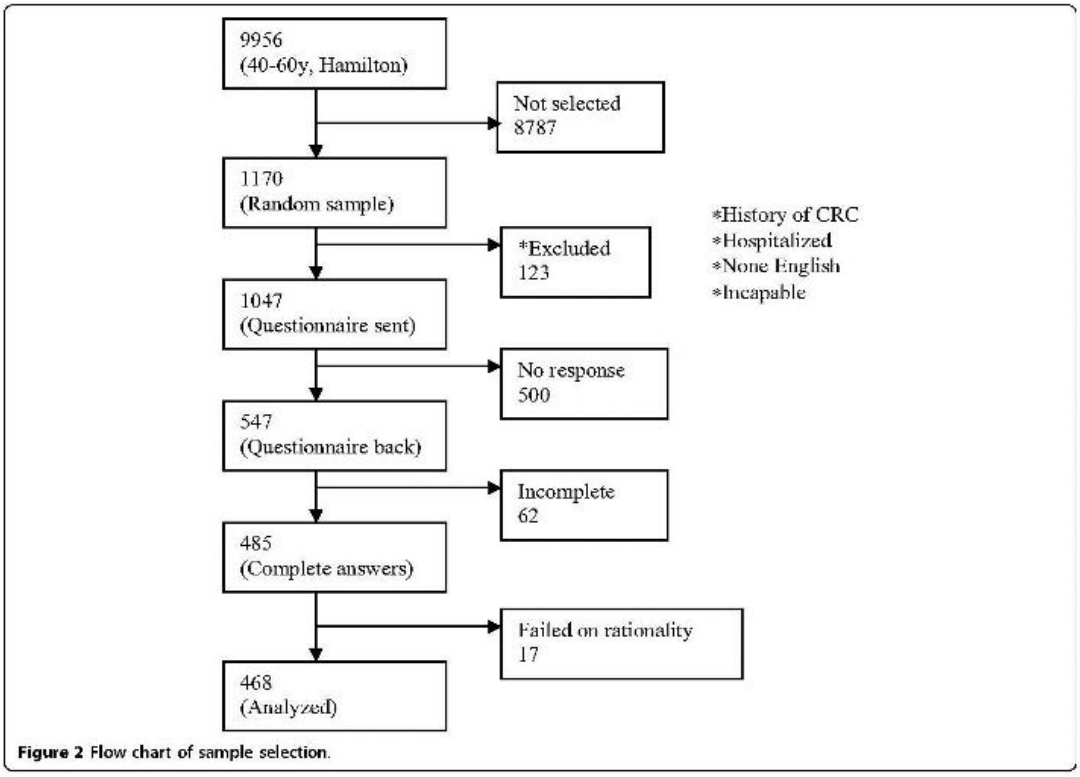


Figure 2 Flow chart of sample selection.

Table 3 Demographic characteristic of respondents

Personal Characteristics	(n = 468)
Age in years: Mean (SD)	50.8 (5.95)
Gender	
Male	52%
Female	48%
Health Status	
Excellent	14%
Very good	42%
Good	33%
Fair	9%
Poor	2%
Family history of CRC	
Yes	12%
No	82%
I don't know	6%
Diagnosed with CRC	
Yes	0.4% (2 patients)
No	99%
I don't know	0.6%

one), the estimates of the β coefficients were similar in magnitude and direction across different statistical models. The most preferred screening test had the following features: stool sample, no preparation, 100% specificity, 70% sensitivity, without pain and with an associated cost of $50. The least preferred screening test had the combination of colonoscopy, special diet for preparation, 80% specificity, 90% sensitivity, with mild pain and no associated cost (Table 4 and Table 5).

When assessing the impact of certain levels of each attribute on patient choice of participating or opt-out (stage-two), the β coefficient estimates for 90% sensitivity and no preparation had a significantly positive effect on uptake and this was consistent across all models. For other attributes and levels, results appeared similar across all three global analysis approaches: the random-effects and GEE logistic models and the random-effects and GEE probit models (Table 6); MNL with clustered robust SE, MNL random-effects and MNP with clustered robust SE (Table 7); and logistic with clustered robust SE, probit with clustered robust SE and bivariate probit (Table 6 and Table 7). The following two examples showed the estimates across models could differ by

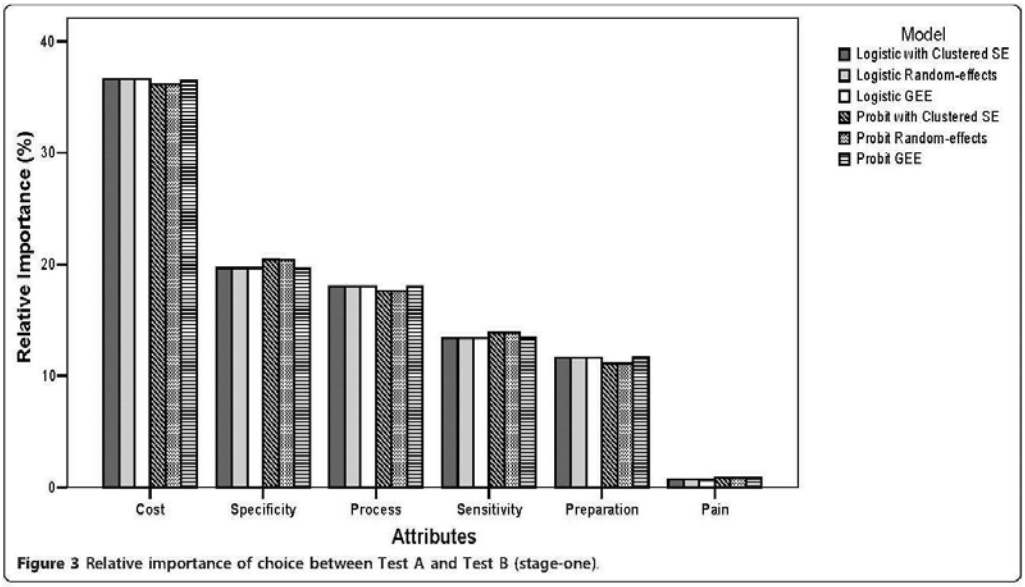


Figure 3 Relative importance of choice between Test A and Test B (stage-one).

magnitude and direction. The magnitude of estimates of the effect of 90% sensitivity varied by model, but the direction was similar across all models. When comparing the cost of $50 to no cost, logistic and probit random-effects and GEE models reported that participants preferred no cost. MNL with clustered robust SE, MNL random-effects and MNP with clustered robust SE model reported that participants preferred the $50 cost. For other models, no significant statistical differences were found (Figure 6). We also found that unlike the

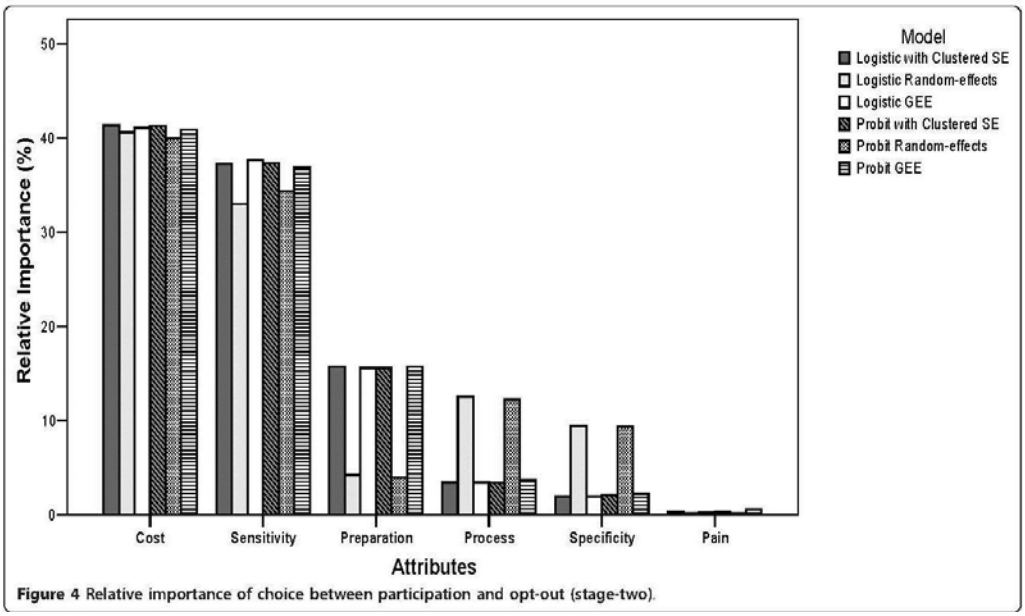


Figure 4 Relative importance of choice between participation and opt-out (stage-two).

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

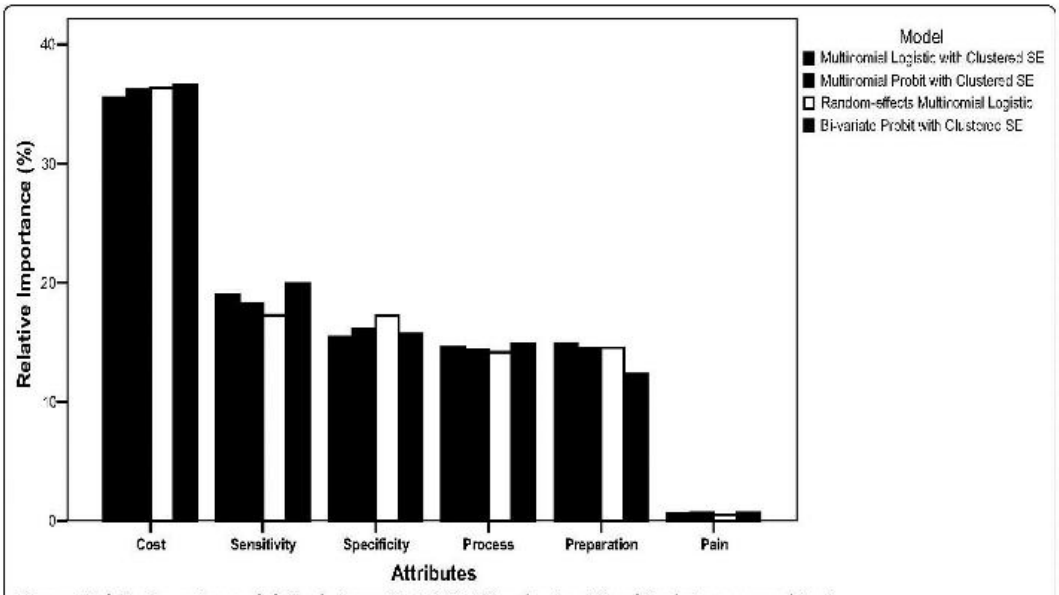


Figure 5 Relative importance of choice between Test A, Test B and opt-out (combined stage one and two).

results from the stage-one data (Table 4 and Table 5), for the stage-two data there was noticeable difference between the β coefficient estimates from the models with and without incorporating the intra-class correlation (Table 6 and Table 7).

When assessing the clustering effect, we found that intra-class correlation was small among the stage-one data (ICC ≈ 0) and relatively large among the stage-two data (ICC = 0.659). For this survey, it appears as though many patients had predetermined their participation for CRC screening. For example, among the 468 participants included in the analyses, 48% always chose to undertake the screening program and 15% always chose no participation regardless of how the screening modalities varied at the first stage. Although Test A and Test B were generic terms of the combinations of the different levels of six attributes and they were randomly assigned to appear first or second in one choice task, we found that 24% more participants chose Test A over Test B. All the design limitations had some impact on our interpretation of the analysis results.

Discussion

We applied six statistical models to binary outcomes, three models to nominal multinomial outcomes and one model to bivariate binary outcomes to estimate the ranking of key attributes of CRC screening tests using data from DCE survey conducted in Hamilton, Ontario, Canada in 2002. We used three methods to adjust the within-cluster correlations: clustered robust standard error, random-effects, and GEE methods. The results showed consistent answers for estimating subject preference for CRC screening tests, both on ranking the importance of the attributes and identifying the significant factors influencing subject choice between testing modalities. For estimating subject willingness to participate or undertake CRC screening (i.e. incorporating "out-put" option), models disagreed both on ranking the importance of the attributes and identifying the significant factors (i.e. attributes and levels) affecting whether or not subjects would participate.

Overall, our analyses showed that participants preferred a CRC screening test with the following characteristics: stool sample, no preparation, 100% specificity, 70% sensitivity and without pain. The CRC test with such a combination of attribute levels would be the FOBT test [18]. Thus, our findings appear to be consistent with the results from Nelson and Schwartz's survey in 2004 [47] which showed FOBT to be the most preferred option for CRC screening. In that survey, they also reviewed 12 previous studies, all of which showed FOBT to be a preferred choice by most patients.

The reason for the consistency in estimating the choice between screening tests and the discrepancy in estimating the choice between participation and "out-put" might be due to the model's ability to adjust the

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 4 Estimates of coefficients of patient choice between Test A and Test B (Two-point outcome from stage-one)

	Two-point outcome							
	Logistic Model				Probit Model			
	Simple SE	Robust SE	Random-effects	GEE	Simple SE	Robust SE	Random-effects	GEE
Process ref = Enema/X-ray)								
Stool	0.27 (0.16, 0.37)	0.27 (0.19, 0.35)	0.27 (0.16, 0.37)	0.27 (0.16, 0.37)	0.16 (0.10, 0.23)	0.16 (0.12, 0.21)	0.16 (0.10, 0.23)	0.16 (0.10, 0.23)
Scope	-0.20(-0.31,-0.09)	-0.20(-0.27,-0.13)	-0.20(-0.31,-0.09)	-0.20(-0.31,-0.09)	-0.12(-0.19,-0.05)	-0.12(-0.17,-0.08)	-0.12(-0.19,-0.05)	-0.12(-0.19,-0.05)
CT	-0.13(-0.24,-0.03)	-0.13(-0.23,-0.04)	-0.13(-0.24,-0.03)	-0.13(-0.24,-0.03)	-0.08(-0.15,-0.02)	-0.08(-0.14,-0.02)	-0.08(-0.15,-0.02)	-0.08(-0.15,-0.02)
No Pain (ref = mild pain)	0.04(-0.03, 0.10)	0.04(-0.01, 0.08)	0.04(-0.03, 0.10)	0.04(-0.03, 0.10)	0.02(-0.01, 0.06)	0.02(-0.001, 0.05)	0.02(-0.01, 0.06)	0.02(-0.01, 0.06)
Preparation (ref = Enema/Lax)								
None	0.17 (0.09, 0.25)	0.17 (0.11, 0.23)	0.17 (0.09, 0.25)	0.17 (0.09, 0.25)	0.10 (0.05, 0.16)	0.10 (0.07, 0.14)	0.10 (0.05, 0.16)	0.10 (0.05, 0.16)
Special Diet	-0.17(-0.26,-0.09)	-0.17(-0.24,-0.11)	-0.17(-0.26, -0.09)	-0.17 (-0.26, -0.09)	-0.11 (-0.16, -0.05)	-0.11 (-0.14, -0.07)	-0.11 (-0.16, -0.05)	-0.11 (-0.16, -0.05)
Specificity (ref = 50%)								
100%	0.14 (0.05, 0.22)	0.14 (0.06, 0.21)	0.14 (0.05, 0.22)	0.14 (0.05, 0.22)	0.09 (0.03, 0.14)	0.09 (0.04, 0.13)	0.09 (0.03, 0.14)	0.09 (0.03, 0.14)
80%	-0.26 (-0.35, -0.18)	-0.26 (-0.33, -0.20)	-0.26 (-0.35, -0.18)	-0.26 (-0.35, -0.18)	-0.17 (-0.22, -0.11)	-0.17 (-0.21, -0.12)	-0.17 (-0.22, -0.11)	-0.17 (-0.22, -0.11)
Sensitivity (ref = 40%)								
90%	-0.07 (-0.15, 0.01)	-0.07 (-0.14, 0.01)	-0.07 (-0.15, 0.01)	-0.07 (-0.15, 0.01)	-0.04 (-0.09, 0.01)	-0.04 (-0.08, 0.01)	-0.04 (-0.09, 0.01)	-0.04 (-0.09, 0.01)
70%	0.24 (0.15, 0.33)	0.24 (0.15, 0.33)	0.24 (0.15, 0.33)	0.24 (0.15, 0.33)	0.15 (0.03, 0.14)	0.15 (0.10, 0.20)	0.15 (0.03, 0.14)	0.15 (0.03, 0.14)
Cost (ref = $10)								
$50	0.63 (0.47, 0.79)	0.63 (0.0.52, 0.74)	0.63 (0.47, 0.79)	0.63 (0.47, 0.79)	0.39 (0.29, 0.49)	0.39 (0.32, 0.46)	0.39 (0.29, 0.49)	0.39 (0.29, 0.49)
$250	0.17 (0.01, 0.33)	0.17 (0.07, 0.28)	0.17 (0.01, 0.33)	0.17 (0.01, 0.33)	0.10 (0.01, 0.20)	0.10 (0.04, 0.17)	0.10 (0.01, 0.20)	0.10 (0.01, 0.20)
$500	0.44 (0.24, 0.63)	0.44 (0.30, 0.58)	0.44 (0.24, 0.63)	0.44 (0.24, 0.63)	0.27 (0.15, 0.39)	0.27 (0.19, 0.36)	0.27 (0.15, 0.39)	0.27(0.15, 0.39)

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 5 Estimates of coefficients of patient choice of Test A and Test B (Stage-one from three-point outcome)

	Three-point of outcome							
	Nominal					Bivariate		
	MNL Simple SE	MNL Robust SE	MNL Random-effects	MNP Simple SE	MNP Robust SE	Bivariate probit Simple SE	Bivariate probit Robust SE
Process (ref = Enema/X-ray)							
Stool	0.24 (0.11, 0.36)	0.24 (0.14, 0.33)	0.23 (0.09, 0.36)	0.18 (0.09, 0.28)	0.18 (0.11, 0.26)	0.16 (0.10, 0.23)	0.16 (0.12, 0.21)
Scope	-0.25 (-0.38, -0.13)	-0.25 (-0.34, -0.17)	-0.29 (-0.42, -0.15)	-0.19 (-0.29, -0.09)	-0.19 (-0.26, -0.11)	-0.12 (-0.18, -0.05)	-0.12 (-0.16, -0.08)
CT	-0.14 (-0.27, -0.01)	-0.14 (-0.25, -0.02)	-0.13 (-0.27, -0.01)	-0.10 (-0.20, -0.001)	-0.10 (-0.19, -0.01)	-0.08 (-0.15, -0.02)	-0.08(-0.14, -0.01)
No Pain (ref = mild pain)	0.04 (-0.03, 0.11)	0.04 (-0.01, 0.09)	0.04 (-0.04, 0.12)	0.04 (-0.01, 0.09)	0.03 (-0.01, 0.07)	0.02 (-0.01, 0.06)	0.02 (-0.01, 0.05)
Preparation (ref = Enema/Lax)							
None	0.18 (0.09, 0.28)	0.18 (0.11, 0.26)	0.18 (0.08, 0.28)	0.14 (0.07, 0.22)	0.14 (0.09, 0.20)	0.10 (0.05, 0.16)	0.10 (0.07, 0.14)
Special Diet	-0.21 (-0.32, -0.11)	-0.21 (-0.29, -0.14)	-0.24 (-0.35, -0.13)	-0.16 (-0.25, -0.08)	-0.16 (-0.22, -0.10)	-0.11 (-0.16, -0.05)	-0.11 (-0.14, -0.07)
Specificity (ref = 50%)							
100%	0.14 (0.04, 0.25)	0.14 (0.04, 0.25)	0.15 (0.04, 0.26)	0.11 (0.03, 0.19)	0.11 (0.03, 0.19)	0.08 (0.03, 0.14)	0.08 (0.04, 0.13)
80%	-0.29 (-0.39, -0.19)	-0.29 (-0.37, -0.21)	-0.33 (-0.44, -0.23)	-0.22 (-0.30, -0.14)	-0.22 (-0.29, -0.16)	-0.17 (-0.22, -0.11)	-0.17 (-0.21, -0.12)
Sensitivity (ref = 40%)							
90%	-0.07 (-0.17, 0.03)	-0.07 (-0.14, 0.01)	-0.07 (-0.17, 0.04)	-0.04 (-0.12, 0.03)	-0.04 (-0.10, 0.01)	-0.04 (-0.10, 0.02)	-0.04 (-0.08, 0.01)
70%	0.20 (0.09, 0.31)	0.20 (0.10, 0.30)	0.20 (0.08, 0.32)	0.16 (0.07, 0.25)	0.16 (0.09, 0.24)	0.15 (0.09, 0.21)	0.15 (0.10, 0.20)
Cost (ref=$10)							
$50	0.59 (0.40, 0.78)	0.59 (0.46, 0.72)	0.66 (0.46, 0.85)	0.45 (0.30, 0.60)	0.45 (0.35, 0.55)	0.39 (0.29, 0.49)	0.39 (0.32, 0.46)
$250	0.15 (-0.04, 0.35)	0.15 (0.02, 0.29)	0.14 (-0.06, 0.35)	0.11 (-0.05, 0.26)	0.11 (-0.01, 0.21)	0.10 (0.00, 0.20)	0.10 (0.04, 0.17)
$500	0.52 (0.28, 0.76)	0.52 (0.34, 0.70)	0.57 (0.32, 0.83)	0.39 (0.20, 0.57)	0.39 (0.25, 0.52)	0.27 (0.15, 0.39)	0.27 (0.19, 0.36)

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 6 Estimates of coefficients of patient choice of participation or opt-out (Two-point outcome from stage-two)

	Two-point outcome							
	Logistic Model				Probit Model			
	Simple SE	Robust SE	Random-effects	GEE	Simple SE	Robust SE	Random-effects	GEE
Process (ref = Enema/X-ray)								
Stool	0.00 (-0.11, 0.11)	0.00 (-0.10, 0.09)	-0.13 (-0.34, 0.08)	-0.04 (-0.10, 0.03)	0.00 (-0.07, 0.07)	0.00 (-0.06, 0.06)	-0.08 (-0.19, 0.04)	-0.02 (-0.06, 0.02)
Scope	-0.02 (-0.14, 0.09)	-0.02 (-0.13, 0.08)	-0.20 (-0.42, 0.01)	-0.06 (-0.13, 0.01)	-0.01 (-0.08, 0.06)	-0.01 (-0.08, 0.05)	-0.10 (-0.22, 0.02)	-0.04 (-0.08, 0.01)
CT	0.08 (-0.03, 0.20)	0.08 (-0.03, 0.20)	0.62 (0.40, 0.85)	0.18 (0.11, 0.25)	0.05 (-0.02, 0.12)	0.05 (-0.02, 0.12)	0.35 (0.22, 0.47)	0.11 (0.07, 0.15)
No Pain (ref = mild pain)	0.01 (-0.05, 0.08)	0.01 (-0.04, 0.07)	0.04 (-0.09, 0.16)	0.01 (-0.03, 0.05)	0.01 (-0.03, 0.05)	0.01 (-0.02, 0.04)	0.02 (-0.05, 0.09)	0.01 (-0.02, 0.03)
Preparation (ref = Enema/Lax)								
None	0.15 (0.06, 0.24)	0.15 (0.07, 0.22)	0.25 (0.09, 0.42)	0.08 (0.03, 0.14)	0.09 (0.03, 0.14)	0.09 (0.04, 0.13)	0.14 (0.04, 0.23)	0.05 (0.02, 0.08)
Special Diet	-0.03 (-0.13, 0.06)	-0.03 (-0.13, 0.07)	-0.01 (-0.20, 0.17)	-0.01 (-0.07, 0.05)	-0.02 (-0.08, 0.04)	-0.02 (-0.07, 0.04)	-0.01 (-0.11, 0.09)	-0.01 (-0.04, 0.03)
Specificity (ref = 50%)								
100%	0.01 (-0.08, 0.10)	0.01 (-0.07, 0.09)	0.34 (0.17, 0.51)	0.10 (0.04, 0.15)	0.01 (-0.05, 0.07)	0.01 (-0.04, 0.06)	0.19 (0.09, 0.29)	0.01 (0.02, 0.09)
80%	0.04 (-0.05, 0.13)	0.04 (-0.03, 0.11)	0.08 (-0.09, 0.25)	0.02 (-0.03, 0.08)	0.02 (-0.03, 0.08)	0.03 (-0.02, 0.07)	0.05 (-0.05, 0.14)	0.01 (-0.02, 0.05)
Sensitivity (ref = 40%)								
90%	0.21 (0.12, 0.30)	0.21 (0.13, 0.29)	0.71 (0.53, 0.89)	0.21 (0.16, 0.27)	0.13 (0.07, 0.18)	0.13 (0.08, 0.18)	0.40 (0.31, 0.50)	0.13 (0.10, 0.16)
70%	0.01 (-0.08, 0.11)	0.01 (-0.07, 0.09)	0.04 (-0.14, 0.22)	0.01 (-0.04, 0.07)	0.01 (-0.05, 0.07)	0.01 (-0.04, 0.06)	0.03 (-0.07, 0.13)	0.01 (-0.03, 0.04)
Cost (ref = $10)								
$50	-0.09 (-0.26, 0.08)	-0.09 (-0.21, 0.02)	-0.40 (-0.71, -0.08)	-0.12 (-0.24, -0.02)	-0.05 (-0.16, 0.05)	-0.05 (-0.12, 0.01)	-0.21 (-0.39, -0.04)	-0.07 (-0.13, -0.01)
$250	-0.16 (-0.36, 0.02)	-0.16 (-0.36, 0.04)	-1.21 (-1.56, -0.86)	-0.36 (-0.47, -0.25)	-0.10 (-0.20, 0.01)	-0.09 (-0.22, 0.02)	-0.67 (-0.87, -0.48)	-0.22 (-0.13, -0.01)
$500	-0.54 (-0.74, -0.34)	-0.54 (-0.74, -0.34)	-1.69 (-2.08, -1.30)	-0.53 (-0.65, -0.41)	-0.33 (-0.45, -0.21)	-0.33 (-0.45, -0.21)	-0.95 (-1.16, -0.73)	-0.32 (-0.40, -0.25)

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

Table 7 Estimates of coefficients of patient choice of participation or opt-out (Stage-two from three-point outcome)

	Three-point of outcome							
	Nominal				Bivariate			
	MNL Simple SE	MNL Robust SE	MNL Random-effects	MNP Simple SE	MNP Robust SE	Bivariate probit Simple SE	Bivariate probit Robust SE
Process (ref = Enema/X-ray)							
Stool	0.09 (-0.03, 0.21)	0.09 (-0.01, 0.19)	0.07 (-0.07, 0.20)	0.08 (-0.01, 0.18)	0.08 (0.01, 0.16)	0.00 (-0.07, 0.07)	0.00 (-0.06, 0.06)
Scope	-0.14 (-0.27, -0.01)	-0.14 (-0.25, -0.03)	-0.23 (-0.37, -0.09)	-0.10 (-0.20, 0.00)	-0.10 (-0.01, -0.19)	-0.02 (-0.08, 0.05)	-0.02 (-0.08, 0.04)
CT	0.04 (-0.09, 0.17)	0.04 (-0.08, 0.15)	0.13 (-0.01, 0.27)	0.02 (-0.08, 0.12)	0.02 (-0.07, 0.11)	0.05 (-0.02, 0.12)	0.05 (-0.02, 0.12)
No Pain (ref = mild pain)	0.03 (-0.04, 0.11)	0.03 (-0.02, 0.09)	0.03 (-0.05, 0.11)	0.03 (-0.03, 0.09)	0.03 (-0.02, 0.07)	0.01 (-0.03, 0.05)	0.01 (-0.02, 0.04)
Preparation (ref = Enema/Lax)							
None	0.23 (0.13, 0.33)	0.23 (0.15, 0.31)	0.20 (0.09, 0.31)	0.17 (0.09, 0.25)	0.17 (0.11, 0.24)	0.09 (0.03, 0.14)	0.09 (0.04, 0.13)
Special Diet	-0.13 (-0.24, -0.02)	-0.13 (-0.24, -0.02)	-0.17 (-0.29, -0.05)	-0.09 (-0.18, -0.01)	-0.09 (-0.17, -0.01)	-0.02 (-0.08, 0.04)	-0.02 (-0.08, 0.04)
Specificity (ref = 50%)							
100%	0.07 (-0.03, 0.17)	0.07 (-0.02, 0.17)	0.16 (0.05, 0.27)	0.06 (-0.02, 0.14)	0.06 (-0.01, 0.13)	0.01 (-0.05, 0.06)	0.01 (-0.04, 0.06)
80%	-0.09 (-0.19, 0.02)	-0.09 (-0.17, -0.01)	-0.17 (-0.28, -0.05)	-0.06 (-0.22, 0.09)	-0.06 (-0.13, -0.01)	0.03 (-0.03, 0.09)	0.03 (-0.02, 0.07)
Sensitivity (ref = 40%)							
90%	0.19 (0.09, 0.29)	0.19 (0.10, 0.27)	0.22 (0.11, 0.33)	0.14 (0.06, 0.22)	0.14 (0.07, 0.0.21)	0.13 (0.08, 0.18)	0.13 (0.08, 0.18)
70%	0.10 (-0.01, 0.21)	0.10 (0.01, 0.20)	0.13 (0.01, 0.25)	0.08 (-0.01, 0.16)	0.08 (0.01, 0.15)	0.01 (-0.05, 0.07)	0.01 (-0.04, 0.06)
Cost (ref=$10)							
$50	0.16 (-0.03, 0.35)	0.16 (0.04, 0.28)	0.25 (0.05, 0.45)	0.12 (-0.03, 0.27)	0.12 (0.02, 0.21)	-0.05 (-0.16, 0.05)	-0.05 (-0.12, 0.02)
$250	-0.08 (-0.28, 0.12)	-0.08 (-0.27, 0.11)	-0.26 (-0.48, -0.04)	-0.07 (-0.22, 0.09)	-0.07 (-0.22, 0.09)	-0.10 (-0.20, 0.01)	-0.10 (-0.22, 0.02)
$500	-0.32 (-0.54, -0.09)	-0.32 (-0.52, -0.11)	-0.28 (-0.63, -0.04)	-0.25 (-0.43, -0.08)	-0.25 (-0.42, -0.09)	-0.33 (-0.45, -0.21)	-0.33 (-0.45, -0.21)

Cheng et al. BMC Medical Research Methodology 2012, 12:15
http://www.biomedcentral.com/1471-2288/12/15

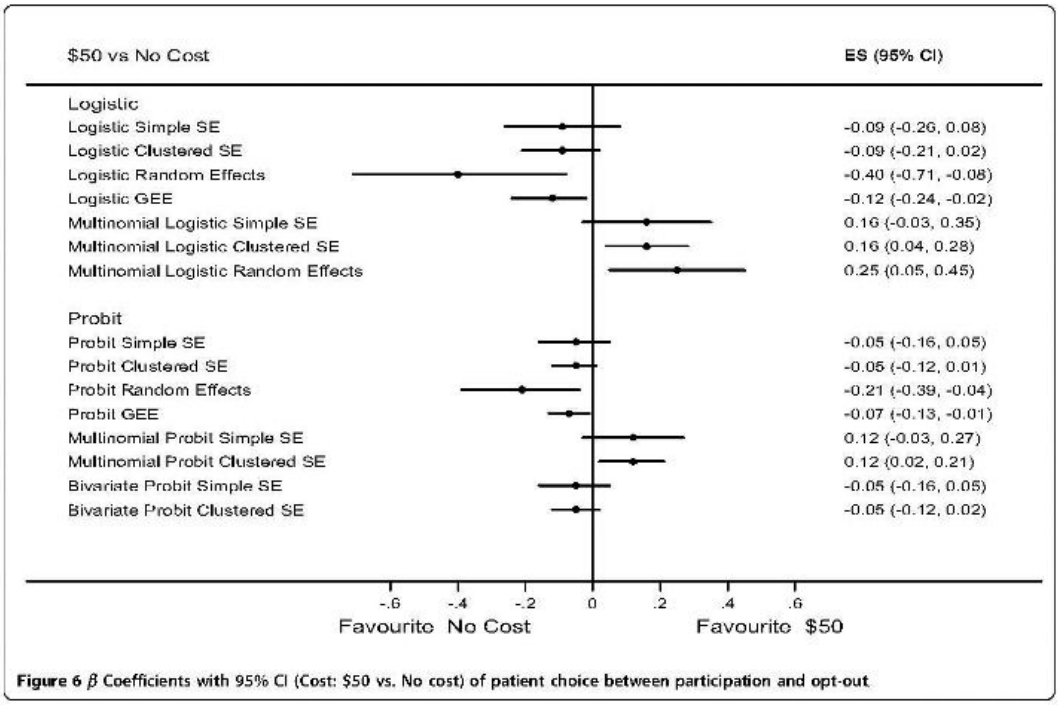


Figure 6 β **Coefficients with 95% CI (Cost: $50 vs. No cost) of patient choice between participation and opt-out**

within-participant (cluster) correlation. When the within-cluster correlation is small (choice between Test A and Test B), the assumption of the independently and identically distributed error term ε_{in} is held. Therefore, it might not be necessary to take the clustering effects into account and thus the estimates are similar across statistical models. However, when the intra-class correlation presents, the analysis needs to account for both the within-cluster variance and between-cluster variance [48].

To the best of our knowledge, this is the first empirical study to compare different methods to address the within-participant correlation in the analysis of DCE data. However, many authors have emphasized the importance of adjusting for clustering in analysis of clustered data or repeated measurements for binary outcomes [49,50]. When intra-class correlations are present in clustered or longitudinal data, the random-effects and GEE models are two commonly recommended approaches. Although they are estimating different parameters (the estimates from random-effects model are interpreted for the observations in the same cluster; the estimates from GEE model are interpreted as the mean across entire sample), the results from these two models are similar most of the time [41,51]. Some researchers generally prefer random-effects model when the results from these two approaches disagree. However, some researchers argue that the random-effects model could provide biased results due to unverifiable assumptions about the data distribution [52].

Comparing to the models for analyzing correlated binary data, statistical software seldom has ready-to-use statistical models developed for multinomial outcomes or multi-variate outcomes. The multinomial probit model is routinely used to deal with correlation between alternatives [53], but it does not take intra-class or intra-respondent correlation into account. Robust standard error can be specified for multinomial logistic or probit and bivariate logistic models to adjust the estimate of standard error, but this would not correct the bias related to point estimates (coefficients). A simulation study has shown that the bias and the inconsistency for estimating the within-cluster correlation increase with the size of the cluster [54]. The newly developed generalized linear latent and mixed model (***gllamm***) procedure in STATA has the ability to run random-effects multinomial logistic model [55] to address the intra-class correlation issue, but this model has yet to be evaluated for performance (i.e. whether or not yields unbiased estimates). Some researchers have suggested

using Bayesian hierarchical random-effects logistic and probit regression for clustered or panel data [56]. Although the Bayesian approach allows the flexibility to specify random effects, it requires considerable skill in programming.

This study has some limitations. First, this study is an empirical comparison of the analytic models and therefore we cannot know which model performs the best. Such an analysis would require simulation studies to assess the performance of the models in terms of the bias, precision, and coverage. Second, some estimates of the cost attribute in our study were inexplicable. For the test associated cost, participants' preference had a non linear order: $50, $0, $500 and $250. This could be a result of as the violation of the model assumptions or model misspecification. Most DCE analyses assume a linear utility function, but some recent studies have shown that this assumption may not be true for price-related attributes. A study of MPS players found that the utility function of the price and storage size had W-shaped curves rather than smooth linear trends [57]. A local travel mode study also found that the preference of time savings followed a non-linear utility function [58]. Another reason which may cause inaccurate results in our study is the use of two-staged design. The two-staged design had the advantage of maximizing the information gained by forcing participants to make a choice at the first stage, but it also gave us some artificial information. Third, many respondents in this survey seemed to have predetermined their participation in CRC screening before seeing the questionnaire. This may have caused an unusually high with-in cluster correlation when choosing between participation and opt-out. We also doubt that the predetermination might cause the ordering effect [59] when choosing the preferred screening tests. When individuals are forced to make a choice between products which they have decided that they do not want, the answer might not resemble the truth. Therefore, the results need to be interpreted cautiously–replication from similar studies is needed to better understand participant preferences for CRC screening and the willingness to undertake the screening program.

Conclusion

Responses from the same participant are likely to be more similar than the responses between participants in DCE data leading to possible intra-class or intra-participant correlation. Therefore, it is important to investigate the size of intra-class correlation before fitting any statistical model. We found that when within-cluster correlation is very small, all models gave consistent results both on the estimates ranking and coefficients. Therefore, the simplest logistic regression and multinomial logistic regression are recommended for the computation advantage being ease. Multinomial probit model may be a preferred choice method of analysis if we assume the existence of the correlation between alternatives.

When within-cluster correlation is high, sensitivity analyses are needed to examine the consistency of the results. Instead of making generalized inferences according to the estimate from any single statistical model, results from the sensitivity analyses based on different models can provide some insight about the robustness of the findings.

Our study empirically compared some commonly used statistical model on taking intra-class correlation into account when analyzing DCE data. To completely understand the necessity of accounting for the intra-class correlation for DCE data, particularly on analyzing nominal type of outcomes, simulation studies are needed.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgements

The original study was funded by a research grant from the Canadian Institutes for Health Research (MOB 53716) and the Cancer Research Foundation of America. We thank the reviewers for their insightful comments and suggestions that led to improvements in the manuscript.

Author details

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada. [2]Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada. [3]Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada. [4]Department of Medicine, Division of Gastroenterology, McMaster University, Hamilton, ON, Canada. [5]Biostatistics Unit/FSORC, 3rd Floor Martha, Room H325, St. Joseph's Healthcare Hamilton, 50 Charlton Avenue East, Hamilton, ON L8N 4A6, Canada.

Authors' contributions

JC (cheng7@mcmaster.ca) conducted literature review, preformed the statistical analyses and composed the draft of the manuscript. LT (thabane@mcmaster.ca) designed the original study, oversaw the statistical analysis and revised the manuscript. EP (oullena@mcmaster.ca) assisted planning statistical analyses and revised the manuscript. DAM (damersha@ucalgary.ca) and JKM (marshlj@mcmaster.ca) designed the original study and revised the manuscript. All authors read and approved the final manuscript.

Received: 28 May 2011 Accepted: 20 February 2012
Published: 20 February 2012

References

1. Longo MF, Cohen DR, Hood K, Edwards A, Robling M, Elwyn G, Russell IT: Involving patients in primary care consultations: assessing preferences using discrete choice experiments. *Br J Gen Pract* 2006, 56(522):35-42.
2. Ryan M, Major K, Skatun D: Using discrete choice experiments to go beyond clinical outcomes when evaluating clinical practice. *J Eval Clin Pract* 2005, 11(4):328-338.
3. Lancaster KJ: A new approach to consumer theory. *J Polit Econ* 1966, 74(2):132-157.

4. Montgomery DC: *Design and analysis of experiments*. 5 edition. New York: Wiley; 2000.
5. Louviere J, Henshen D: On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transp Res Rec* 1982, 890:11-17.
6. Louviere J, Woodworth G: Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *J Mark Res* 1983, 20:350-367.
7. de Bekker Grob EW, Ryan M, Gerard K: Discrete choice experiments in health economics: a review of the literature. *Health econ* 2010, doi:10.1002/hec.1697.
8. Ryan M, Gerard K: Using discrete choice experiments to value health care programs: current practice and future research reflections. *Appl Health Econ Health Policy* 2003, 2(1):55-64.
9. Marshall DA, Bridges JFP, Hauber B, Cameron R, Donnalley L, Fyie K, Johnson FR: Conjoint analysis applications in health–how are studies being designed and reported? an update on current practice in the published literature between 2005 and 2008. *The Patient. Patient-Centered Outcomes Research* 2010, 3:249-256.
10. Louviere JJ, Lancsar E: Choice experiments in health: the good, the bad, the ugly and toward a brighter future. *Health Econ Policy Law* 2009, 4(Pt 4):527-546.
11. Ryan S, Dolan P: Discrete choice experiments in health economics. For better or for worse? *Eur J Health Econ* 2004, 5(3):199-202.
12. Louviere JJ, Pihlens D, Carson R: Design of discrete choice experiments: a discussion of issues that matter in future applied research. *Journal of Choice Modelling* 2010, 4:1-8.
13. Ryan M, Gerard K, Amaya-Amaya M: *Using discrete choice experiments to value health and health care*. Dordrecht, The Netherlands: Springer; 2008.
14. Bridges J-F, Hauber B, Marshall DA, Lloyd A, Prosser LA, Regier DA, Johnson FR, Mauskopf J: Conjoint analysis applications in health–a checklist: a report of the ISPOR good research practices for conjoint Analysis task force. *Value Health* 2011, 14:403-413.
15. Lancsar E, Louviere J: Conducting discrete choice experiments to inform healthcare decision making: a user's guide. *PharmacoEconomics* 2008, 26(8):661-677.
16. Mehndiratta SR, Hansen M: Analysis of discrete choice data with repeated observations: comparison of three Techniques in intercity travel Case. *Transp Res Rec* 1997, 1607:69.
17. Marshall DA, Johnson FR, Phillips KA, Marshall JK, Thabane L, Kulin NA: Measuring patient preferences for colorectal cancer screening using a choice format survey. *Value Health* 2007, 10(5):415-430.
18. Colorectal Cancer Association of Canada. , [http://www.colorectal-cancer.ca/en/just-the-facts/colorectal/] and [http://www.colorectal-cancer.ca/en/screening/fobt-and-fit/].
19. Anonymous From the Centers for Disease Control and Prevention: Colorectal cancer test use among persons aged > or = 50 years–United States, 2001. *JAMA* 2003, 289(19):2492-2493.
20. Walsh JM, Terdiman JP: Colorectal cancer screening: scientific review. *JAMA* 2003, 289(10):1288-1296.
21. Slomski A: Expert panel offers advice to improve screening rates for colorectal cancer. *JAMA* 2010, 303(14):1356-1357.
22. Labianca R, Merelli B: Screening and diagnosis for colorectal cancer: present and future. *Tumori* 2010, 96(6):889-901.
23. Kuhfeld WF: Discrete choice (SAS Technical Papers: Marketing research, MR2010F). [http://support.sas.com/techsup/tech note/mr2010f.pdf]. (Date of last access: January 7, 2012).
24. Street DJ, Burgess L: Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. *Journal of Statistics Planning and Inference* 2004, 118:185-199.
25. Louviere J, Flynn F, Carson R: Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* 2010, 2(2):57-72.
26. Neuhaus JM: Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res* 1992, 1(3):249-273.
27. Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, Fisher MR: A survey of methods for analyzing clustered binary response data. *Int Stat Rev* 1996, 64:89-118.
28. Huber S, Ervin LH: Using heteroscedastic consistent standard errors in the linear regression model. *Am Stat* 2000, 54:203-206.
29. Rogers W: Regression standard errors in clustered samples. *Stata Technical Bulletin* 1994, 3:19-23.
30. Larsen K, Petersen JH, Budtz-Jørgensen E, Endahl L: Interpreting parameters in the logistic regression model with random effects. *Biometrics* 2000, 56:909-914.
31. Hedeker D, Gibbons RD, Flay BR: Random-effects regression models for clustered data with an example from smoking prevention research. *J Consult Clin Psychol* 1994, 62(4):757-765.
32. Stata online help. [http://www.stata.com/help.cgi?xtmelogit].
33. Rabe-Hesketh S, Skrondal A: *Multilevel and longitudinal modeling using Stata*. 2 edition. USA: A Stata Press Publication; 2008.
34. Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T: A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Stat Med* 2009, 28(18):2338-2355.
35. Hanley JA, Negassa A, Edwardes MD, Forrester JE: Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003, 157(4):364-375.
36. Ballinger GA: Using generalized estimating equations for longitudinal data analysis. *Organ Res Methods* 2004, 7:127-150.
37. Long JS, Freese J: *Regression models for categorical dependent variables using STATA*. 2 edition. Texas, USA: Stata Press; 2006.
38. Chib S, Greenberg E: Analysis of multivariate probit models. *Biometrika* 1998, 85:347-361.
39. Pizzo E, Pezzoli A, Stockbrugger R, Bracci E, Vagnoni E, Gullini S: Screenee perception and health-related quality of life in colorectal cancer screening: a review. *Value Health* 2011, 14(1):152-159.
40. Cheng S, Long J: Testing for IIA in the multinomial logit model. *Sociological Methods Research* 2007, 35(4):583-600.
41. Crouchley R, Ganjali M: A comparison of GEE and random effects models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binary event series. *Stat Model* 2002, 2:39-57.
42. Hedeker D: A mixed-effects multinomial logistic regression model. *Stat Med* 2003, 22(9):1433-1446.
43. Daganzo C: *Multinomial probit: the theory and its application to demand forecasting*. New York: Academic; 1979.
44. Kaplan D, Venezky RL: Literacy and voting behavior: a bivariate probit model with sample selection. *Soc Sci Res* 1994, 23:350-367.
45. Watt DI, Kayis B, Willey K: The relative importance of tender evaluation and contractor selection criteria. *International Journal of Project Management* 2010, 28:51-60.
46. Veitman SJ, Hilsden RJ, Au T, Dowden S, Manns BJ.: Colorectal cancer screening for average-risk North Americans: an economic evaluation. *PLoS Med* 2010, 7(11):e1000370.
47. Nelson RL, Schwartz A: A survey of individual preference for colorectal cancer screening technique. *BMC Cancer* 2004, 4:76.
48. Campbell MJ, Donner A, Klar N: Developments in cluster randomized trials and Statistics in Medicine. *Stat Med* 2007, 26(1):2-19.
49. Ma J, Thabane L, Kaczorowski J, Chambers L, Dolovich L, Karwalajtys T, Levitt C: Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: the Community Hypertension Assessment Trial (CHAT). *BMC Med Res Methodol* 2009, 9:37.
50. Schuster YH, Grohn YT, McDermott B, McDermott JJ: Analysis of correlated discrete observations: background, examples and solutions. *Prev Vet Med* 2003, 59(4):223-240.
51. Neuhaus JM, Kalbfleisch JD, Hauck WW: A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev* 1991, 59:25-35.
52. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, Bruckner T, Satariano WA: To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010, 21(4):467-474.
53. Munizaga MA, Heydecker BG, de Dios Ortúzar J: Representation of heteroskedasticity in discrete choice models. *Transp Res* 2000, 34:219-240.
54. Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JA: Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *Int J Epidemiol* 2003, 32(5):840-846.
55. Heil SP: A review of multilevel and longitudinal modeling using stata. *J Educ Behav Stat* 2009, 34:559-560.
56. Burda M, Harding M, Hausman J: A Bayesian mixed logit probit model for multinomial choice. *J Econ* 2008, 147:232-246.

57. Ferguson S, Olewnik A, Cormier P: *Proceedings of the exploring marketing to engineering information mapping in mass customization: a presentation of ideas, challenges and resulting questions: August 28-31; Washington, DC, USA* USA: ASME; 2011.
58. Kato H: *Proceedings of the non-linearity of utility function and value of travel time savings: empirical analysis of inter-regional non-business travel mode choice of Japan: September 18-20; Strasbourg* European Transport Conference: France; 2006.
59. Kjaer T, Bech M, Gyrd-Hansen D, Hart-Hansen K: **Ordering effect and price sensitivity in discrete choice experiments: need we worry?** *Health Econ* 2006, 15(11):1217-1228.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2288/12/15/prepub

doi:10.1186/1471-2288-12-15
Cite this article as: Cheng *et al.*: **An empirical comparison of methods for analyzing correlated data from a discrete choice survey to elicit patient preference for colorectal cancer screening.** *BMC Medical Research Methodology* 2012 12:15.

CHAPTER 3

THE IMPACT OF INCLUDING OR EXCLUDING BOTH-ARMED ZERO-EVENT STUDIES ON USING STANDARD META-ANALYSIS METHODS FOR RARE EVENT OUTCOME: A SIMULATION STUDY

Ji Cheng [1, 2], Eleanor M Pullenayegum [1, 3], John K Marshall [4,], Alfonso Iorio [1,4], Lehana Thabane [1, 2]

[1] Department of Clinical epidemiology & Biostatistics, McMaster University, Hamilton, ON, Canada.

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

[3] Hospital for Sick Children, Toronto ON, Canada

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

Keywords: Both-armed zero-event, Meta-analysis, rare event outcome, simulation

Word count: 3941

Corresponding Author:

Lehana Thabane PhD

Tel: 905.522.1155 x 33720

Fax: 05.308.7386

email: thabanl@mcmaster.ca

St. Joseph's Healthcare Hamilton

3rd Floor, Martha Wing, Room H-325

50 Charlton Avenue East

Hamilton ON L8N 4A6

Abstract

Objectives: There is no consensus on whether studies with no observed events in both the treatment and control arms, the so-called both-armed zero-event studies, should be included in a meta-analysis (MA) of randomized controlled trials (RCTs). Current analytic approaches handled them differently depending on the choice of effect measures and authors' discretion. Our objective is to evaluate the impact of including or excluding both-armed zero-event (BAZE) studies in MA of RCTs with rare outcome events through a simulation study.

Method: We simulated 2500 datasets for different scenarios varying the parameters of baseline event rate, treatment effect and number of patients in each trial, and between-study variance. We evaluated the performance of commonly used pooling methods in classical MA—namely, Peto, Mantel-Haenszel (M-H) with fixed-effects and random-effects models, and inverse variance (IV) method with fixed-effects and random-effects models—using bias, root mean squared error (RMSE), length of 95% confidence interval [CI] and coverage.

Results: The overall performance of the approaches of including or excluding BAZE studies in meta-analysis varied according to the magnitude of true treatment effect.

Including BAZE studies introduced very little bias, decreased MSE, narrowed the 95% CI, and increased the coverage when no true treatment effect existed. However, when a true treatment effect existed, the estimates from the approach of excluding BAZE studies led to smaller bias than including them. Among all evaluated methods, the Peto method excluding BAZE studies gave the least biased results when a true treatment effect existed.

Conclusion: We recommend including BAZE studies when treatment effects are unlikely, but excluding them when there is a decisive treatment effect. Providing results of both including and excluding BAZE studies to assess the robustness of the pooled estimated effect is a sensible way to communicate the results of a MA when the treatment effects are unclear.

Strengths and limitations of this study

- A simulation study thoroughly investigated the impacts of including or excluding both-armed zero-event studies in meta-analyses by comparing all commonly used pooling methods
- The simulation parameters were chosen according to the characteristics of meta-analyses in the Cochrane Database of Systematic Reviews to closely reflex the reality
- Our results not only confirmed the findings from the previous empirical studies but also added more details on how including or excluding both-armed zero-event may impact the estimates of meta-analyses differently depending on the magnitude of true treatment effects
- Only odds ratio was investigated through simulations, thus the findings from this study may not be able to be fully extended to other effect measures such as relative risk or absolute risk difference

Background

Systematic review (SR) with meta-analysis (MA) has become an important research tool for the health research literature which synthesizes evidence from individually conducted studies that assess the same outcomes on the same topic. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement[1] adopted the definition used by Cochrane Group[2] which defines SR as a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review. Meta-analysis refers to the use of statistical techniques in a systematic review to integrate the results of included studies. Therefore, the results of MAs from randomized controlled trials (RCT) are considered to be the best quantitative clinical evidence in the literature.[3,4] Studies included in a SR are selected rigorously according to predefined exclusion and inclusion criteria. Thus all identified studies in a SR with available data should be included in the MA. However, there is no consensus among researchers whether this principle should be fully applied and how to apply to the MAs using dichotomous outcomes.

The outcomes of dichotomous data are events. The number of observed events in a RCT using dichotomous outcomes is most affected by the event rate and sample size, and also affected by the length of the study period. When the event rate is low, the sample size is

small and the study period is short, it is possible that no outcome event is observed in the RCT although the probability of the event happening is not zero. A study with no outcome event observed in either treatment or control arms is called a zero-event study. Both-armed zero-event (BAZE), also called double-zero event or zero-total-event, is an extreme case of zero-event, which is defined as no event is observed in both treatment and control arms.

When rare adverse events or rare diseases are used as the study outcomes, it is not an uncommon phenomenon that no outcome events are observed at the end of the study. In the United States, a rare adverse event is defined as one per 1000 patients.[5] In the European Union, a rare disease is defined as one per 2000 people.[6] To obtain a representative number of outcomes for a rare event study, a large number of patients are needed. However, very often, RCTs are either not designed primarily to investigate adverse events or do not have the resources to recruit the sample size required for such events. A published review of the Cochrane Database of systematic reviews showed that the median sample size for dichotomous outcomes was 102 (inter-quartile range of 50-243).[7] Therefore, when the primary outcome in a MA is a rare event, zero-event studies could be among the qualified studies. Warren (2011) and colleagues conducted a systematic review of meta-analyses published between 1994 and 2006 where rare events were a primary outcome.[8] Among 166 MAs, 65 (39%) included zero-event studies, and

41 (25%) included BAZE studies. Amongst the 41 MAs with BAZE studies, 19 MAs (46%) included them in the primary or sensitivity analyses, 18 (44%) excluded them and 4 (10%) were unclear. This review also found that the continuity correction was most used approach to incorporate zero-event studies, and 0.5 was the common choice of the correction factor (93%).

For single-armed zero-event studies, there is consensus on their inclusion in MAs. Bradburn (2007) and colleagues reported a simulation study comparing commonly used methods of handling zero-event studies in MAs.[9] This provides a good guideline for the subsequent MAs. However, when BAZE studies were present in systematic reviews, the practice of handling varies.[8,10]

There are two major reasons why BAZE are handled variably in meta-analyses. First, the statistical methods and software such as RevMan[11], Stata's metan module[12] and Comprehensive Meta-analysis[13]to handle BAZE studies differ according to the choice of effect measures. BAZE studies are included in the pooled results when risk difference (RD) is used, but automatically excluded by all statistical software used for MA when odds ratio (OR) or relative risk (RR) is used. Second, there is no guideline for handling BAZE studies in MAs. A few published papers examined various approaches using

empirical data have produced ambiguous results. In 2007, Friedrich and colleagues empirically compared the statistical methods of handling BAZE studies in MA and recommended that BAZE studies should be included in all MAs. They concluded that including BAZE studies could narrow the confidence interval and increase the precision of the pooled estimates.[14] In 2008, Dahabreh and colleagues conducted a sensitivity analysis to re-evaluate the treatment effect of Rosiglitazone and found that including BAZE studies changed the pooled odd ratio of myocardial infarction between treatment and control groups from significant to not significant statistically.[15] Although the above empirical studies showed us that including BAZE studies could impact the results of MAs, the impact may not be beneficial towards the truth in all scenarios. In addition to the above empirical studies, a recently published simulation study argued that incorporating BAZE studies using a relatively complicated Beta-binomial regression could generate unbiased estimates for MAs.[16] However, due to its complexity and lack of available procedures in commonly used statistical software, this model may not a practical choice.

Since number of events observed in studies using dichotomous outcomes is determined by event rates and number of subjects, zero-events are more likely to occur with the conditions of extremely low event rates or small sample sizes even though the event rates are different between two study groups. In the intuitive way, the arithmetical difference

between two study groups with no observed events is null. Therefore, we believe that depending on the magnitude of true treatment effects, including BAZE studies in MA may affect the pooled estimates of treatment effects in two different ways. When there is no true treatment effect, i.e. the event rates are similar in treatment and control arms, including BAZE studies can narrow the confidence interval of the pooled studies of a MA. But on the other hand, we suspect that when a true treatment effect exists, including BAZE studies could moderate the magnitude of the pooled estimate and lead to the underestimation of the treatment effect.

To test this hypothesis, we conducted a simulation study to evaluate the impact of excluding and including BAZE studies. Although it is not difficult to statistically deduct deduce that the bias brought by including BAZE studies is affected by the following factors: 1) low event rate, 2) large treatment effect, and 3) small sample size, stimulation is still needed to quantifying the magnitude of the bias. Our investigation was focused on comparing the statistical pooling methods adopted by the commonly used software such as RevMan and Stata for meta-analyzing aggregated data.

Method

Odds ratio and relative risk are the most commonly used effect measures for assessing the treatment effect for dichotomous outcomes in meta-analyses. The results of these two effect measures are similar when the event probability is less than 20%.[16,17] Since the event rates used in our simulation study were much lower, we chose OR as the effect measure to engage the Peto method in our investigation. Bradburn (2007) *et al* have shown that the Peto method was a better choice for rare event meta-analyses for dichotomous outcomes when only one-armed zero event studies were included.[9]

Simulation Scenarios

The simulation scenarios in our study were chosen based on a combination of several simulation parameters. Three types of parameters were used in this simulation study: fixed, varied and derived. We believed some parameters had more impact on the simulation results than others. We chose fixed values for the low impact parameters across all simulation scenarios and let the values of those high impact parameters vary in certain ranges. The parameter values were drawn from the published literature (Table 1). The derived parameters were calculated by the input parameters according to a statistical formula. For the fixed parameters, we tested the following values. The numbers of studies (m) in each MA was set at 5. The review published in 2011 reported that the median

(interquartile) of the numbers of studies included in the meta-analysis in the Cochrane Database was 3 with inter quartile range (IQR) from 2 to 6.[7] For the treatment and control arm ratio (r), we only considered 1:1 allocation. A review paper have shown that 78% of clinical trials were conducted with equal patient allocation strategies.[19] To reduce the number of simulation scenarios, we deliberately chose to use the same number of patients across all studies in each MA.

For the following parameters, we chose to input multiple values instead of constants. The control arm event probabilities (p) investigated in this simulation were 0.001, 0.005, and 0.01. They are chosen according to the varying definitions of rare events.[5,6] The treatment effects measured as odds ratio (OR) were set as no effect (OR =1), medium sized (0.8), large (0.5) and extremely large (0.2). [20] The numbers of patients (n) in each individual study included were 50, 100 and 200 based on the same review mentioned above,[7] which revealed that the median (Q1, Q3) of the sample size in each individual study was 102 (100, 243). We also considered the potential impact of between study variance in our simulation design. We set the between study standard deviation (SD) as 0.1, 0.5 and 1, which represented little, moderate and large between study variance.[20] The between study variation was added in the OR, i.e. the treatment effect.

In this simulation study, the treatment arm event probabilities were calculated through the control arm event probabilities and treatment effects (OR).

$$p_{T_i} = \frac{\left(\frac{p_C}{1-p_C)}\right)^{\Omega_i}}{1 + (\frac{p_C}{1-p_C})^{\Omega_i}}$$

Where p_T = treatment arm probability, p_C = control arm probability, Ω = odds ratio, i = 1, 2, ..., study.

Number of simulations

We simulated 2500 data sets for each scenario to ensure the accuracy of our simulation results.[21]

Analysis Methods

Five pooling procedures were used to meta-analyze each simulated data set. They were Peto, Mantel-Haenszel (M-H) with fixed-effects and random-effects models, and inverse variance (IV) method with fixed-effects and random-effects models.[2]

Methods to Including Both-armed Zero-events

To implement the above 5 pooled methods to incorporate studies with BAZE in MA, a continuity correction factor was add to each of the four cells of the 2 x 2 table for a BAZE study, i.e. event in the treatment arm, non-event in the treatment arm, event in the control arm, and non-event the in control arm. We chose to use the constant continuity factor 0.5. It is common and plausible choice when the group ratio is balanced between treatment and control arms.[22]

Evaluating simulation performance

Four measures were used to assess the performance of this simulation study [21] (Table2): 1) percentage bias, which is calculated as the percentage of the difference between the average of the estimated value and the true value (absolute bias) over the true value; 2) root mean square error (RMSE), which measures the average distance of estimated treatment effects from the parameter value; 3) the average length of 95% confidence intervals (CI) is also used to compare the precisions between pooling methods; 4) coverage, which measures the percentage of the true treatment effects included in the available 95% confidence intervals (CI) over all generated data sets. The RMSE and average 95% CI length were reported the log OR scale. The performances of the simulation were compared across the five pooling methods used for the approaches of including and excluding BAZE studies in the meta-analyses. We also reported the

inclusiveness of the approach of excluding BAZE studies in MA, which reported the percentage of number of studies included in the pooling process.

Statistical Software and Program

The data sets for each simulation scenario are generated using R 2.15.2 (The R Foundation for Statistical Computing). The meta-analyses were conducted using Stata 13.1 (College Station, TX). The estimates summarizing the overall performance of this simulation were also calculated using Stata.

Results

In this study, we ran 57 simulated scenarios. The scenarios were grouped to investigate the impact of the value changing on the following variable parameters while holding the number of studies (m = 5) and allocation ratio (1:1) fixed: i) the treatment effect (OR = 1, 0.8, 0.5, 0.2), ii) the control arm event probability (p= 0.001, 0.005, 0.01), the number of patients in each individual study (n = 50, 100, 200) and the between-study standard deviation (SD = 0.1, 0.5, 1). When examining the changes on one variable parameter, we held the other variable parameters on the common scenario, which was set as (OR = 0.5, p = 0.001, n = 100 and between-study SD = 0.5). We assessed the simulation results by

comparing bias, RMSE, the length of 95% CI, and coverage. We also reported the inclusiveness of the approach of excluding BAZE studies

Including BAZE studies

Our simulation results supported our hypothesis that when there is no true treatment effect (OR =1), the approach of including BAZE studies in meta-analyses had the best overall performance regardless of the choice of pooling methods, which gave the smallest bias (<0.1%) (Table 3a) and RMSE (Figure 1), and narrowest 95% CI (Figure 2). However, when there was true treatment effect, this approach gave the larger bias compared to the alternative approach of excluding BAZE studies. The magnitude of the bias increased with an increase in the treatment effect. Compared to the approach of excluding BAZE studies, the result obtained by including BAZE studies had smaller RMSEs when the treatment effects were small (OR =0.8) or moderate (0.5), but when the treatment effect was large (OR = 0.2), RMSEs were also larger (Figure 1). The changes of the treatment effect also impacted the coverage. For all methods, the coverage was high (> 99%) when the treatment effect was zero (OR = 1) to moderate (OR = 0.5), but then dropped to 95% when the treatment effect was large (OR = 0.2). We also found that the bias of the pooled estimates increased with decreasing control arm probability (Table 3b) and number of patients (Table 3c) and increasing between-study variance (Table 3d).

Excluding BAZE studies

Similarly excluding BAZE studies for meta-analyses introduced little bias on the pooled estimates (0.7-1.4%) when there was no true treatment effect (Table 3a) When a true treatment effect existed, the pooled estimates obtained using this approach yielded smaller bias compared to including BAZE studies. Again the magnitude of bias increased with a decrease in the control arm probability (Table 3b) and number of patients (Table 3c) and an increase in between-study variance (Table 3d). We also noticed that excluding BAZE studies didn't have much impact on RMSE, the length of confidence intervals and coverage, except for Peto method—which had slightly wider confidence interval (Figure 2) and lower coverage (91%) when large treatment effect presented (OR =0.2). However, the inclusiveness, i.e. the number of studies included in MA dropped noticeably (72%, 67%, 59%, 46%) with the increase of the treatment effects (OR = 1, 0.8, 0.5, 0.2), respectively.

Peto method excluding BAZE studies

Among all five pooling methods, the Peto method excluding BAZE studies provided the most reliable results (percentage bias < 0.8) for this rare event setting (control arm probability = 0.001, 0.005, 0.01) when the true treatment effect and between-study variance were small to moderate and number of patients were equal or greater than 100 in each individual study (Table 3a-3d).

In summary, our simulation study verified that when there was no true treatment effect (OR = 1), the approach of including BAZE studies consistently outperformed the approach of excluding BAZE studies across all five pooling methods by providing less biased results with smaller RMSE, narrower 95% CI and higher coverage regardless of the changes of control arm probability, number of patients and between-study variance. However, whenever a true treatment effect was present, the results from the approach of including BAZE studies introduced larger bias than the approach of excluding them.

Discussion

This simulation study investigated the impact of including or excluding BAZE studies in MAs for rare event outcomes when odds ratio is used as the effect measure for pooled estimates of dichotomous outcomes. We found that including BAZE studies provided more accurate overall pooled estimates than excluding them when there was no true treatment effect. However, when there was a true treatment effect, the results from both approaches underestimated the true treatment effect, and including BAZE studies increased bias further. Amongst the pooling methods, Peto's method with exclusion of BAZE studies provided the pooled OR considerably closer to the true treatment effect for small to moderate treatment effects under the condition of small to moderate between-study variance and relatively large sample size.

Our simulation study confirmed the empirical findings obtained by Friedrich et al. (2007). They recommended including BAZE studies in all meta-analyses for the benefits of providing conservative point estimates and increasing the study integrity.[14] However, the "conservative" estimate is a double-edged sword. In the sense of drawing the estimates towards null hypothesis, although underestimating benefit may delay or deny patient's access to a new treatment[23] when evaluating the beneficial treatment effect for a new drug, with the patient safety as physician's priority concern, the conservative result might be a the safer choice. With many uncertainties unchecked, quickly shifting from the standard care to a new treatment based on the findings from a small study (even it is a MA) can be a dangerous move. Some studies have showed that the treatment effect tend to be over estimates when the trials were underpowered.[21,22] On the other hand, when the result of a MA is regarding the safety measures such as serious adverse event, the conservative result means underestimating the harm, which could lead to expose patients to unnecessary danger.[26] Therefore, depending the purpose of the SR (evaluating benefits or harms), including BAZE studies in MA could have different implications.

This simulation study confirmed that among all five commonly used pooling methods, only the Peto method without inclusion of BAZE studies produces a pooled OR approaching the true treatment effect when sample size are relatively large. This finding is consistent with the simulation study conducted by Bradburn *et al* (2007),[9] which

evaluated performances of the common methods used to meta-analyze the sparse data for binary outcomes. In addition to their findings, our simulation study also shows that compared to the random-effects model (IV or H-M), the Peto method as a fixed-effect model gave the least biased estimates when the between-study variance is from small to moderate.. The reason of the Peto method outperforming the random-effects model is that as Sweeting *at el* [22]has shown in their simulation study, the heterogeneity was difficult to estimate for the rare event data. Therefore, the benefit of using random-effect model doesn't overcome the bias introduced by the IV or H-M methods, which were proven by the simulation study conducted by Bradburn and *at el.*[9]

This simulation study clearly showed that including both-armed (and even single armed) zero-event studies in MA could do more harm than benefit when the treatment effect is comparing harmful outcomes. However, in reality, it is not easy or sometimes even impossible to know whether a true treatment effect exists or not. Therefore, a comprehensive approach of a series of sensitivity analyses need to be conducted when performing systematic reviews that include zero-event studies. An example could be used is Dahabreh at el (2008) who re-analyzed the cardiovascular events in randomized trials of rosiglitazone.[15] Although, the results showed that including BAZE studies turned the pooled odds of myocardial infarction (MI) from statistically significant to not significant. Their conclusion that rosiglitazone increased MI was made after assessing the

consistency of results from different methods. The above example demonstrates that when MAs are conducted to evaluate rare events, it is difficult to get a concordant result. To assist readers to make their own informative decision about the results of a MA, its methods should be communicated in full transparency. In addition to reporting the result following the PRISMA guideline,[1] the eligible studies with zero-event and the methods used to deal with zero-event studies need to be clearly described. We believe that an extension of the PRISMA guideline on how to report MAs on rare event outcomes with zero event studies needs to be developed to include a section of reporting the methods used to deal with zero-event studies and impact on the overall estimates of MAs.

Although we chose the values of simulation parameters from literature review, we realize that the results of our simulation study cannot be generalized to all situations in MA. To reduce the simulation scenarios to a manageable level, we used fixed values for some parameters. We only considered the balanced group ratio between treatment and control arms, but only 22% of RCTs used unbalanced design among previous in a recent review.[19] Within each simulated MA data set, we fixed the number of studies to five, each with the same number of patients. This approach might be over simplified. Although we chose to investigate OR using common pooling methods, we believe that our findings can be applied to RR under similar condition for the estimates of OR and RR are similar when event rates are less than 0.2.[16,17] For the continuity correction

approach to incorporate zero-event studies, we only used 0.5 as continuity correction factor, which works well when the trial arms are balanced, but will increase the bias when there is a big difference on the numbers of patients between two arms and the treatment effect are large. [22]

The commonly used MA pooling methods we discussed in this simulation are based on parameter estimation, which requires the use of continuity correction to include zero events. The likelihood maximization based Poisson Regression can incorporate zero events without continuity correction and supposedly generates an unbiased estimate of RR. The simulation from Spittal and *et al*[27] showed that random-effects Poisson Regression outperformed the standard pooling methods when meta-analyzing the incidence rate ratio for zero events data. We ran the random-effects model Poisson Regression on our stimulated data, and there was a convergence issue. The reason could be that there were a large proportion of zero-event (either in one arm or both arms) studies presented in a relatively smaller number of studies in each MA due to extremely low event rate. This convergence problem may not be a problem for MAs with larger number of studies. However, the most commonly used MA software such as RevMan doesn't have the capacity to conduct any advanced statistical model, which may present a challenge for researchers who use the standard MA analysis packages. Similar to random-effect Poisson Regression, Bayesian approach using none-informative prior as an

alternative of the standard classical MA method we investigated in this study has the advantage of incorporating zero-event studies without applying a continuity correction. [22] How including BAZE studies in Bayesian MA impacts the pooled estimates will be studied in subsequent simulations.

Conclusion

To conclude, we recommend including BAZE studies in MA using OR as effect measure when treatment effects are unlikely to preserve data integrality of the systematic review. When treatment effects are clearly present, excluding BAZE studies and using the Peto method is a safer choice for evaluating rare events. However, most of the time, the real situation about the treatment effect is hard to foresee from the available data, it is important to conduct sensitivity analyses using alternative approaches to assess the robustness of the primary analysis. And the purpose of the SR also need to be considered when deciding on how to deal with BAZE studies in MAs. Furthermore, the results of MAs for rare events need to be interpreted within the clinical content.

Contributors: JC, EMP, JKM and LT designed the study. JC wrote the simulation program, conducted the statistical analysis and drafted the manuscript. JC, EMP, IA and LT provided input on statistical concept. JKM and IA provided the clinical expertise. All authors revised the manuscript for important clinical and statistical contents and approved the final manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors

Competing interests: There is no competing interest declared by any authors.

Data sharing statement: No additional data are available from this simulation study.

Reference

1 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Phys Ther* 2009;**89**:873–80. doi:10.1136/bmj.b2535

2 Cochrance Group. Cochrane handbook: Meta-analysis of dichotomous outcomes. http://handbook.cochrane.org/chapter_9/9_4_4_meta_analysis_of_dichotomous_outcomes.htm

3 Evans D. Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs* 2003;**12**:77–84. doi:10.1046/j.1365-2702.2003.00662.x

4 Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;**126**:376–80.http://www.ncbi.nlm.nih.gov/pubmed/9054282

5 Marodin G, Goldim JR. Confusions and ambiguities in the classification of adverse events in the clinical research. *Rev Esc Enferm USP* 2009;**43**:690–6.

6 Institution of Medicine. Adverse Drug Event Reporting: The Roles of Consumers and Health Care Professionals. In: *Workshop Summary 2007*.

7 Davey J, Turner RM, Clarke MJ, *et al.* Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol* 2011;**11**:160. doi:10.1186/1471-2288-11-160

8 Warren FC, Abrams KR, Golder S, *et al.* Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC Med Res Methodol* 2012;**12**:64. doi:10.1186/1471-2288-12-64

9 Bradburn MJ, Deeks JJ, Berlin J a., *et al.* Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Stat Med* 2007;**26**:53–77. doi:10.1002/sim.2528

10 Keus F, Wetterslev J, Gluud C, *et al.* Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. *Am J Gastroenterol* 2009;**104**:546–51. doi:10.1038/ajg.2008.22

11 The Cochrane Collaboration. RevMan 5.1 User Guide. 2011.

12 Harris RJ, Bradburn MJ, Deeks JJ, *et al.* Metan: Fixed- and random-effects meta-analysis. *Stata J* 2008;**8**:3–28.

13 Altman D, Duval S, Lipsey M, *et al.* Comprehensive Meta Analysis Version 3 . 0.

14 Friedrich JO, Adhikari NKJ, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol* 2007;**7**:5. doi:10.1186/1471-2288-7-5

15 Dahabreh IJ, Economopoulos K. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. *Clin Trials* 2008;**5**:116–20. doi:10.1177/1740774508090212

16 Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Stat Med* 2015;**34**:1097–116. doi:10.1002/sim.6383

17 Grimes D a, Schulz KF. Making sense of odds and odds ratios. *Obstet Gynecol* 2008;**111**:423–6. doi:10.1097/01.AOG.0000297304.32187.5d

18 Egger M, Smith GD, Phillips a N. Meta-analysis: principles and procedures. *BMJ* 1997;**315**:1533–7. doi:10.1136/bmj.315.7121.1533

19 Dumville JC, Hahn S, Miles JN V, *et al.* The use of unequal randomisation ratios in clinical trials: A review. *Contemp Clin Trials* 2006;**27**:1–12. doi:10.1016/j.cct.2005.08.003

20 Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. 1988.

21 Burton A, Altman D, Royston P, *et al.* The design of simulation studies in medical statistics. *Stat Med* 2006;**25**:4279–92.

22 Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med* 2004;**23**:1351–75. doi:10.1002/sim.1761

23 Hayward R a, Kent DM, Vijan S, *et al.* Reporting clinical trial results to inform providers, payers, and consumers. *Health Aff (Millwood)*;**24**:1571–81. doi:10.1377/hlthaff.24.6.1571

24 Nuesch E, Trelle S, Reichenbach S, *et al.* Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. 2010;**341**. doi:10.1136/bmj.c3515

25 Zhang Z, Xu X, Ni H. Small studies may overestimate the effect sizes in critical care meta-analyses : a meta- epidemiological study. *Crit Care* 2013;**17**:R2. doi:10.1186/cc11919

26 Miller DB, Humphries KH. A new way to evaluate randomized controlled trials ? New approach does more harm than good. 2005;:241–4.

27 Spittal MJ, Pirkis J, Gurrin LC. Meta-analysis of incidence rate data in the presence of zero events. *BMC Med Res Methodol* 2015;**15**:42. doi:10.1186/s12874-015-0031-0

Table 1 Simulation Parameter Setup

Parameter	Assigned Value	Rational	Reference
Odds Ratio (OR)	1, 0.8, 0.5, 0.2	Small, medium and large treatment effect	Cohen J. Statistical Power Analysis for the Behavioral Sciences. Second ed. Hillsdale, NJ: Erlbaum; 1988.
Control group event probability (p)	0.001,0.005,0.01	1 in 2000 rare disease in EU; 1 in 1000 rare adverse event;	1. Marodin G, Goldim JR. Confusions and ambiguities in the classification of adverse events in the clinical research. Rev Esc Enferm USP 2009;43:690–6. 2. Institution of Medicine. Adverse Drug Event Reporting: The Roles of Consumers and Health-Care Professionals: Workshop Summary (2007)
Number of studies in each meta-analysis (m)	5	median = 3; interquartile: 2-6; less than 1% >29	Davey J. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC 2011
Number of patients in individual study (n)	50, 100, 250	Median=102; interquartile 50-243	Davey J. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis; BMC Medical Research Methodology 2011, 11:160
Between study standard deviation (SD)	0.1, 0.5, 1	Small, moderate, large	Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. 1988
Ratio of group size (r)	1:1	78% trials had equal group ratio	Dumvillev JC. The use of unequal randomisation ratios in clinical trials: A review. Contemporary Clinical Trials 27 (2006) 1-12

Table 2 Measures for evaluating simulation performance

criteria	Formula
Percentage bias ($(\frac{\delta}{\beta})$ %)	$\left(\frac{\overline{\hat{\beta}} - \beta}{\beta}\right) \times 100$
Root mean square error (RMSE)	$\sqrt{(\left(\overline{\hat{\beta}} - \beta\right)^2 + \left(SE\ (\hat{\beta})\right)^2)}$
Average length of 95% CI	$\frac{\sum_{i=1}^{B} 2Z_{1-\alpha/2} SE(\hat{\beta}_i)}{B}$ for i = 1, 2 ,…, B, where B = the number of meta-analyses conducted using simulated data sets
Coverage of 95% CI	Percentage of times the 95% CI of $\hat{\beta}_i$ include β, for i = 1, 2 ,…, M, where M = the number of meta-analyses conducted using simulated data sets
Inclusiveness	Average percentage of number of studies included in the meta-analysis.

β: the true value of estimate of interest; $\hat{\beta}$: estimate of β; $\overline{\hat{\beta}}$: mean of $\hat{\beta}$ in simulation

δ: bias

SE: standard error

$Z_{1-\alpha/2}$: $(1 - \alpha/2)$ quantile of the standard normal distribution

Table 3a Impact of the treatment effect changes on bias

Number of studies = 5	Number of patients = 100	Group ratio = 1	Control arm probability = 0.001	Number of simulated data sets = 2500	Between-study SD = 0.5

	Excluding BAZE studies								**Including BAZE studies**							
	OR = 1		OR 0.8		OR =0.5		OR = 0.2		OR = 1		OR 0.8		OR =0.5		OR = 0.2	
Methods	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$
IV Random effects	1.01	0.8	0.88	-9.9	0.70	-40.6	0.47	-133.1	1.00	< 0.1	0.99	-23.2	0.97	-93.0	0.94	-370.7
IV Fixed effects	1.01	0.7	0.88	-9.9	0.70	-40.6	0.47	-133.1	1.00	<0.1	0.98	-23.0	0.96	-92.3	0.93	-367.4
M-H Radom effects	1.01	0.8	0.88	-9.9	0.70	-40.6	0.47	-133.1	1.00	<0.1	0.99	-23.2	0.97	-93.0	0.94	-370.7
M-H Fixed effects	1.01	0.8	0.88	-9.9	0.70	-40.6	0.47	-133.1	1.00	<0.1	0.98	-23.0	0.96	-92.3	0.93	-367.4
Peto	1.01	1.4	0.80	0.2	0.54	-7.8	0.26	-30.6	1.00	<0.1	0.95	-22.6	0.90	-90.6	92.2	-360.9

Note: $\overline{\hat{\beta}}/\beta$ = 1 + Taylor expansion (bias_log); bias_log: bias calculated on log scale.

Table3b Impact of the control arm probability changes on bias

Number of studies = 5	Number of patients = 100		Group ratio = 1		OR = 0.5		Number of simulated data sets = 2500				Between-study SD = 0.5	
Methods	**Excluding BAZE studies**						**Including BAZE studies**					
	pc = 0.001		pc = 0.005		pc = 0.01		pc = 0.001		pc = 0.005		pc = 0.01	
	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$
IV Random effects	0.70	-40.6	0.68	-35.1	0.64	-28.5	0.97	-93.0	0.85	-70.8	0.76	-51.3
IV Fixed effects	0.70	-40.6	0.67	-34.9	0.64	-27.3	0.96	-92.3	0.84	-68.5	0.74	-48.0
M-H Radom effects	0.70	-40.6	0.68	-35.1	0.64	-28.5	0.97	-93.0	0.85	-70.7	0.76	-51.3
M-H Fixed effects	0.70	-40.6	0.67	-34.9	0.64	-27.3	0.96	-92.3	0.84	-68.5	0.74	-48.0
Peto	0.54	-7.8	0.52	-4.6	0.51	-1.1	0.90	-90.6	0.80	-59.5	0.67	-33.2

Table 3c: Impact of the number of patient changes on bias

Number of studies = 5	Control group probability = 0.001		Group ratio = 1		OR = 0.5		Number of simulated data sets = 2500				Between-study SD = 0.5	
Methods	**Excluding BAZE studies**						**Including BAZE studies**					
	n = 50		n = 100		n = 200		n = 50		n = 100		n = 200	
	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$
IV Random effects	0.73	-45.7	0.70	-40.6	0.68	-36.5	0.98	-96.8	0.97	-70.7	0.93	-86.0
IV Fixed effects	0.73	-45.8	0.70	-40.6	0.68	-36.3	0.98	-96.5	0.96	-68.5	0.92	-84.5
M-H Radom effects	0.73	-45.7	0.70	-40.6	0.68	-36.5	0.98	-96.8	0.97	-70.7	0.93	-86.0
M-H Fixed effects	0.73	-45.8	0.70	-40.6	0.68	-36.3	0.98	-96.5	0.96	-68.5	0.92	-84.5
Peto	0.58	-15.2	0.54	-7.8	0.51	-2.4	0.98	-95.8	0.95	-59.5	0.90	-80.7

Table 3d: Impact of the between-study variance changes on bias

Number of studies = 5	Control group probability = 0.001		Group ratio = 1		OR = 0.5		Number of simulated data sets = 2500		number of patients per arm = 100			
Methods	**Excluding BAZE studies**						**Including BA0E studies**					
	SD = 0.1		SD = 0.5		SD = 1		SD = 0.1		SD = 0.5		SD = 1	
	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$	$\overline{\widehat{OR}}$	$(\frac{\delta}{\beta})\%$
IV Random effects	0.68	-35.3	0.70	-40.6	0.88	-76.7	0.96	-92.5	0.97	-93.0	0.99	-97.3
IV Fixed effects	0.68	-35.3	0.70	-40.6	0.88	-76.7	0.96	-91.6	0.96	-92.3	0.99	-97.0
M-H Radom effects	0.68	-35.3	0.70	-40.6	0.88	-76.7	0.96	-92.5	0.97	-93.0	0.99	-97.3
M-H Fixed effects	0.68	-35.3	0.70	-40.6	0.88	-76.7	0.96	-91.6	0.96	-92.3	0.99	-97.0
Peto	0.50	-0.9	0.54	-7.8	0.80	-60.5	0.95	-89.9	0.90	-90.6	0.98	-96.4

Figure 1: Comparing root mean square error (RMSE)

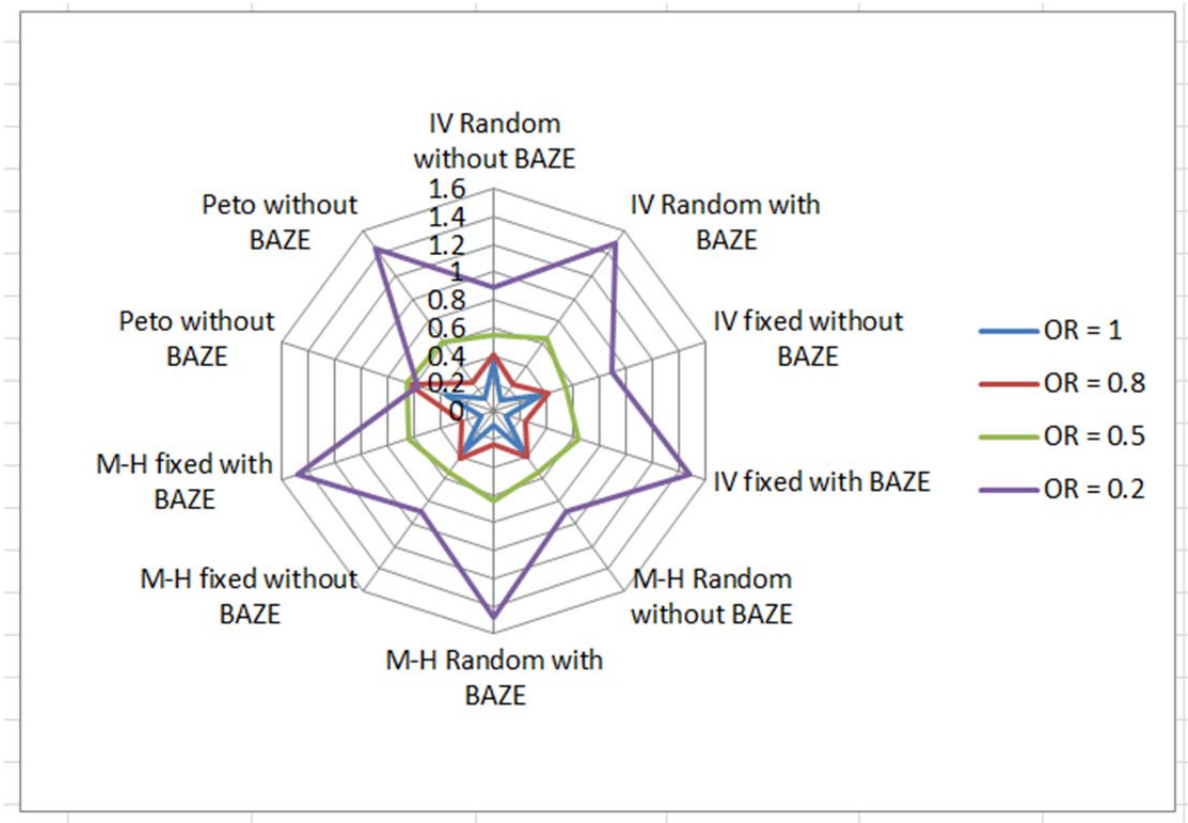
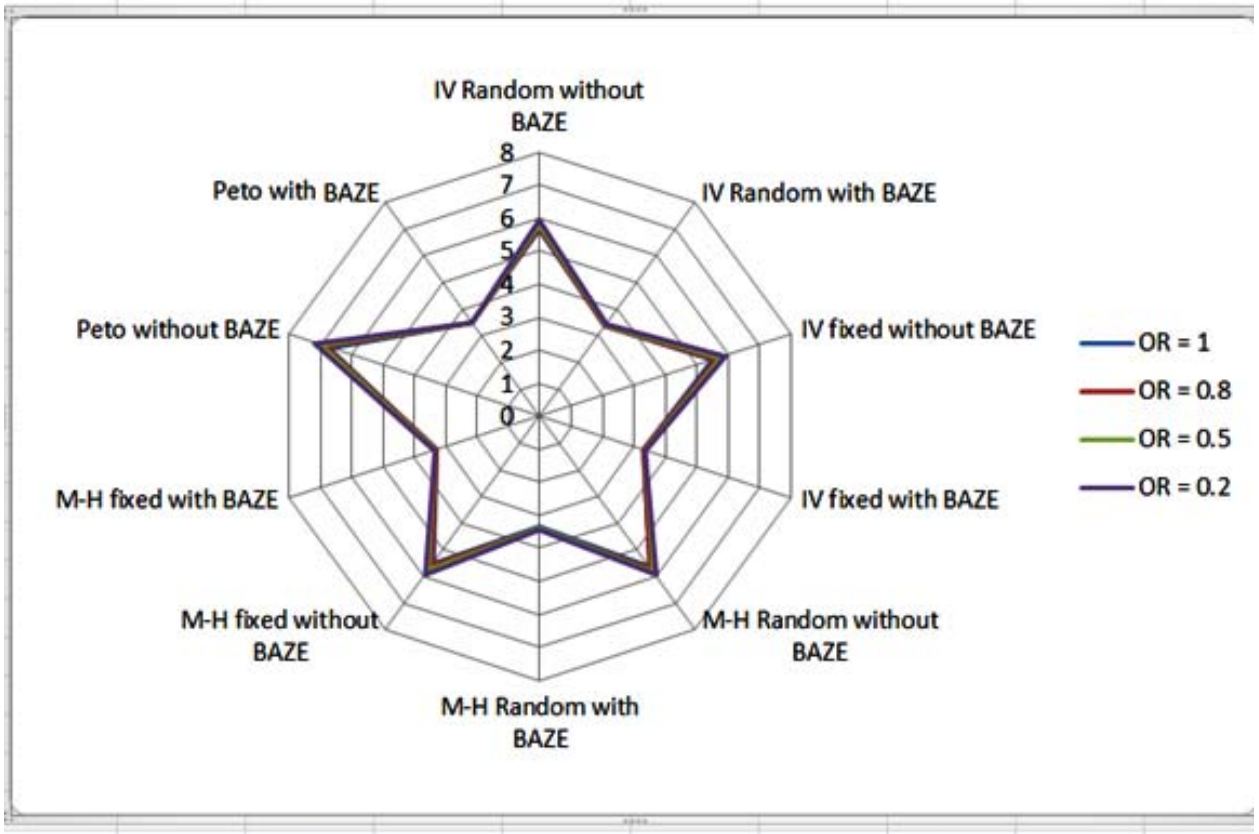


Figure 2: Comparing length of 95% confidence interval (CI)



CHAPTER 4

BAYESIAN APPROACH TO THE ASSESSMENT OF THE POPULATION SPECIFIC RISK OF INHIBITORS IN HEMOPHILIA A PATIENTS: A PRIMER FOR CLINICIANS

Ji Cheng,[1,2] Alfonso Iorio,[2,3] Maura Marcucci,[4] Vadim Romanov,[5] Eleanor M Pullenayegum,[6,7] John K Marshall, [3,8] Lehana Thabane[1,2*]

Affiliations:

1) St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada

2) Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada

3) Department of Medicine, McMaster University, Hamilton, ON, Canada

4) Geriatrics, Fondazione Ca' Granda Ospedale Maggiore Policlinico & Università degli Studi di Milano, Milan, Italy

5) Baxter HealthCare, Global Medical Affairs, Westlake Village, California, USA

6) Child Health Evaluation Sciences, Hospital for Sick Children, Toronto, ON, Canada

7) Dalla Lana School of Public health, University of Toronto, Toronto, ON, Canada

8) Hamilton Health Science, Hamilton, ON, Canada

Keywords: inhibitor rate; meta-analysis; multicentric study; Bayesian; hemophilia A

Word count: 5325

*Corresponding author

Lehana Thabane, PhD

Biostatistics Unit/FSORC, St Joseph's Healthcare--Hamilton

3rd Floor Martha Wing, Room H325

50 Charlton Avenue East, Hamilton, Ontario L8N 4A6, CANADA

Telephone: 1-905-522-1155 ext. 33720/34905; Office Fax: 1-905-528-7386

Email Address: thabanl@mcmaster.ca

ABSTRACT

Background: Bayesian modelling empowers analysis of rare events via incorporation of external data. To illustrate how the approach will i) compare with classical one; ii) change with different priors; and enable testing iii) thresholds and iv) size of information.

Methods: We used three different scenarios: s1) a single cohort of previously treated patients (PTPs), s2) a meta-analysis of PTPs cohorts, and s3) a previously unexplored clinical setting (patients with positive inhibitor history). Patient population: Hemophilia A patients from the ADVATE Post Authorization Surveillance Studies. Outcome: Any inhibitors. Statistical analysis: Non-informative and informative priors were applied to Bayesian standard (s1) and random-effects (s2,s3) logistic models (i.ii). Bayesian probabilities of satisfying three meaningful thresholds of the risk of developing a clinical significant inhibitor (10/100, 5/100 [high rates] and 1/86 [FDA mandated cut-off rate in PTPs])(iii) were estimated. The effect of scaling up the study data size by 2 and 10 times was evaluated (iv).

Results: Results based on non-informative priors were similar to the classical approach. Using priors from PTPs lowered the point estimate and narrowed the credible intervals

(s1: from 1.3 [0.5, 2.7] to 0.8 [0.5, 1.1]; s2: from 1.9 [0.6, 6.0] to 0.8 [0.5, 1.1]; s3: 2.3 [0.5, 6.8] to 0.7 [0.5, 1.1]). All probabilities of satisfying a threshold of 1/86 were above 0.65. Increasing the number of patients by 2 and 10 times substantially narrowed the credible intervals for the single cohort study (1.4 [0.7, 2.3] and 1.4 [1.1, 1.8], respectively). Increasing the number of studies by 2 and 10 times for the multiple-studies scenarios (s2: 1.9 [0.6, 4.0] and 1.9 [1.5, 2.6]; s3: 2.4 [0.9, 5.0] and 2.6 [1.9, 3.5], respectively) had a similar effect.

Conclusion: Bayesian approach as a robust, transparent and reproducible analytic method can be efficiently used to answer the complex clinical questions.

Background

Developing inhibitors against factor VIII concentrates is the most severe and costly complication of the treatment of hemophilia A.[1] Patients who develop inhibitors have more episodes of bleeding and require larger doses of factor replacement to achieve hemostasis.

There are several reasons that complicate studying the determinants of inhibitor development.[2] The first is that the development of inhibitory antibodies is a combination of different events, more than a single one, with nothing as simple as black and white.[3–6] The second is the multifactorial nature of the phenomenon with both known and unknown risk factors and only some modifiable.[7,8] The third is the rarity of the disease, which hampers the opportunity to obtain substantial comparative data.[2]

In this challenging scenario, it is important to determine the risk associated with specific brands or classes of factor concentrates because the type of product is one of the few actionable risk factors in the field.[9,10] Other characteristics of the treatment regimen like dose, frequency, indication, and concomitant treatments or exposures also contribute to the risk of inhibitor development.[11–13]

Progress in this field requires a close collaboration with complementary expertise. Knowledge of immunology and basic science can help gain a broader and deeper understanding of the molecular and cellular mechanisms driving the development or breach of tolerance.[14–17] Clinical investigators can work to dissect the common characteristics among the heterogeneous clinical manifestations of inhibitory responses. Epidemiologists and biostatisticians can develop more powerful and efficient ways of looking at the available data and generating new ones.

There are several unmet needs in the statistical models used to analyze observational data about inhibitor development, which relate to the rarity of adverse events in an already rare disease.[18–24] The first critical issue is the scarcity of evidence, which emphasizes the need for incorporating external evidence to increase the power and the informative value of small and otherwise weak cohorts.[25] A second issue is the need for an efficient way to analyze the intricate relationship between treatment, time and the varying risk of events over time.[26] A third is the need to adjust for covariates (known risk factors) when performing multivariable exploration of, for example, inhibitor rates in previously untreated patients. The fourth and last, is the proper assessment and comparison of event

rates generated by non-parallel cohorts[27–29] In the present paper, we will address the first and fourth issues.

A powerful approach to the above problems might be a Bayesian framework. The Bayesian approach to interpreting experimental data from a clinical study consists of modeling the logical process leading to a change in opinion from before to after the availability of new information (the evidence provided by a new observation).

Here is a simple example of the Bayesian approach. Assume it is 7:00 AM a day in March. You look out of the windows and you see overcast. What is your estimate of the chance of snow? With nothing more than that, you would probably say that the chance of snow that day is 25%, chance of rain 25%, the chance of clearing up 25%, and the chance of staying the same 25%. Assume now you are in Toronto. Your estimate of the chance of snow that day would probably become 50%, chance of rain about 1%, the chance of cleaning up 20%, and chance of staying the same about 30%. If instead you were in Hong Kong, you might estimate the chance of snow at 0%, chance of rain at 80%, etc. The same exact information gets a different interpretation based on your previous knowledge about the city and it would be improper not to take it into account. To add another level of

complexity, you might imagine you will get different estimates if it is your first day in Toronto or in Hong Kong, or instead if you are familiar with the area. In fact, conditional on your previous knowledge, you will have a different level of confidence in your forecast estimates. In clinical practice, it is very useful and ethical to express some degree of credibility or confidence in your forecast with a patient. The power of the Bayesian approach is in formalizing and making transparent the way you define your previous knowledge, translate it into a technical language, and incorporate the new information. This process then provides a way to express the credibility of your forecast that is analogous to the classical measure of confidence in the result.

The power of the Bayesian approach derives from the opportunity of making use of -existing knowledge in the assessment of data that extends to either incorporating that knowledge in the final results or using it as a standard for comparison. That knowledge could be a similar measure in a similar unrelated trial, or a threshold of clinical importance.

Here is another example. It is still 7:00 AM, in March, overcast, in Toronto. The weather forecast is for a 50% chance of heavy rain, and you are ready to go for a walk. You don't

mind walking under a light rain, and you are a risk taker, but you always carry your coat and umbrella when the chance of heavy rain is over 80%. Based on the observation of historical trends, the average chance of heavy rain is not higher than 20% that day in Toronto. That said, the horizon looks unusually dark and the US east coast has been recently hit by the most powerful tornado of the last century. Now you ask yourself: can the chance of heavy rain be higher than 80% today? Sure it can. After putting all pieces of information together, you update your own estimates of the chance of heavy rain for today to 80% (Figure 1). This is enough for you to consider taking your coat and umbrella.

This paper demonstrates how the Bayesian approach works in comparison with a classical (frequentist) approach; how it can incorporate external evidence in the analysis of a single cohort of patients and a pooled analysis of a set of studies; whether it can help increase credibility of the results and the understanding of the underlying mechanisms; and how can we generate probabilities to be used in a physician-to-patient interaction. Recently published data [30–32] were used to work out three examples.

Methods

Overall study design: This paper is built around three case studies, and uses a standardized multistep approach to show i) how the Bayesian results compare with those based on the commonly used classical approach; ii) the impact of different sources of external information used to construct a Bayesian prior; iii) the use of different sources of external information as thresholds against which to benchmark Bayesian posterior estimates of risk; and iv) the impact of the size of information on the Bayesian posterior estimates. In the material and methods section we will describe the three case studies, the statistical details, and the data source we used for the simulation.

Case Study scenarios:

Case one: analyzing a rare adverse event in a single cohort (inhibitor rate in a cohort of previously treated patients, PTPs). The first example was set to represent the analysis of a single study where all patients were treated with the same FVIII product, aiming to assess the rate of inhibitor development in cohort of Hemophilia A PTPs. For this example, we re-analyzed the same cohort already published by Oldenburg and colleagues.[30] We used this example to explore and discuss the basics of the Bayesian approach and the pros and cons of choosing different priors.

Case two: analyzing a rare adverse event by pooling a set of studies in a meta-analysis. The second example was conceived to represent a meta-analysis of studies assessing the rate of inhibitors in a set of independent but similar studies in comparable populations of Hemophilia A patients; for this example we used a previous paper we published.[31] The main goal of this example -was to show how the Bayesian approach can -be a natural framework for a meta-analytical process.

Case three: analyzing the inhibitor rate in a previously unexplored setting. The third example illustrates the Bayesian analysis, interpretation and reporting of a small cohort study exploring a new clinical setting for which no obvious priors are available in the literature. We chose as a working example a previous report, studying the rate of inhibitor development in patients with low titer inhibitor at baseline or positive personal history of inhibitors. Although the study design and data collection were similar to the multi-national studies described in the previous case, the patient population was definitely different and not overlapping, thus adding to the complexity of data not directly comparable to any existing.[32] Another challenge presented in this example was pooling extremely sparse data in a multi-center study where no outcomes were observed in some centers (so-called zero-event).

By using the three study examples, the steps of conducting Bayesian analyses, and the potential benefits of using the Bayesian approach are showed step-by-step hereafter.

The Concepts behind Bayesian Inference:

Statistical inference is the process of fitting a probability model to a set of observed samples from a population to summarize the results by a probability distribution on the parameters of interests to make a general statement -about the population and predictions for new observations. In the classical (frequentist) approach, the statistical modeling only involves fitting a probability distribution to the observed experimental data to model the likelihood of the observed experimental data for a given estimate of interest such as treatment effect, incidence rate and *etc*. Unlike the classical approach, the Bayesian approach combines experimental and prior or external information via the Bayes theorem, to produce the posterior distribution which is used to make all inferences about the estimate of interest.

$$p(\delta|\text{data}) = p(\text{data}|\delta) \times p(\delta), \quad \textit{where } \delta \textit{ is the parameter of interest}$$

Posterior distribution = Data Likelihood x Prior distribution

As shown in the textbox, $p(\delta)$ represents prior distribution of *the parameter of interest (*hereforth to be referred to as *"the parameter"),* which present prior or external information about the estimate of treatment effect, incidence rate and *etc.*; $p(\text{data}|\ \delta\ )$, the likelihood function, specifies the statistical model of the observed experimental data given *the parameter* and $p\ (\delta\ |\text{data})$ is the posterior distribution of *the parameter* — which is essentially a combination of the evidence provided by the observed experimental data and prior relevant data from clinical experience or past research evidence.[33,34] In many cases, the posterior distribution $p(\delta|\text{data})$ is intractable and therefore to make inferences about *the paramete ,* the Bayesian approach uses Monte Carlo Markov Chain (MCMC) to obtain samples from the posterior $p(\delta|\text{data})$.[35] MCMC is an iterative process, with each iteration yielding a realization or observation from the posterior distribution $p(\delta|\text{data})$. Typically, investigators will conduct a large number of iterations or simulations: 1,000, 10,000, or even more. These are used to inform posterior inferences about *the parameter.* For example, the posterior mean or median is used to estimate *the parameter,* while the 2.5^{th} and 97.5^{th} observations are used as the 95% credible interval for *the parameter.* To calculate the probability that the estimate of *the parameter*$< K,$ where K is some threshold is given by the proportion of observations less than K. Table 1 provides a brief summary of the comparison main features of the frequentist and Bayesian approaches in clinical trials.

How can our questions be framed using the abovementioned Bayesian framework? We take our first scenario as an example. In this example, we are interested in estimating the inhibitor rate from the collected data. The likelihood $p(\text{data}|\ \delta)$ in our Bayesian model has a binomial distribution

$$\text{data}|\ \delta\ \sim \text{Binomial}\ (\delta,\ n)$$

where δ represents the inhibitor rate, and n is the total number of patients on some underlying treatment. The prior, $p(\delta)$, could be the inhibitor rate reported in an external study. After we run our Bayesian model to combine the information on the inhibitor rate contained in our data and the knowledge on the inhibitor rate found in the external study, we will have an updated estimate of the inhibitor rate that is represented through the posterior distribution.

How probabilities are used in the frequentist and the Bayesian approach is fundamentally different. For instance, let's say that we have a clinically meaningful reason to consider as sufficiently low an inhibitor rate less than 10%; thus, we want to test - if "the inhibitor rate (in our population) is less than 10%". In fact the frequentist frames test whether the null hypothesis (known) hold, i.e. "the inhibitor rate is greater or equal to 10%"at an arbitrary acceptable probability p that the null hypothesis may be wrongly rejected, say

0.05 [36]. In our example, if the "probability" of "the inhibitor rate is greater or equal to 10%" *is less* than 0.05, we conclude that this hypothesis can be rejected. However, this probability is not in fact a probability directly related to the acceptance of the testing hypothesis, but a level of credibility that, given the rate of inhibitor in our sample, and given the frequentist theoretical construct based on the normal distribution of the means of the infinite possible samples of the theoretical population, we can refuse the null hypothesis. In fact, when p =< 0.05, we will reject the null hypothesis, but we are never able to say that the probability of "the inhibitor rate being less than 10%" is truly 0.95. On the other hand, the Bayesian probability is a quantity of the testing hypothesis. The Bayesian can really test the probability that the rate of inhibitor in our sample is less than 10%. If p = 0.95, we are confident that the probability of "the inhibitor rate being less than 10%" is actually 0.95. The estimates associated with the probability are confidence interval (CI) in classical approach and credible interval (CrI) in Bayesian approach. Back to our example, the 95% CI is interpreted as "the estimates of the inhibitor rate will fall in between these two boundaries 95% of the time if the data can be repeated infinitely". It cannot be used to make an assertion about the current test based on a single sample set without the assumption of the infinite repetition. In comparison, the 95% CrI tells us a straight forward story, "given the data and the model, the chance of the true inhibitor rate fall in this interval is 95%".

Setting basic models with non-informative priors: For all three cases, Bayesian statistical models with non-informative priors were introduced first as the basic starting model. Non-informative priors are vague priors that carry relatively minimal information; consequently, the posterior estimates are derived predominantly from the study data, and directly comparable to the results obtained through the classical frequentist approach. For the first example (one single cohort of patients) classical logistic and Bayesian logistic models were used. For the second and third examples (multiple studies setting) classical random-effects logistic and Bayesian hierarchical random-effects logistic models were adopted, through which the patients from the same cohort were clustered. The random-effects model was adopted as the commonest choice for individual patient data meta-analyses.

Choosing informative priors: As a second step, we replaced non-informative priors with information-rich priors to incorporate the pre-existing external information/knowledge from previous studies into the analysis of the current study data. Unlike non-informative priors, informative ones contribute information to the posterior estimates, which can be looked at as a "combination" of the pre-existing evidence with -evidence generated by the current experiment. . In fact, posterior estimates are weighted averages based on prior and current experimental evidence/data, with the weights determined by the precision of the

corresponding evidence. For examples one and two, the main goal of using prior was to incorporate the existing evidence and increase the comprehensiveness of the conclusion. To this scope, we sought relevant comparable priors, and tested two different sets of informative priors. The first set (a) comprised data obtained during the treatment with a certain molecule (e.g. rAHF-PFM) in different studies; specifically a-i) estimates of inhibitor rates from the manufacturer pivotal studies;[37] a-ii) estimates of inhibitor rates from a meta-analysis [38] and a-iii) estimates of inhibitor rates from an independent prospective multicentric cohort.[39] The second set of informative priors (b) comprised pooled inhibitor rates for any FVIII concentrate, including: b-i) a meta-analysis [38] and b-ii) an independent prospective multicenter cohort.[39]

The third study example was specifically chosen not to have a study on the same patient population already available; thus, no obvious informative priors can be located in the literature. Notwithstanding, we wanted to show the value of the Bayesian approach in exploring how the rate of inhibitor development in this population would change when the known rate in previously treated patients (PTP) and that in previously untreated patients (PUPs) are added in. Consequently, in addition to the informative priors used in the first two examples, we also added the inhibitor rates for PUPs reported in the EUHASS study [39] for a) the specific molecule and for b) all products. The key rationale

was to assess the robustness or sensitivity of the posterior inhibitor rates see when a prior based on a truly high-risk population is used. The details of generating priors can be found in Appendix A.

Calculating probabilities: Unlike the frequentist probability model, which tests whether the null hypothesis can be rejected successfully, the Bayesian probability approach generates a quantitative estimation of the "degree of truth" of the study hypothesis. Another interesting characteristic of the posterior probability is its nature of conditional probability, which lends it to be continually updated upon the availability of new data. The most informative example of using the conditional probability framework is where the inhibitor rate among the patients with low titer inhibitor at baseline or personal history of inhibitor had been given little consideration to date. Therefore, comparing the posterior estimates of the inhibitor rate obtained from the study data commonly used as clinical thresholds will provide clinicians with meaningful ways to interpret the results. To make this more evident, and show another peculiar property of the Bayesian framework, we further calculated for the third example the Bayesian probabilities of the posterior inhibitor rates being lower than three specific clinically meaningful thresholds, two high rates (10/100 and 5/100) and the FDA mandated cut-off rate in PTPs (1/86).[40,41]

Testing more complex hypotheses: We moved then to show the effect of scaling down or discounting the value of the prior information. We used the third example, for which due to the inability to -use -a full consistent informative prior one might want to assign less weight to the - information carried by the selected - priors. The weight of the prior could be reduced in at least 2 ways:

1)by decreasing the precision i.e. enlarging the variance of the priors depending on how relevant the particular piece of information is to the study we are assessing [34,42]. In our third example, we will discount the precision of the rates of inhibitor in PUPs in EUHASS for the specific molecule and for any factor VIII concentrate by 75% and 95% each, respectively. This equals to the human process of any perceived information: you told me that the rate of event is this, but I only 25% trust your information. 2) a second approach to obtain the same objective, i.e. to undervalue the contribution of the priors, is scaling up the weight of study data by increasing the precision assigned to the experimental data. One easy and understandable way to do this is to simulate the impact on posterior estimates of increasing the study sample size. Thus, we showed the effect of increasing the study data sample size by 2 times and 10 times respectively, for all three studies examples when using pooled estimates as priors. The increment of sample was done in two ways: i) increasing the number of events and number of patients in each center proportionally; ii) increasing the number of centers and keeping the number of

patients in each center the same. We re-ran the Bayesian model with non-informative priors using new inflated data for all three examples, and we further re-analyzed the effect of data inflation in the third example for all informative priors previously used.

Analysis and reporting: Throughout this study, posterior inhibitor rates (our results) were reported as percentage rates with 95% associated confidence interval (CI) in the case of classic statistics, or 95% credibility interval (CrI) in the case of Bayesian statistics. Graphic, descriptive statistics and classical meta-analyses were performed using Stata 13.1 (Statacorp, College Station, Tx, US). Bayesian analyses were performed using WinBUGS software 1.4.3 (http://www.mrc-bsu.cam.ac.uk/bugs/). In every Bayesian analysis, two chains were run simultaneously and the convergence of the Bayesian models was assessed based on the history trace, posterior density and auto-correlation plots for parameters of interest. The codes of Bayesian models detailed setups on the Bayesian simulations can be found in Appendix B.

Source data used in the present paper: The individual data sets used to build our examples for illustration purposes [30–32] were from the ADVATE PASS (Post-Authorization Safety Studies) program. The study population in PASS studies was

severe-moderate Hemophilia A patents undergoing treatment (prophylaxis or on-demand) with ADVATE in routine clinical use. The primary safety outcome in these studies is defined as measurable inhibitors levels during the study period, including *de novo*, recurrent and persistent inhibitors. We adopted the cut-offs specified in the original PASS protocols: 1.0 Bethesda Unit (BU) for USA- EU- and Australia-PASS; and 0.6 BU for Japan-, Italy-, Korea and Taiwan-PASS (studies adopting the Nijmegen modification).[30,43]

Results

Description of the three datasets used for the examples. For Example 1, six inhibitors were reported among 428 patients (all severity patients, de-novo and recurrent, in 4 PTPs and 2 PUPs). For Example 2, five cohorts were included the IPD meta-analysis, and 21 inhibitors were reported in 1188 patients. For Example 3, individual patient data were extracted from seven PASS studies and 6 inhibitors were reported in 219 patients.

Comparing results from Bayesian approach to classical approach

As expected, the results obtained from classical analytical approach and Bayesian statistical model using non-informative priors were similar for all the three examples. For the single cohort study (example one), the estimates were the same to one decimal place

(percent rate (95% CI): 1.3% (0.5, 2.7); percent rate (95% CrI): 1.3% (0.5, 2.7)). For the pooled analysis (example two), the Bayesian posterior estimates gave a slightly wider 95% CrI (1.9% (0.5, 4.5)) toward the lower end as compared to the estimate from the classical approach (1.9% (0.8, 4.5)). For the cohort of patients with previous/current inhibitor (example three), the posterior estimates from the Bayesian model showed a slightly lower percent rate and wider 95% CrI (2.3% (0.5, 6.8)) as compared to the estimate from classical approach (2.6% (1.0, 6.8)). (Table 2, Figure2a, 2b, 2c).

Impact of using informative priors in Bayesian analysis

For example one and two, using external information as priors consistently narrowed the credible intervals and lowered the central estimate of percentage rates. The range of the inhibitor percentage rates for the single cohort (example one) was 0.8% to 1.3% and for the pooled analysis was 0.8% to 1.9%. For the cohort of patients with previous/current inhibitor (example three), the posterior estimates of inhibitor percentage rates changed depending on the external information brought in by priors. The lowest percentage rates with narrowest 95% CrI was obtained using the informative prior generated from EUHASS study of PTP patients for all FVIII products: 0.7% (0.5, 1.1). The highest percentage rate was gained using the informative priors generated from EUHASS study of PUP patients for all FVIII products: 24.9% (21.1, 29.2). (Table 2, Figure2a, 2b, 2c).

Comparing posterior inhibitor rates to thresholds

For the cohort of patients with previous/current inhibitor (example three), the posterior estimates of inhibitor rates were compared to the selected thresholds and the probabilities of posterior inhibitor rates lower than the thresholds were then calculated. A probability of 1 means that the calculated rate of inhibitors is certainly below the threshold, a probability of 0.5 would mean a 50% likelihood that the rate is below the threshold. Testing a threshold for the rate of inhibitors of 10%, six out of eight comparisons (when non-informative prior and informative priors were generated from the studies of PTP patients) showed a probability greater than 0.99. In contrast, when EUHASS PUPs study results were used as priors, the probabilities of a rate lower than 10% dropped dramatically to less than 0.001. Similar findings were obtained when the threshold was dropped to lower than 5%. When the threshold was dropped further to the FDA approved rate for PTPs of 1/86, only the probability using priors for EUHASS study in PTPs for all products was greater than 0.9, but all estimations using PTPs as priors were still above 0.65. (Table 3)

Impact of sample size of data

We found that increasing the number of patients narrowed the credible interval for example one (i.e. the mimic of single center studies), but had little impact on the posterior estimates of example two and three, which represented multi-center study settings. However, when more centers were added to get to the same sample size, the credible intervals noticeably narrowed. Another interesting observation is that when the number of patients was increased in each center while the number of centers remained the same, the posterior inhibitor rates decreased for example three, in which three of seven centers reported no inhibitor event in the original data Table 2. (More exploratory results can be found in Appendix C.)

Discussion

In this paper we used three real-world examples to guide the reader to appreciate the power of the Bayesian approach to analyze and interpret rare events observed in a rare population. Initially using vague priors (not having or ignoring prior knowledge), we showed how the Bayesian estimation process generates point estimates very similar to the frequentist approach. We demonstrate how the iterative process typical of the Bayesian estimation can be used to generate credible intervals around the point estimate, which are

the range of possible values of the estimate. We show how the credibility interval has a precise probabilistic distribution of discrete values, which can be used to assess whether the likelihood of the event is above or below a given value. We discuss how this is of much greater interest for the clinician and the researcher than the distribution of the point estimate in hypothetical repetitions of the experiment, which is what the confidence intervals represent. Subsequently, we moved to show how the point-estimate and the credibility interval change when we consider a specific set of experimental data in light of what we already know of a given or similar phenomenon.

Some further considerations are hopefully of value. For example one and two, the Bayesian models with non-informative priors yielded results comparable to the classical approach. For third study case, the point estimate of inhibitor rate obtained from the Bayesian random-effects logistic model was lower than that obtained from the classical random-effects logistic model. The reason is that the data used for this example are extremely sparse. In three out of seven pooled studies, there were no inhibitors observed. The classical logistic model directly takes event as outcome and thus fails to generate the estimates when no event is in the data. Therefore, when classical random-effect logistic was used to pool the data from seven individual studies, the three studies without outcomes were ignored, and the inhibitor rate was estimated from the four studies with

observed inhibitor. Unlike the classical model, the Bayesian model re-samples data for certain times (say 100,000) based on the information provided by the current data and then generate the estimates in accordance before reporting the posterior estimates which usual is the median of the entire estimates. In our example, when data reported no event, the Bayesian model resampled data using the probability of event sampled around zero. By doing so, the Bayesian model was able to incorporate those studies reporting no inhibitor into the posterior estimates and thus gave a lower inhibitor rate. On the other hand, for the same example, the 95 CrIs were wider than the 95% CI. This is because the Bayesian model introduced more random uncertainty through non-informative priors -that had very large variance. When the study data are not large enough, random uncertainty will be added in the posterior estimates. In our example, we had seven studies that were not even able to provide saturated information for estimating the between study variance. Therefore, the model borrowed information from non-informative priors that only added uncertainty to estimate the between study variance.

In example three, we show how our set of experimental data can or cannot change our previous belief. We modeled the effect of observing six inhibitors in about 200 patients from the unlikely expectation that the inhibitor rate would have been as in RODIN, to the optimistic expectation that the rate would have not been different from that in PTPs. We

also showed how we could model the "strength" of this belief, by "discounting" the previous information. Critics of the Bayesian approach would certainly say- that by adding "discounts" you may play with data until you show what you want. We would object that this would be the case if you were only using one set of priors (maybe even discounted). If you instead show the results produced by using a whole range of priors, you explore the relevance of your previous beliefs and assumptions. Along the same lines, we showed how Bayesian modeling can be used to simulate the effect of repeating the study or doubling the population, either by increasing the number of patients in the same centers or by increasing the number of centers. All of this richness of information is completely unavailable when using the frequentist approach. The reader needs to be aware, at this point, that most of the modeling of the impact of health care interventions on economics of health care systems or quality of life of patient population is generally obtained via Monte Carlo chain simulations which are, in essence, Bayesian probability applications [34,44].

To come back to the clinical ground, we showed how the Bayesian posterior distribution can be interrogated to get, for example, the posterior probability that the rate of inhibitors in a population like the one we studied (e.g. patients with previous history of inhibitors) was above or below a given (clinically meaningful) threshold. This is what, in our

opinion, is needed for clinical decisions, and, indirectly, for policy making decisions like taking into account the 30% of patients with a previous history of inhibitors in the proportion of population to be suitable to switch concentrate as a result of a tender process.

Although the interpretation of probability is clinically intuitive, we are painfully aware that the wisdom of probability is a difficult concept to grasp. However, we would like to think that most of the difficulty is in the limited number of attempts made in the past to present the basics of the Bayesian approach in a practical and simplified manner. We made such an attempt, targeting practicing hematologists as our audience, by using three real-world examples in the field of hemophilia. We used real data to generate new evidence via a Bayesian simulation, and we added as much educational value as well. The Bayesian approach offers a great opportunity to move science forward in the rare disease field by maximizing the use of existing knowledge. If we guess about today's probability of rain blinded to where we are and when, we have a very high chance of getting soaked, or uselessly carrying our umbrella.

Conclusion

The Bayesian estimates of the inhibitor rate of patients undergoing treatment with ADVATE provide a broader understanding for the clinicians, which can be utilized to inform clinical decisions in management of patients with Hemophilia A. Bayesian approach as a robust, transparent and reproducible analytic method can be efficiently used to answer the complex clinical questions.

Competing interests: JC, MM, EMP and JKM have no conflict to declare. AI has worked as a consultant for, and held research contracts sponsored by Bayer, Biogen Idec, NovoNor- disk, Pfizer. VR is an employee of Baxter HealthCare. LT has worked as a consultant for, and held research contracts sponsored by many pharmaceutical companies including GlaxoSmithKline Inc, AstraZeneca, Sorono Canada Inc, F. Hoffman-La Roche Ltd, Pfizer, Theralase Inc, CanReg Inc, Merck Frosst - Schering Pharmaceuticals and Proctor and Gamble Pharmaceuticals Canada Inc.

Funding: This work was supported by a research contract agreement between Baxter Health-Care and St Joseph HealthCare Hamilton.

Contributors: JC, AI, MM, VR and LT designed the study and drafted the manuscript. JC conducted the statistical analysis. AI, MM, VR, JKM, provided the clinical expertise. JC, AI, MM, EMP, LT provided input on statistical concept. All authors revised the manuscript for important clinical and statistical contents and approved the final manuscript.

Acknowledgements: The authors would like to acknowledge Jessica Fowler (BAXTER Healthcare) for editorial services in reviewing the manuscript.

References

1. W Collins P, Chalmers E, Hart D, et al. Diagnosis and management of acquired coagulation inhibitors: A guideline from UKHCDO. *Br J Haematol*. 2013;162(July):758-773. doi:10.1111/bjh.12463.

2. Iorio A. Epidemiology of inhibitors in hemophilia. In: *Textbook of Hemophlia*.; 2014:53-58.

3. Caram C, de Souza RG, de Sousa JC, et al. The long-term course of factor VIII inhibitors in patients with congenital haemophilia A without immune tolerance induction. *Thromb Haemost*. 2011;105(1):59-65. doi:10.1160/TH10-04-0231.

4. Hay CRM, Palmer B, Chalmers E, et al. Incidence of factor VIII inhibitors throughout life in severe hemophilia A in the United Kingdom. *Blood*. 2011;117(23):6367-6370. doi:10.1182/blood-2010-09-308668.

5. Stieltjes N, Torchet MF, Misrahi L, et al. Epidemiological survey of haemophiliacs with inhibitors in France: orthopaedic status, quality of life and cost – the "Statut Orthopédique des Patients Hémophiles" avec Inhibiteur study. *Blood Coagul Fibrinolysis*. 2009;20(1):4-11. doi:10.1097/MBC.0b013e328313fc8e.

6. Gringeri A, Monzini M, Tagariello G, et al. Occurrence of inhibitors in previously untreated or minimally treated patients with haemophilia A after exposure to a plasma-derived solvent-detergent factor VIII concentrate. *Haemophilia*. 2006:128-132. doi:10.1111/j.1365-2516.2006.01201.x.

7. Astermark J, Altisent C, Batorova a, et al. Non-genetic risk factors and the development of inhibitors in haemophilia: a comprehensive review and consensus report. *Haemophilia*. 2010;16(5):747-766. doi:10.1111/j.1365-2516.2010.02231.x.

8. Coppola A, Santoro C, Tagliaferri A, Franchini M, DI Minno G. Understanding inhibitor development in haemophilia A: towards clinical prediction and prevention strategies. *Haemophilia*. 2010;16 Suppl 1:13-19. doi:10.1111/j.1365-2516.2009.02175.x.

9. Iorio A, Puccetti P, Makris M. Clotting factor concentrate switching and inhibitor development in hemophilia A. *Blood*. 2012;120(4):720-727. doi:10.1182/blood-2012-03-378927.

10. Lee CA, Lillicrap D, Astermark J. Inhibitor development in hemophiliacs: the roles of genetic versus environmental factors. *Semin Thromb Hemost*. 2006;32 Suppl 2:10-14. doi:10.1055/s-2006-946909.

11. Gouw SC, van den Berg HM, Fischer K, et al. Intensity of factor VIII treatment and inhibitor development in children with severe hemophilia A: the RODIN study. *Blood*. 2013;121(20):4046-4055. doi:10.1182/blood-2012-09-457036.

12. Gouw SC, van den Berg HM. The multifactorial etiology of inhibitor development in hemophilia: genetics and environment. *Semin Thromb Hemost*. 2009;35(8):723-734. doi:10.1055/s-0029-1245105.

13. Gouw SC, van der Bom JG, Marijke van den Berg H. Treatment-related risk factors of inhibitor development in previously untreated patients with hemophilia A: the CANAL cohort study. *Blood*. 2007;109(11):4648-4654. doi:10.1182/blood-2006-11-056291.

14. Pavlova A, Delev D, Lacroix-Desmazes S, et al. Impact of polymorphisms of the major histocompatibility complex class II, interleukin-10, tumor necrosis factor-alpha and cytotoxic T-lymphocyte antigen-4 genes on inhibitor development in severe hemophilia A. *J Thromb Haemost*. 2009;7(12):2006-2015. doi:10.1111/j.1538-7836.2009.03636.x.

15. Astermark J, Lacroix-Desmazes S, Reding MT. Inhibitor development. *Haemophilia*. 2008;14 Suppl 3:36-42. doi:10.1111/j.1365-2516.2008.01711.x.

16. Saint-Remy J-M, Reipert BM, Monroe DM. Models for assessing immunogenicity and efficacy of new therapeutics for the treatment of haemophilia. *Haemophilia*. 2012;18 Suppl 4:43-47. doi:10.1111/j.1365-2516.2012.02828.x.

17. Matino D, Lillicrap D, Astermark J, et al. Switching clotting factor concentrates: considerations in estimating the risk of immunogenicity. *Haemophilia*. 2014;20(2):200-206. doi:10.1111/hae.12283.

18. Kesselheim AS, Myers J a, Avorn J. Characteristics of clinical trials to support approval of orphan vs nonorphan drugs for cancer. *JAMA*. 2011;305(22):2320-2326. doi:10.1001/jama.2011.769.

19. Iorio A, Marcucci M. Clinical trials and haemophilia : does the Bayesian approach make the ideal and desirable good friends ? *Haemophilia*. 2009;15(4):900-903. doi:10.1111/j.1365-2516.2009.02031.x.

20. Chow S-C, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis*. 2008;3:11. doi:10.1186/1750-1172-3-11.

21. Honkanen VE, Siegel AF, Szalai JP, Berger V, Feldman BM, Siegel JN. A three-stage clinical trial design for rare disorders. *Stat Med*. 2001;20(20):3009-3021.

22. Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ*. 1995;311(7020):1621-1625.

23. Behera M, Kumar A, Soares HP, Sokol L, Djulbegovic B. Evidence-based medicine for rare diseases: implications for data interpretation and clinical trial design. *Cancer Control*. 2007;14(2):160-166.

24. Committee on Strategies for Small-Number-Participant Clinical Research Trials, Board on Health Sciences PolicyCommittee on Strategies for Small-Number-Participant Clinical Research Trials B on HSP, ed. *Small Clinical Trials: Issues and Challenges*. National Academy Press; 2001.

25. Iorio A, Marcucci M. Clinical trials and haemophilia : does the Bayesian approach make the ideal and desirable good friends ? *Haemophilia*. 2009;15(4):900-903. doi:10.1111/j.1365-2516.2009.02031.x.

26. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

27. Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363(9422):1728-1731. doi:10.1016/S0140-6736(04)16261-2.

28. Hornberger J, Wrone E. When to base clinical policies on observational versus randomized trial data. *Ann Intern Med*. 1997;127(8 Pt 2):697-703.

29. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Epidemiology*. 2007;18(6):800-804. doi:10.1097/EDE.0b013e3181577654.

30. Oldenburg J, Goudemand J, Valentino L, et al. Postauthorization safety surveillance of ADVATE [antihaemophilic factor (recombinant), plasma/albumin-free method] demonstrates efficacy, safety and low-risk for immunogenicity in routine clinical practice. *Haemophilia*. 2010;16(6):866-877. doi:10.1111/j.1365-2516.2010.02332.x.

31. Marcucci M, Cheng J, Oldenburg J, et al. Meta-analysis of Post Authorization Safety Studies: worldwide postmarking surveillance of hemophilia A patients treated with antihemophilic factor recombinant plasma/albumin-free method

rAHF-PFM. *J Thromb Haemost*. 2013;11(Suppl 2):1075. http://onlinelibrary.wiley.com/doi/10.1111/jth.12443/pdf.

32. Romanov V, Marcucci M, Cheng J, Thabane L, Iorio A. Evaluation of Safety and Effectiveness of factor VIII treatment in Hemophilia A patients with low titer inhibitors or a personal history of inhibitor. *Thromb Haemost*. 2015;113(3):Inpress.

33. Kalil AC, Sun J. Bayesian Methodology for the Design and Interpretation of Clinical Trials in Critical Care Medicine. *Crit Care Med*. 2014;42:2267-2277. doi:10.1097/CCM.0000000000000576.

34. Spiegelhater DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. 1st ed. Chichester, West Sussex, England: John Wiley & Sons Ltd.; 2004:391.

35. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput*. 2000;10:325-337. doi:10.1023/A:1008929526011.

36. Greenland S, Poole C. Living with P Values. *Epidemiology*. 2013;24(1):62-68. doi:10.1097/EDE.0b013e3182785741.

37. Tarantino MD, Collins PW, Hay CRM, et al. Clinical evaluation of an advanced category antihaemophilic factor prepared using a plasma/albumin-free method: pharmacokinetics, efficacy, and safety in previously treated patients with haemophilia A. *Haemophilia*. 2004;10(5):428-437. doi:10.1111/j.1365-2516.2004.00932.x.

38. Xi M, Makris M, Marcucci M, Santagostino E, Mannucci PM, Iorio A. Inhibitor development in previously treated hemophilia A patients: a systematic review, meta-analysis, and meta-regression. *J Thromb Haemost*. 2013;11(9):1655-1662. doi:10.1111/jth.12335.

39. Fischer K, Lassila R, Peyvandi F, et al. Inhibitor development in haemophilia according to concentrate Four-year results from the European HAemophilia Safety Surveillance (EUHASS) project. *J Thromb Haemost*. 2015;113(5):968-975. doi:10.1160/TH14-10-0826

40. Aledort LM. Harmonization of clinical trial guidelines for assessing the risk of inhibitor development in hemophilia A treatment. *J Thromb Haemost*. 2011;9(3):423-427. doi:10.1111/j.1538-7836.2010.04169.x.

41. Department Of Health And Human Services Food And Drug Administration Center For Biologics Evaluation And Research. Workshop On Factor VIII Inhibitors. Lister Hill Center National Institutes of Health. 2003;2802(202).

42. Salpeter SR, Cheng J, Thabane L, Buckley NS, Salpeter EE. Bayesian Meta-analysis of Hormone Therapy and Mortality in Younger Postmenopausal Women. *Am J Med*. 2009;122(11):1016-1022.e1. doi:10.1016/j.amjmed.2009.05.021.

43. Iorio A, Marcucci M, Cheng J, et al. Patient data meta-analysis of Post-Authorization Safety Surveillance (PASS) studies of haemophilia A patients treated with rAHF-PFM. *Haemophilia*. 2014;20(6):777-783. doi:10.1111/hae.12480.

44. Stimson RH. An Introduction to Bayesian Inference and Decision. *J Oper Res Soc*. 1974;25(2):336-337. doi:10.1057/jors.1974.62.

Table 1: Brief comparison of the frequentist and Bayesian approaches in clinical trials

(adopted and modified from several sources [33,34,44])

Feature		**frequentist Approach**	**Bayesian Approach**
Interpretation of probability		The proportion of times an event will occur in an infinitely long series of repeated identical situations	The "degree of belief" of an event (or a number of repeatable events) will occur
Main question		What is the probability of data (trial result), given the hypothesis (treatment effect)?	What is the probability of the hypothesis (treatment effect), given the data (trial result)?
Design features		Hypotheses, type I and II errors	Hypotheses, Prior or external information
Reasoning paradigm		Deductive reasoning	Inductive reasoning
Trial monitoring		Pre-specified with adjustments for type I error for interim analyses	Adaptive by design based on accumulating evidence
Condition of drawing statistical inference		Inference based observed experimental data	Inference based on observed experimental data and prior information
Information for Analysis	Use of external information/pre-belief	Informally considered only at study design stage, e.g. sample size calculation	Formally incorporated in the design, analysis and interpretation as a prior
	Experimental data	Summarized via the likelihood function, which captures all information provided by data regarding any unknown population parameters	

Results summaries	Point estimate	The "best estimate" obtained from observed experimental data	An “weighted point estimate from the posterior distribution derived by combining all relevant sources of information including the external information and observed experimental data
	Interval estimates	95% confidence interval (CI)—an interval that we are 95% confident that the true value of the unknown parameter would be as low as its lower bound and as high as its upper bound	95% credible interval (CrI)—an interval in which the unknown parameter would lie with probability 0.95 given the observed experimental data
	Probabilities	P-value, the chance of observing a result as extreme as what is seen in the experiment when the null hypothesis of no effect is true	Posterior probabilities
Decision-making	Frame-work	Not straightforward and hard to apply in clinical practice	Intuitive and based on minimizing expected losses; easy to apply in clinical practice

Table 2: Inhibitor rates for three different examples

	Example 1	Example 2	Example 3
Method	Single study	Meta-analysis	Multicenter cohort– no appropriate priors
Test data (number of inhibitors/number of patients)	PASS data[30] (6/428)	PASS data[31] (21/1188)	PASS data[32] (6/219)
Classical Statistical Analysis: percent rate (95% CI)	1.3 (0.5, 2.7)	1.9 (0.8, 4.5)	2.6 (1.0, 6.8)
Bayesian Statistical Analysis: percent rate (95% CrI)			
Non-informative prior	1.3 (0.5, 2.7)	1.9 (0.6, 6.0)	2.3 (0.5, 6.8)
Informative prior: Baxter Pivot Study (1/102)	1.3 (0.5, 2.5)	1.6 (0.6, 4.1)	1.8 (0.5, 4.8)
Informative prior: meta-analysis of OS (7 ADVATE studies) (3/569)	0.9 (0.4, 1.9)	1.0 (0.4, 2.2)	0.9 (0.3, 2.3)
Informative prior: meta-analysis of OS (38/3866)	1.0 (0.8, 1.4)	1.0 (0.8, 1.4)	1.0 (0.8, 1.4)
Informative prior: EUHASS study *de novo* inhibitor PUPs ADVATE (37/141)	NO	NO	23.4 (17.5, 30.7)
Informative prior: EUHASS study *de novo* inhibitor PUPs (108/417)	NO	NO	24.9 (21.1, 29.2)
Informative prior: EUHASS study inhibitors in PTPs ADVATE (5/707)	1.0 (0.5, 1.8)	1.1 (0.5, 2.1)	1.0 (0.4, 2.1)
Informative prior: EUHASS study inhibitors in PTPs (all FVIII) (26/3736)	0.8 (0.5, 1.1)	0.8 (0.5, 1.1)	0.7 (0.5, 1.1)
Discounted prior: Discounting EUHASS PUPs ADVATE by 75%	NO	NO	16.9 (9.0, 29.4)

Discounted prior: Discounting EUHASS PUPs ADVATE by 95%	NO	NO	5.3 (2.2, 16.0)
Discounted prior: Discounting EUHASS PUPs ALL by 75%	NO	NO	22.2 (15.7, 30.4)
Discounted prior: Discounting EUHASS PUPs ALL by 95%	NO	NO	12.3 (5.4, 25.8)
Enhanced data : Enhancing study data by 2 times - increasing number of patients (with non-informative prior)	1.4 (0.7, 2.3)	2.0 (0.6, 6.4)	2.2 (0.5, 6.6)
Enhanced data: Enhancing study data by 2 times - increasing number of studies (with non-informative prior)	NO	1.9 (0.9, 4.0)	2.4 (0.9, 5.0)
Enhanced data: Enhancing study data by 10 times - increasing number of patients (with non-informative prior)	1.4 (1.1, 1.8)	2.1 (0.6, 6.6)	1.6 (0.4, 5.4)
Enhanced data: Enhancing study data by 10 times - increasing number of studies (with non-informative prior)	NO	1.9 (1.5, 2.6)	2.6 (1.9, 3.5)

PASS: post-authorization safety studies

OS: observational study **CI**: confidence interval **CrI**: credible interval

PUP: Previously untreated patient **PTP**: Previously treated patient

EUHASS: European Haemophilia Safety Surveillance

Table 3: Probabilities for the inhibitor rate from PASS [32] to be lower than pre-specified thresholds

	Example 3: PASS	Threshold 1	Threshold 2	Threshold 3
Bayesian Statistical Analysis: percent rate (95% CrI)	Multicenter study – no appropriate priors	<10/100	<5/100	<1/86
Non-informative prior	2.3 (0.5, 6.8)	0.994	0.921	0.165
Informative prior: Baxter Pivot Study (1/102)	1.8 (0.5, 4.8)	>0.999	0.979	0.225
Informative prior: meta-analysis of OS (7 ADVATE studies) (3/569)	0.9 (0.3, 2.3)	>0.999	>0.999	0.677
Informative prior: meta-analysis of OS (38/3866)	1.0 (0.8, 1.4)	>0.999	>0.999	0.782
Informative prior: EUHASS study *de novo* inhibitor PUPs ADVATE (37/141)	23.4 (17.5, 30.7)	<0.001	<0.001	<0.001
Informative prior: EUHASS study *de novo* inhibitor PUPs (108/417)	24.9 (21.1, 29.2)	<0.001	<0.001	<0.001
Informative prior: EUHASS study inhibitors in PTPs ADVATE (5/707)	1.0 (0.4, 2.1)	>0.999	>0.999	0.658
Informative prior: EUHASS study inhibitors in PTPs (all FVIII) (26/3736)	0.7 (0.5, 1.1)	>0.999	>0.999	0.988
Discounted prior: Discounting EUHASS PUPs ADVATE by 75%	16.9 (9.0, 29.4)	0.051	<0.001	<0.001
Discounted prior: Discounting EUHASS PUPs ADVATE by 95%	5.3 (2.2, 16.0)	0.876	0.449	0.001
Discounted prior: Discounting EUHASS PUPs ALL by	22.2 (15.7, 30.4)	<0.001	<0.001	<0.001

75%				
Discounted prior: Discounting EUHASS PUPs ALL by 95%	12.3 (5.4, 25.8)	0.306	0.016	<0.001
Enhanced data : Enhancing study data by 2 times - increasing number of patients (with non-informative prior)	2.2 (0.5, 6.6)	0.995	0.932	0.161
Enhanced data: Enhancing study data by 2 times - increasing number of studies (with non-informative prior)	2.4 (0.9, 5.0)	0.998	0.967	0.305
Enhanced data: Enhancing study data by 10 times - increasing number of patients (with non-informative prior)	1.6 (0.4, 5.4)	0.998	0.976	0.067
Enhanced data: Enhancing study data by 10 times - increasing number of studies (with non-informative prior)	2.6 (1.9, 3.5)	>0.999	>0.999	<0.001

PASS: post-authorization safety studies

OS: observational study **CI**: confidence interval **CrI**: credible interval

PUP: Previously untreated patient **PTP**: Previously treated patient

EUHASS: European Haemophilia Safety Surveillance

Figure1: Bayesian concept graphic illustration

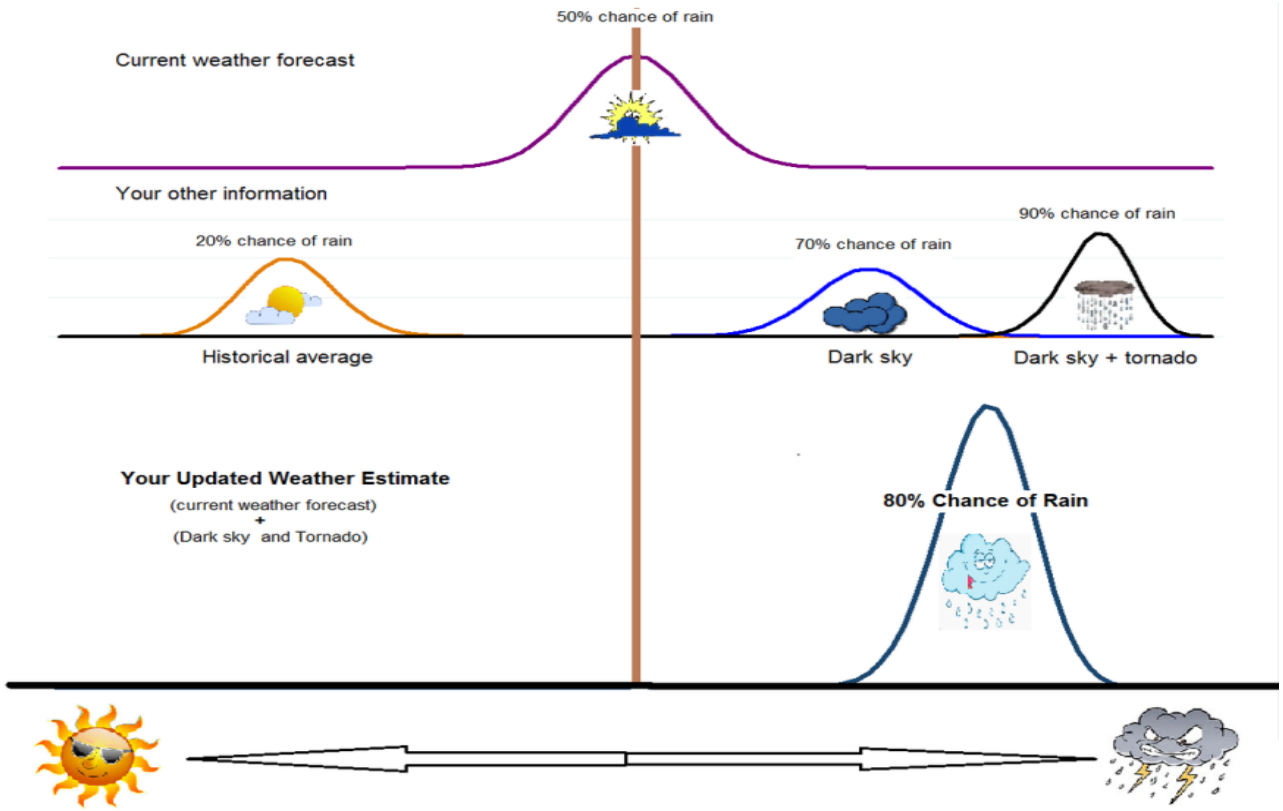


Figure2a: Example 1: Single study

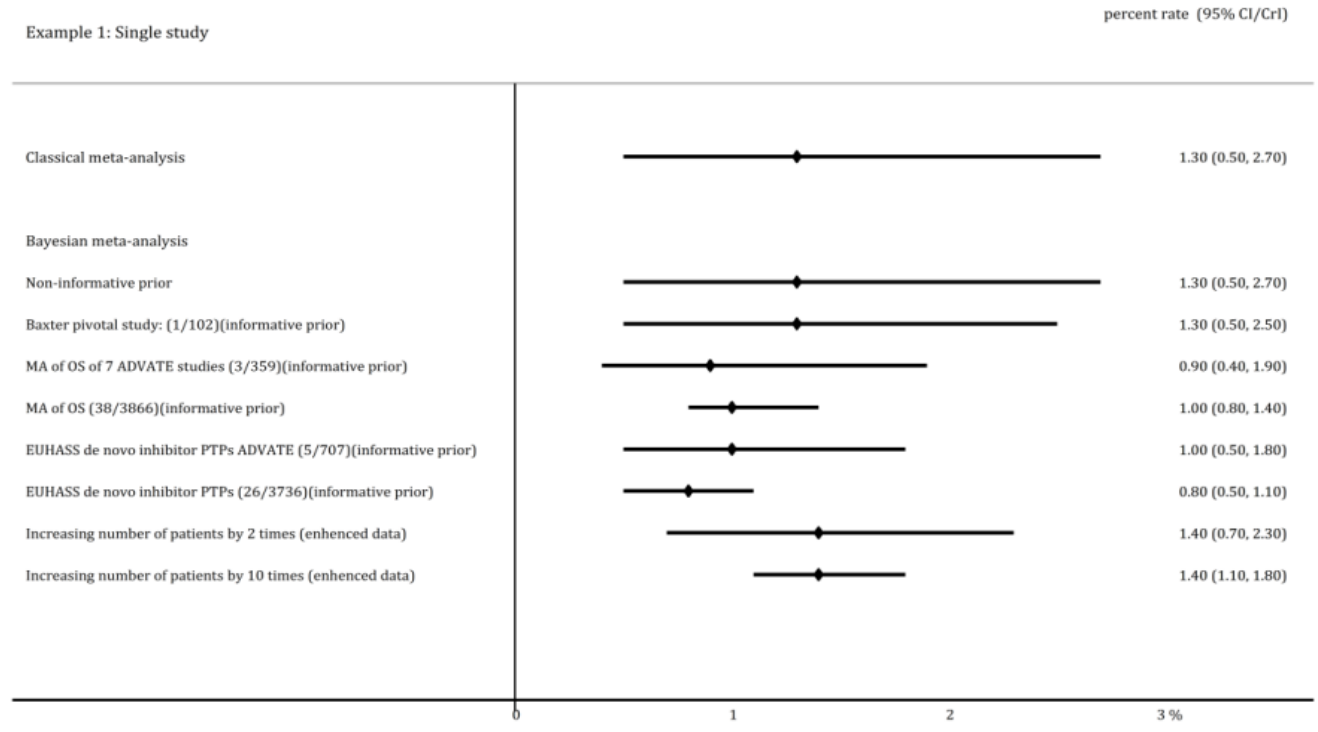


Figure2b: Example 2 – Meta-analysis

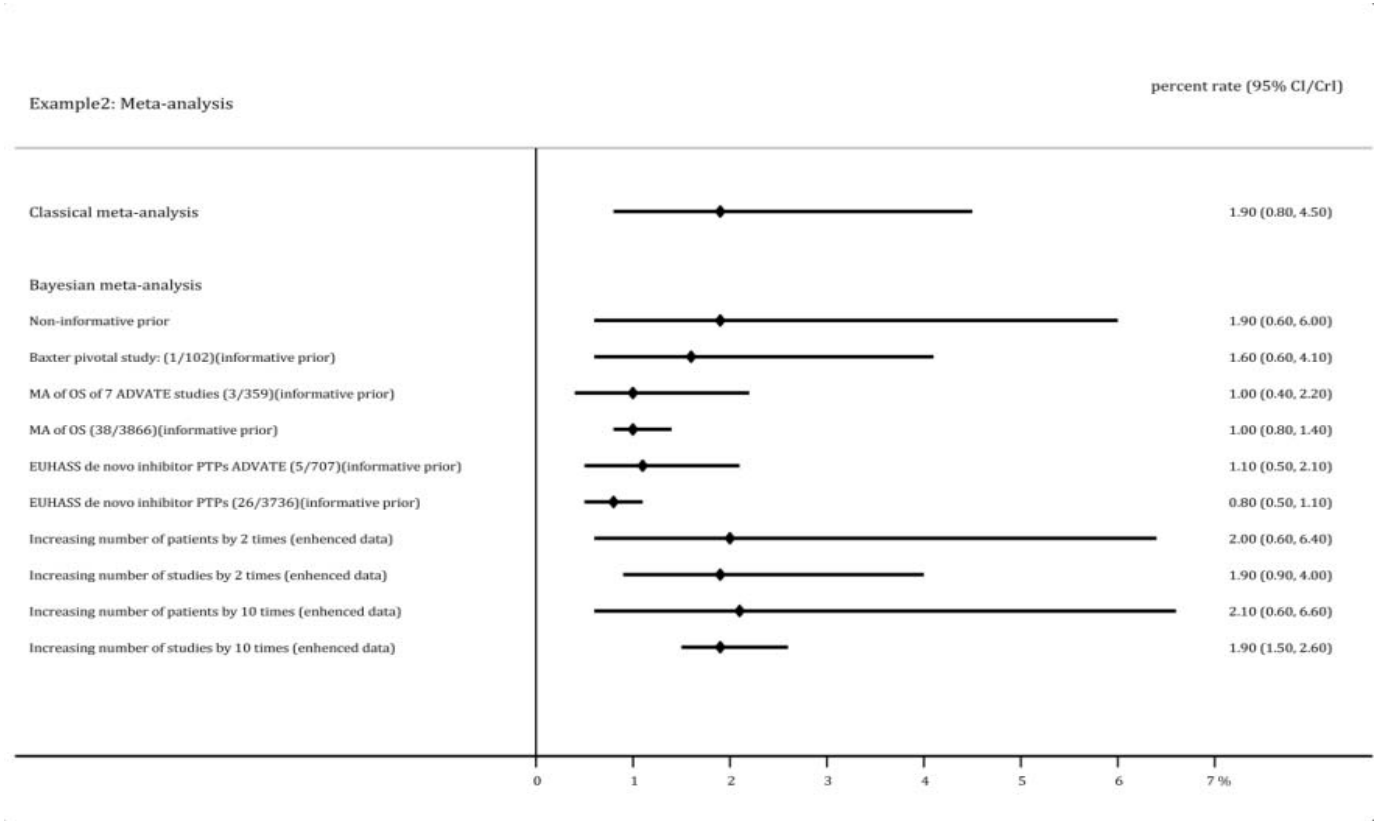
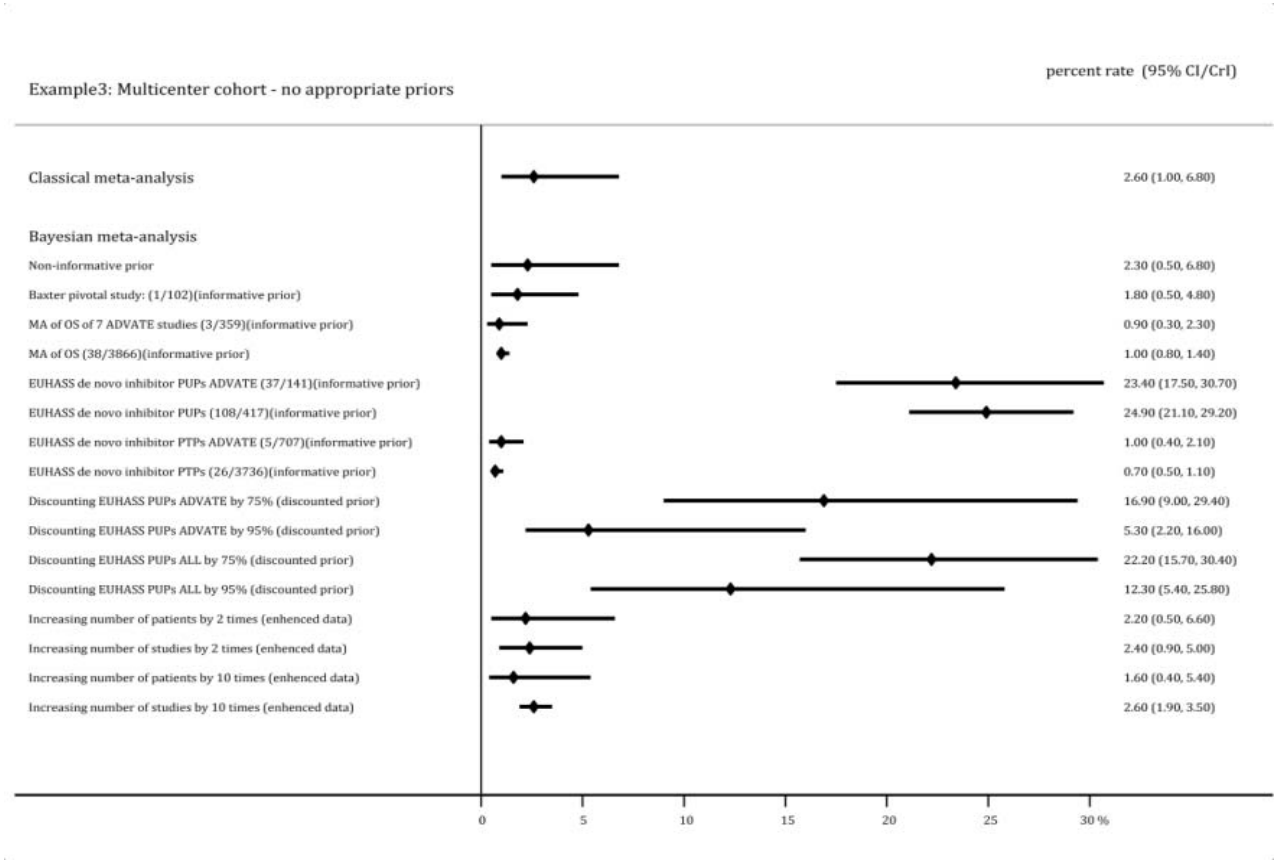


Figure2c: Example 3 – Multicenter cohort – no appropriate priors



Appendix A: Analysis Methods and the choice of priors

	Example 1	Example 2	Example 3
Method	Single study	Meta-analysis	Multi-centric cohort – no appropriate priors
Test data	PASS data [30] (6/428)	PASS data [31] (21/1188)	PASS data [32] (6/219)
Classical Statistical Analysis	Logistic model	Random-effects logistic model	Random-effects logistic model
Bayesian Statistical Analysis	Logistic model	Hierarchical (Random-effects) logistic model	Hierarchical (Random-effects) logistic model
Non-informative prior	OK	OK	OK
Informative prior: Baxter Pivotal Study (1/102)	OK	OK	OK
Informative prior: meta-analysis of OS (7 ADVATE studies) (3/569)	OK	OK	OK
Informative prior: meta-analysis of OS (38/3866) Do you need data per study?	OK	OK	OK
Informative prior: EUHASS study *de novo* inhibitor PUPs ADVATE (37/141)	NO	NO	OK
Informative prior: EUHASS study *de novo* inhibitor PUPs (108/417)	NO	NO	OK
Informative prior: EUHASS study inhibitors in PTPs ADVATE	OK	OK	OK

(5/707)			
Informative prior: EUHASS study inhibitors in PTPs (all FVIII) 22/3736	OK	OK	OK
Discounted prior: Discounting EUHASS PUPs ADVATE by 75%	NO	NO	OK
Discounted prior: Discounting EUHASS PUPs ADVATE by 95%	NO	NO	OK
Discounted prior: Discounting EUHASS PUPs ALL by 75%	NO	NO	OK
Discounted prior: Discounting EUHASS PUPs ALL by 95%	NO	NO	OK
Enhanced data : Enhancing study data by 2 times - increasing number of patients	OK	OK	OK
Enhanced data: Enhancing study data by 2 times - increasing number of studies	NO	OK	YES
Enhanced data: Enhancing study data by 10 times - increasing number of studies	OK	OK	OK
Enhanced data: Enhancing study data by 10 times - increasing number of studies	NO	OK	YES

PASS: post-authorization safety studies

OS: observational study **CI**: confidence interval **CrI**: credible interval

PUP: Previously untreated patient **PTP**: Previously treated patient

EUHASS: European Haemophilia Safety Surveillance

Appendix B: Bayesian codes

Number updates: 300000; Number of chain: 2; Number of thin: 5

Burn in: 10000; Seed: 314159

Model 1 simple logistic regression

```
model {
    r ~ dbin(p, n)
    logit(p) <- mu #log odds
    mu ~ dnorm(0, 1.0E-5) # non-informative
    #mu ~ dnorm (-4.615, 0.990) # PIVOT 1/102, log_odds=log(1/101), var=1/1+1/101
    #mu ~ dnorm (-5.24, 2.984) # Meta-OS ADVATE, 3/569
    #mu ~ dnorm (-4.613, 37.626) # Meta-OS ADVATE, 38/3866
    #mu ~ dnorm (-4.944, 4.965) # EUHASS PTP ADVATE, 5/707
    #mu ~ dnorm (-4.961, 25.819) # EUHASS PTP ALL, 26/3736
    odds<-exp(mu)
    prop<-odds/(1+odds)
    perc<-prop*100
        }
```

Model 2: Random-effects logistic model

```
model {
    for( i in 1 : Num ) {
        r[i] ~ dbin(p[i], n[i])
        logit(p[i]) <- mu[i] #log odds
        mu[i] ~ dnorm(d, tau)
          }
    d ~ dnorm(0,1.0E-5) # Non-informative prior
    # Prior1: Baxter Pivot Trial, 1/102
    #d ~ dnorm(-4.62,0.99) # log_odds, variance=1/1+1/(102-1),
    percision=1/var
     #d ~ dnorm (-5.24, 2.984) # Meta-OS ADVATE, 3/569
    #d ~ dnorm (-4.613, 37.626) # Meta-OS ADVATE, 38/3866
    #d ~ dnorm (-1.033, 27.291) # EUHASS PUP ADVATE, 37/141
    #d ~ dnorm (-1.051,80.029) # EUHASS PUP ALL, 108/417
     #d ~ dnorm (-4.944, 4.965) # EUHASS PTP ADVATE, 5/707
    #d ~ dnorm (-4.961, 25.819) # EUHASS PTP ALL, 26/3736
    #d ~ dnorm (-1.033, 6.822) # EUHASS PUP ADVATE(-75%), 9.25/35.25
     #d ~ dnorm (-1.033, 1.365) # EUHASS PUP ADVATE(-95%), 1.85/7.05
    #d ~ dnorm (-1.051, 20.007) # EUHASS PUP ALL(-75%), 27/104.25
```

```
 #d ~ dnorm (-1.052, 4.001) # EUHASS PUP ALL(-95%), 5.4/20.85
tau<-1/(sigma*sigma)
sigma~dunif(0,2) # between study variance is estimated from PASS1
odds<-exp(d)
prop <- exp(d)/(1+exp(d))
perc<-prop*100
ppos1<-step(10/100-prop)
ppos2<-step(5/100-prop)
ppos3<-step(1/86-prop)
}
```

Appendix C: Assessing the impact of sample size change and priors choice on the Bayesian posterior estimates (Example 3 [32])

	Original data	Increasing sample size by 2 times		Increasing sample size by 10 times	
Test data (number of inhibitors/number of patients); number of centers	Example 3 : PASS (6/219); 7	Increasing No. of patients in each center: (12/438); 7	Increasing No. of centers: (12/438); 14	Increasing No. of patients in each center: (60/2190); 7	Increasing No. of centers: (60/2190): 70
Non-informative prior	2.3 (0.5, 6.8)	2.2 (0.5, 6.6)	2.4 (0.9, 5.0)	1.6 (0.4, 5.4)	2.6 (1.9, 3.5)
*Informative prior: Baxter Pivotal Study (1/102)	1.8 (0.5, 4.8)	1.8 (0.5, 4.6)	2.1 (0.8, 4.2)	1.4 (0.4, 3.9)	2.6 (1.8, 3.4)
Informative prior: meta-analysis of OS (7 ADVATE studies) (3/569)	0.9 (0.3, 2.3)	0.9 (0.3, 2.4)	1.3 (0.5, 2.7)	0.8 (0.3, 2.0)	2.3 (1.5, 3.1)
Informative prior: meta-analysis of OS (38/3866)	1.0 (0.8, 1.4)	1.0 (0.8, 1.4)	1.1 (0.8, 1.5)	1.0 (0.7, 1.4)	1.4 (1.0, 1.8)
Informative prior: EUHASS study *de novo* inhibitor PUPs ADVATE (37/141)	23.4 (17.5, 30.7)	23.2 (17.3, 30.4)	21.4 (15.6, 28.1)	22.8 (17.0, 29.8)	11.2 (7.7, 15.4)
Informative prior: EUHASS study *de novo* inhibitor PUPs (108/417)	24.9 (21.1, 29.2)	24.8 (20.1, 29.1)	24.1 (20.4, 28.3)	24.7 (20.8, 28.9)	19.2 (16.1, 22.5)
Informative prior: EUHASS study inhibitors in PTPs ADVATE (5/707)	1.0 (0.4, 2.1)	1.0 (0.4, 2.1)	1.2 (0.6, 2.4)	0.9 (0.4, 1.9)	2.3 (1.5, 3.0)
Informative prior: EUHASS study inhibitors in PTPs (all FVIII) (26/3736)	0.7 (0.5, 1.1)	0.8 (0.5, 1.1)	0.8 (0.5, 1.2)	0.7 (0.5, 1.1)	1.2 (0.8, 1.7)

PASS: post-authorization safety studies

OS: observational study **CI**: confidence interval **CrI**: credible interval

PUP: Previously untreated patient **PTP**: Previously treated patient

EUHASS: European Haemophilia Safety Surveill

CHAPTER 5

CONCLUSIONS

Properly conducting statistical analysis is one of the essential steps towards the success of any health research project. However, finding a suitable analytic model or approach can sometimes be challenging. Choosing one method over its alternatives often involves comparing and evaluating all available analysis options based on the understanding of the underlying statistical assumptions and the nature of the outcomes. In this manuscript-based thesis, I investigated three situations where the use of different statistical models or approaches could impact the results of the analyses. Situation (1): the choice of statistical models may affect the analysis results. In particular, I compared the commonly used models in analyzing correlated choice experiment (DCE) data. Situation (2): including or excluding a sub-set of data may affect the analysis results. In this case, I assessed the impact of including or excluding both-armed zero-event (BAZE) studies on the pooled results for meta-analyses (MA). Situation (3): incorporating in external information may affect the analysis results of the current study data. In this case, I showed how the Bayesian approaches used to incorporate the external information may

affect the understanding of the evidence contained in the current research data. In this chapter, I will summarize the findings, discuss the implications and limitations and shed light on future studies.

In Chapter2, I empirically compared commonly-used statistical models for analyzing correlated data of DCE survey while accounting for within-participant correlation. The data used in this project were from a choice survey conducted in Hamilton, Ontario, Canada in 2002 (ref), which was designed to evaluate patient preference for the various colorectal cancer (CRC) screening tests to identify the key attributes and levels that may influence the uptake of CRC screening tests. This DCE study used a two-stage design: the choice between two hypothetical tests at the first stage, and the choice between taking the preferred test and opting out. This design gave us the chance to define the outcome in three ways: 1) binary outcome (Test A/B; Yes to the test/ No), 2) multinomial outcome (Test A/B/No), and 3) bivariate outcome (A/B and Yes (to A or B)/No). For the clustered binary outcomes, six models were investigated: logistic and probit regressions using cluster-robust standard error (SE), random-effects and generalized estimating equation approaches. For the clustered multinomial outcomes, three models were applied: multinomial logistic and probit regressions with cluster-robust SE and random-effects

multinomial logistic models. And for the bivariate outcomes: bivariate probit models with cluster-robust SE were fitted.

The following findings and observations of comparison of the results from different models may be worth further consideration in future DCE design and analysis.

1) When participants were repeatedly asked to make a choice between two alternative tests (combinations of the attributes at different levels) at stage one, there was little within-participant correlation. However, when the choices were between participation (on the preferred test chosen at stage one) and opt-out, the within-participant correlation was substantial.
2) The results from different models were consistent when little within-participant correlation was present. Therefore, simple logistic model (for binary outcome) or multinomial logistic model (for multinomial outcomes) is as a good choice as other complicated statistical models.
3) When there was substantial within-participant correlation, the results were inconsistent between different methods used to account for intra-class correlation.
4) The observed within-participant correlation was likely caused by the participant's pre-determination on participation or opt-out for the screening tests. This pre-

determination seemed also lead to the ordering effect which might bias the estimates of participant preference of the screen test.

5) Participant preference on the cost-related attribute may not preform linearly, which violates the linear utility assumption.

The most import contribution of this paper is that, to my best knowledge, this was one of the first studies to investigate the commonly used statistical models in accounting for the within-participant correlation (intra-class correlation) issue in DCE data conducted for health research. It has been well recognized that the data collected for the studies using repeated measurement or cluster design are correlated. Taking this type of correlation into account is crucial in both the study design and data analysis stages. Ignoring the intra-class correlation will lead to under-calculated sample size[1] and biased estimates of the parameter [2]. However, this issue has not drawn enough awareness for designing and analyzing DCE studies. Both ISPOR (International Society for Pharmacoeconomics and Outcomes Research) guidelines for constructing DCE studies[3] and analyzing DCE data [4] did not mention how to deal the potential within-participant correlation. Although this empirical study cannot provide the answer as to which model is superior for accounting for within-participant correlation while analyzing DCE data, it points out that it is an important issue and needs of further investigations. Given the different statistical models

available to analyze DCE studies, I believe that when analyzing DCE data with potential within-participant correlation, the analysis results obtained from the primary statistical model need to be examined thoroughly through sensitivity analysis for robustness by checking the consistency and discrepancy. For policy makers, we recommend exercising caution in interpreting findings from DCE studies.

There are some limitations to this study. First, this is an empirical study. It cannot serve as a direct tool to find the "best" statistical model to analyze DCE data with within-participant correlation. Second, the data were collected through a study with a two-stage design with the forced choice between the screen tests at stage-one and the opt-out option at stage-two. The intention on adopting this design was to maximize the information collected about participant preferences on the screen tests. However, by forcing the participants who have already decided to decline the screen test before even seeing the questionnaire, the data collected to elicit participant preference on the tests may not accurately reflect the facts. Third, I focused on comparing the commonly-used statistical models which were available through standard statistical software. Some complicated models with potential advantages, particularly on dealing with nonlinear utility functions such as Bayesian random-effects and GEE models with polynomial logit function, were omitted from this paper.

In Chapter 3, I conducted a simulation study to investigate the impact of including both-arm zero-event (BAZE) studies in small meta-analysis (MA) for rare event outcomes for standard meta-analysis methods with continuity correction. It is not difficult to logically deduce that including BAZE studies in meta-analysis for rare event outcomes may introduce bias in estimating the treatment effect. This simulation is the first study to confirm and quantify this hypothesis.

The key findings in this chapter include:

1) I confirmed that including BAZE studies in MA using continuity correction methods provided unbiased point estimates of OR and narrowed the 95% confidence interval when there was no true treatment effect existing between treatment and control arms.
2) I verified my hypotheses that when a true treatment effect existed, including BAZE studies in MA added bias by pulling the point estimates of OR towards the null hypothesis in the direction of underestimating treatment effects, and the bias increased substantially with decreasing event rate, number of patients and increasing treatment effect and between study variance.

3) My study once again proved that the Peto method without including BAZE studies generated the least biased results when the event rate was low, treatment effect was small to moderate, and between-study variance was small to moderate.
4) My study also showed that when there was a true between-study variance, the Peto method still out-performed random-effects models by providing the least biased point estimate.

The focus of my study was s to investigate the bias which might be introduced in the estimates of small MA for rare events by including BAZE studies using the standard pooling methods with continuity correction. A certain degree of bias towards the null hypothesis provides more conservative estimates when evaluating beneficial treatment outcomes between new and standard treatments[5–7]. It is considered a safer approach for patient care. My main concern is that when evaluating harmful events such as serious adverse event (SAE), this type of bias, i.e. underestimating harm may hinder the action on stopping the unsafe treatment [8]. Most RCTs are not designed to investigate rare harmful events, e.g. SAE, and thus the sample size is not large enough to detect the true proportion of such events due to low statistical power. With the combination of extremely low event rates and insufficient sample size, zero events are very likely to be observed by chance. Including BAZE studies in MAs for the purpose of evaluating the safety-related

outcomes may lead to the risk of underrating the harm. The impact of including BAZE study may also be different depending on the size of the MA. Currently, with the no guideline being established, the approaches to dealing with the BAZE studies in MAs are varied. Therefore, I recommend sensitivity analyses to assess the consistency and discrepancy by including and excluding BAZE studies in MAs. I believe that an extension of PRISMA statement on reporting the approaches to dealing with zero-event studies (including either-arm zero-event and both-arm zero-event) in MA is necessary to communicate the results of MA on rare event outcomes with full transparency. The next phase of this investigation will involve creating a checklist to summarize the recommendations for dealing with zero-event studies in MAs. Last, but not the least, the results of MAs of rare event outcomes need to be cautiously interpreted within the clinical contents.

I set up my simulation based on the following scenarios: 1) rare event outcomes, 2) small meta-analysis, and 3) standard MA pooling methods for commonly used meta-analysis software. Therefore, the findings in this study cannot be extended to all types of MAs. First, I investigated only the effect measure OR by incorporating BAZE studies using default continuity correction options, i.e. adding 0.5 to all cells. Although the results can be implied to the standard MA using RR (relative risk or risk ratio) as pooled estimates, I

cannot make the similar conclusion for the meta-analysis using a statistical model based on likelihood maximization, such as Poison random-effects model. I also left the Bayesian meta-analysis for future investigation. Bayesian approach has a probability-based sampling mechanism[9] which may provide the means to reduce the bias introduced by including BAZE studies.

In Chapter 4, I illustrated how Bayesian statistical methods can be used to incorporate other relevant evidence via priors to enhance or modify the evidence presented in the current study data. The data I used for my scenario-based analysis were from three published PASS (Post-Authorization Safety Surveillance) studies[10–12], which were single-armed Phase IV trials conducted to evaluate the safety outcomes. The patient-level data were provided by Baxter Healthcare, Global Affairs (Westlake Village, California, USA). The outcome I used in this study was inhibitor, a rare serious adverse event which may report among the patients undertaking the medication treatment for hemophilia A. In this paper, I compared the use of three different types of priors in incorporating external information: non-informative prior, informative prior and discounted prior through three study scenarios: 1) estimating event rate in a single cohort study, 2) pooling estimates for a set of studies using meta-analysis, and 3) generating estimates from small studies in a previously unexplored study population.

The key points demonstrated in this chapter are:

1) Results from Bayesian statistical models with non-informative priors are comparable to the classical (Frequentist) approaches on estimating the rare event rate.
2) Incorporating external information through informative priors can enhance the evidence presented in the study data.
3) Borrowing information from previously studied similar populations through informative priors can create a range of estimates for an unstudied population.
4) Bayesian probability can be directly used to quantify the comparison between the evidence obtained through the current study and a threshold.
5) The evidence can be weighted through discounting the prior information or scaling up the presentation of the study data

This study investigated how Bayesian methods can be used to optimize the evidence for rare event data for current study by maximizing the use of existing knowledge through priors. Furthermore, I demonstrated how Bayesian estimates can be utilized to inform clinical decisions in patient management in complex clinical settings. The success of integrating all relevant evidence through a Bayesian approach depends on two aspects: 1) how to properly choose the clinically relevant priors, and 2) how to statistically formulate the clinical knowledge. These tasks need the joint force of clinicians and statisticians.

Properly implementing the Bayesian results in the clinical decision making depends on the comprehensive understanding of the evidence, in particular for the findings obtained from the first-time-ever exploration regarding new study settings or populations. Therefore, comparing the results obtained using different priors can be useful to strengthen the existing evidence by assessing the consistency and explore the uncertainty for new findings by examining the discrepancy.

In regarding the purpose of serving as examples of how to conduct Bayesian analysis for hematologists to analyze and generate evidence for rare events among rare study populations, I chose to use Bayesian random-effect logistic regression throughout the entire project for three scenarios for simplicity. I am aware that other statistical models such as random-effect Poisson regress may be a better choice for rare event data with zero outcomes. It is worth noting that properly setting up priors can be challenging because it depends on the type of outcomes and Bayesian models.

Statistical analysis is never as simple as “my-way-or-the-highway”. It is a comprehensive process involving assessing, comparing and decision making on study samples, statistical models and relevant information. And for studies with complicated design, data structure

or content, the choice of an appropriate analytical strategy relies on the comparison of the alternatives. In this PhD thesis, I discussed three cases where sensitivity analysis was helpful in this regard. I hope my work will bring awareness to the importance of conducting sensitivity analysis for health research projects.

Reference

1 Thompson DM, Fernald DH, Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: Estimates from the robert wood johnson prescription for health projects. *Ann Fam Med* 2012;**10**:235–40. doi:10.1370/afm.1347

2 Campbell MK, Mollison J, Steen N, *et al.* Analysis of cluster randomized trials in primary care: a practical approach. *FamPract* 2000;**17**:192–6.PM:10758085

3 Johnson FR, Lancsar E, Marshall D, *et al.* Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value Heal* 2013;**16**:3–13. doi:10.1016/j.jval.2012.08.2223

4 Ispor Task Force. Conjoint Analysis Statistical Analysis: An ISPOR Conjoint Analysis Good Research Practices Task Force Report. ;:1–33.

5 Hayward RA, Kent DM, Vijan S, *et al.* Reporting clinical trial results to inform providers, payers, and consumers. *Health Aff (Millwood)*;**24**:1571–81. doi:10.1377/hlthaff.24.6.1571

6 Nuesch E, Trelle S, Reichenbach S, *et al.* Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. 2010;**341**. doi:10.1136/bmj.c3515

7 Zhang Z, Xu X, Ni H. Small studies may overestimate the effect sizes in critical care meta-analyses : a meta- epidemiological study. *Crit Care* 2013;**17**:R2. doi:10.1186/cc11919

8 Miller DB, Humphries KH. A new way to evaluate randomized controlled trials ? New approach does more harm than good. 2005;:241–4.

9 Sutton a J, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;**10**:277–303. doi:10.1191/096228001678227794

10 Oldenburg J, Goudemand J, Valentino L, *et al.* Postauthorization safety surveillance of ADVATE [antihaemophilic factor (recombinant), plasma/albumin-free method] demonstrates efficacy, safety and low-risk for immunogenicity in routine clinical practice. *Haemophilia* 2010;**16**:866–77. doi:10.1111/j.1365-2516.2010.02332.x

11 Marcucci M, Cheng J, Oldenburg J, *et al.* Meta-analysis of Post Authorization Safety Studies: worldwide postmarking surveillance of hemophilia A patients

treated with antihemophilic factor recombinant plasma/albumin-free method rAHF-PFM. *J Thromb Haemost* 2013;**11**:1075.http://onlinelibrary.wiley.com/doi/10.1111/jth.12443/pdf

12 Romanov V, Marcucci M, Cheng J, *et al.* Evaluation of Safety and Effectiveness of factor VIII treatment in Hemophilia A patients with low titer inhibitors or a personal history of inhibitor. *Thromb Haemost* 2015;**113**:Inpress.