BAYESIAN EVIDENCE SYNTHESIS IN HEALTH TECHNOLOGY ASSESSMENT

BAYESIAN APPROACHES FOR SYNTHESISING EVIDENCE IN HEALTH TECHNOLOGY ASSESSMENT

By C. ELIZABETH McCARRON, B.A., M.A., M.Sc.

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

McMaster University DOCTOR OF PHILOSOPHY (2011) Hamilton, Ontario (Health Research Methodology)

TITLE: Bayesian Approaches for Synthesising Evidence in Health Technology Assessment AUTHOR: C. Elizabeth McCarron, B.A., M.A., M.Sc. (McMaster University) SUPERVISOR: Dr. Jean-Eric Tarride NUMBER OF PAGES: ix, 145

ABSTRACT

Background and Objectives: Informed health care decision making depends on the available evidence base. Where the available evidence comes from different sources methods are required that can synthesise all of the evidence. The synthesis of different types of evidence poses various methodological challenges. The objective of this thesis is to investigate the use of Bayesian methods for combining evidence on effects from randomised and non-randomised studies and additional evidence from the literature with patient level trial data.

Methods: Using a Bayesian three-level hierarchical model an approach was proposed to combine evidence from randomised and non-randomised studies while adjusting for potential imbalances in patient covariates. The proposed approach was compared to four other Bayesian methods using a case study of endovascular versus open surgical repair for the treatment of abdominal aortic aneurysms. In order to assess the performance of the proposed approach beyond this single applied example a simulation study was conducted. The simulation study examined a series of Bayesian approaches under a variety of scenarios. The subsequent research focussed on the use of informative prior distributions to integrate additional evidence with patient level data in a Bayesian cost-effectiveness analysis comparing endovascular and open surgical repair in terms of incremental costs and life years gained.

Results and Conclusions: The shift in the estimated odds ratios towards those of the more balanced randomised studies, observed in the case study, suggested that the proposed Bayesian approach was capable of adjusting for imbalances. These results were reinforced in the simulation study. The impact of the informative priors in terms of increasing estimated mean life years in the control group, demonstrated the potential importance of incorporating all available evidence in the context of an economic evaluation. In addressing these issues this research contributes to comprehensive evidence based decision making in health care.

PREFACE

This thesis is a “sandwich thesis” that combines three individual manuscripts prepared for publication in peer-reviewed journals. One of the manuscripts is already published, another has been accepted for publication and the third is in submission. The contributions of C. Elizabeth McCarron to all of the papers in the thesis include: developing the research ideas and research questions, performing the analyses, interpreting the results, writing all of the manuscripts, submitting all of the manuscripts, and responding to reviewers’ comments. The work in this thesis was conducted between the winter of 2009 and the fall of 2011.

ACKNOWLEDGEMENTS

I wish to express my appreciation for the assistance given by my Ph.D. supervisor Dr. Jean-Eric Tarride. I also wish to thank the other members of my Ph.D. supervisory committee, Dr. Eleanor Pullenayegum, Dr. Lehana Thabane, and Professor Ron Goeree for all of the help and guidance they have provided.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

APPENDIX

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

APPENDIX

CHAPTER 1

INTRODUCTION

Background and Rationale

The International Network of Agencies for Health Technology Assessment defines health technology assessment (HTA) as, “a multidisciplinary field of health policy analysis studying the medical, social, ethical, and economic implications of development, diffusion, and use of health technology” [1]. Examples of health technologies include pharmaceuticals, devices, and surgical procedures [1]. In health systems throughout the world, HTA plays an essential role in supporting decision making about access to technology, its diffusion, and innovation [2].

In a cost-containment environment, economic evaluation plays an important role in HTA. The economic evaluation of health care technologies involves the comparison of alternative interventions in terms of their relative costs and effects [3]. By comparing costs and effects, economic evaluations inform decision making regarding the efficient allocation of scarce resources. Cost-effectiveness research is used as formal inputs into decisions about which interventions and programmes should be funded from collective resources by health systems around the world [4]. The increasing use of economic evaluations to inform health care decision making raises important methodological issues for this area of research. One of these issues is the need to synthesise evidence on effects from all sources of available evidence [4]. Depending on the technologies being compared, the body of available evidence could include a variety of different sources (e.g., randomised controlled trials (RCTs), non-randomised/observational studies).

Bayesian statistical methods represent a valuable set of analytical tools for combining evidence from different sources [4]. Based on Bayes' theorem, named after the 18th-century Presbyterian minister Thomas Bayes, Bayesian statistics have enjoyed a revival in recent years [5]. While the application of Bayesian methods to the economic evaluation of health care technology is relatively new, the potential for these methods to take into account all available evidence to inform decision making is profound. At its core, Bayes' theorem describes a process of how to modify existing beliefs as additional information becomes available. With this comes the opportunity for a more iterative approach to health care decision making, one flexible enough to take advantage of all available evidence. An example of such an approach, within a real world health policy setting, is seen in the Programs for Assessment of Technology in Health (PATH) Research Institute reduction of uncertainty through field evaluation (PRUFE) framework established in Ontario, Canada [6].

The objective of the PRUFE framework is to provide a comprehensive evidence base upon which informed decisions regarding the cost-effectiveness of new health technologies can be made [6]. This process relies on the use of both systematic literature reviews and patient level field evaluations to provide research comparing the costs and effects of these technologies. A recent example revolved around the decision by the Ontario Ministry of Health and Long-Term Care to reimburse endovascular aneurysm repair (EVAR) for abdominal aortic aneurysms in patients at a high risk for morbidity or mortality following surgery. The available evidence consisted of a systematic review, including both randomised and non-randomised studies [7], as well as a patient level economic evaluation comparing EVAR and open surgical repair (OSR) in high risk patients [8]. However, the randomised and non-randomised studies were analysed separately and the patient level data were analysed using standard techniques (i.e., non-parametric bootstrap methods) that gave no consideration to external evidence. In both instances, the analyses failed to take advantage of all of the available evidence. Critical to the success of approaches such as the PRUFE framework, is the existence of methods capable of combining evidence from different sources.

Methodological Issues in Evidence Synthesis

Depending on the types of evidence being combined, a researcher may face various methodological challenges. The specific issues addressed in this thesis are: 1) how to combine evidence from randomised and non-randomised studies, and 2) how to combine patient level data from a trial based economic evaluation with additional evidence from the literature. The objective of this dissertation is to investigate the use of Bayesian methods for combining evidence from these different sources. The methods are examined against the background of a comparison of EVAR and OSR for the treatment of abdominal aortic aneurysms.

Issue 1: ***Combining randomised and non-randomised studies***

Beyond the importance of basing health care decision making on all available evidence, there may be other practical reasons to combine different types of comparative evidence (e.g., randomised and non-randomised studies). For certain health care technologies, especially non-drug technologies, there may be a lack of randomised studies [4]. In addition, when synthesising evidence on effects as part of an economic evaluation used to inform decision making at a population level, issues may arise concerning effectiveness relative to efficacy. RCTs are designed to provide estimates of efficacy in an ideal setting, while non-randomised or observational studies may better reflect estimates of the effectiveness of the treatments in the real world. In exchange for the greater generalisability associated with non-randomised studies, there is also an increased likelihood of imbalances among patient characteristics due to the non-randomised nature of the studies [9]. These imbalances, if not accounted for in some way, could bias the results. The extent to which bias in the results is affected by factors such as the impact of the imbalances, the relative number of randomised and non-randomised studies and the study arm sizes must also be understood.

Issue 2: ***Combining patient level trial data and additional evidence***

Economic evaluations of patient level data refer to studies involving primary data collection, usually from alongside a RCT [3]. Traditional approaches for analysing patient level economic evaluations rely solely on the information contained in the trial data [3]. The consequence of which is that these analyses effectively ignore all other external sources of evidence. When the results of these economic evaluations are used to inform decision making, the failure to take into consideration all of the available evidence could have important health policy implications. Not only could the results influence decisions regarding the funding of one intervention compared to another, but they could also have an impact on decisions regarding the need for future research. Another important issue relates to how to value the additional evidence relative to the patient level data, which would require a careful consideration of potential differences between the two sources of information. Despite the potential for Bayesian methods to take advantage of all available evidence, a recent review of Bayesian patient level economic evaluations found that only half of the included studies used some type of informative prior in their analysis [10]. Where there are sources of evidence in addition to the trial data, not incorporating these into the analysis fails to exploit the full potential of the Bayesian approach and could undermine the results.

Outline for the Thesis

This thesis consists of three papers that are related to the issues discussed above. The three papers are separated into different chapters beginning from Chapter 2.

Using a case study, Chapter 2 proposes a new approach for combining evidence from randomised and non-randomised studies by adjusting for imbalances in patient characteristics between study arms. The proposed approach extends the Bayesian three-level hierarchical model initially developed by Prevost et al. [11] for combining different types of study designs. Imbalances in patient characteristics are adjusted for using differences in the patient characteristics between study arms. The proposed method is compared to four other Bayesian approaches: 1) three-level hierarchical model unadjusted for potential imbalances as in Prevost et al. [11], 2) three-level hierarchical model adjusted using aggregate study values [11], 3) informative prior downweighting the non-randomised studies [12], and 4) prior constraint downweighting the non-randomised studies [11]. This comparison is undertaken using data from the previously published systematic review of EVAR versus OSR for the treatment of abdominal aortic aneurysms [7].

A simulation study is conducted in Chapter 3 to assess the performance of the newly proposed Bayesian approach beyond the single case study used in Chapter 2. Sets of balanced randomised and imbalanced non-randomised studies were generated using simulation techniques. The same models considered in Chapter 2 are also investigated in this chapter with the exception of the prior constraint model. Six scenarios were examined to assess the sensitivity of the results to changes in the impact of the imbalances and the relative number and size of studies of each type. The values selected for the various criteria were meant to reflect realistic scenarios such as might be encountered in practice (e.g., more patients enrolled in non-randomised studies than randomised studies [12]), yet at the same time testing the robustness of the proposed approach.

Chapter 4 explores the use of informative Bayesian prior distributions as a mechanism by which additional evidence can be incorporated into a trial based economic evaluation. The additional evidence from the literature is combined in a Bayesian meta-analysis and is then used to inform the prior distributions for effects in the two arms of the trial. The patient level data were taken from the previously published economic evaluation that compared EVAR and OSR in high risk patients using a one year prospective observational study [8]. The additional evidence represents information that was available at the time of the original economic evaluation and was taken from the same published systematic review used in Chapter 2 [7]. Two types of informative priors were examined to represent different potential valuations of the additional evidence relative to the patient level data. Namely, the external information was taken both at face value as well as being treated with scepticism by explicitly downweighting its information relative to that from the patient level data.

Chapter 5, the concluding chapter of the thesis, provides an overall summary of the research. This chapter also identifies potential areas of future research as well as discussing the implications and contributions of the work undertaken in this thesis.

CHAPTER 2

The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms

C Elizabeth McCarron[1,2§], Eleanor M Pullenayegum [1,3], Lehana Thabane[1,3], Ron Goeree[1,2], Jean-Eric Tarride[1,2]

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

[2]Programs for Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[3]Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[§]Corresponding author

Citation: McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Medical Research Methodology*. 2010; 10:64. doi: 10.1186/1471-2288-10-64.

Abstract

Background: Informing health care decision making may necessitate the synthesis of evidence from different study designs (e.g., randomised controlled trials, non-randomised/observational studies). Methods for synthesising different types of studies have been proposed, but their routine use requires development of approaches to adjust for potential biases, especially among non-randomised studies. The objective of this study was to extend a published Bayesian hierarchical model to adjust for bias due to confounding in synthesising evidence from studies with different designs.

Methods: In this new methodological approach, study estimates were adjusted for potential confounders using differences in patient characteristics (e.g., age) between study arms. The new model was applied to synthesise evidence from randomised and non-randomised studies from a published review comparing treatments for abdominal aortic aneurysms. We compared the results of the Bayesian hierarchical model adjusted for differences in study arms with: 1) unadjusted results, 2) results adjusted using aggregate study values and 3) two methods for downweighting the potentially biased non-randomised studies. Sensitivity of the results to alternative prior distributions and the inclusion of additional covariates were also assessed.

Results: In the base case analysis, the estimated odds ratio was 0.32 (0.13,0.76) for the randomised studies alone and 0.57 (0.41,0.82) for the non-randomised studies alone. The unadjusted result for the two types combined was 0.49 (0.21,0.98). Adjusted for differences between study arms, the estimated odds ratio was 0.37 (0.17,0.77), representing a shift towards the estimate for the randomised studies alone. Adjustment for aggregate values resulted in an estimate of 0.60 (0.28,1.20). The two methods used for downweighting gave odd ratios of 0.43 (0.18,0.89) and 0.35 (0.16,0.76), respectively. Point estimates were robust but credible intervals were wider when using vaguer priors.

Conclusions: Covariate adjustment using aggregate study values does not account for covariate imbalances between treatment arms and downweighting may not eliminate bias. Adjustment using differences in patient characteristics between arms provides a systematic way of adjusting for bias due to confounding. Within the context of a Bayesian hierarchical model, such an approach could facilitate the use of all available evidence to inform health policy decisions.

Background

Health technology assessment has been defined as a multidisciplinary field of policy analysis studying the medical, social, ethical, and economic implications of development, diffusion, and use of health technology [1]. Evidence on the effects of interventions from comparative studies is a critical component of this process. The different types of study designs (e.g., randomised, non-randomised/observational) used to assess the effects of interventions can be arranged into a hierarchy, at the top of which is the randomised controlled trial (RCT) [2]. Randomisation increases the likelihood that the treatment groups will be balanced in terms of known and unknown prognostic or confounding variables. Consequently the treatment effects estimated from RCTs are less subject to the potential confounding effects of extraneous variables [3]. Evidence from RCTs alone, however, may not be sufficient to inform decision makers. In particular, the strict inclusion and exclusion criteria which are often applied in RCTs may limit their generalisability relative to non-randomised studies [4,5]. In some cases, compliance to randomisation, among the randomised studies, might also be an issue. Furthermore, the scarcity of randomised studies for certain non-drug technologies, such as medical devices and surgical procedures, may necessitate the use of evidence from non-randomised studies in addition to that available from randomised studies [4]. Contrary to ignoring evidence from non-randomised studies, it has been argued that all available evidence should be used to inform health care decision making [4,5,6,7]. Such an approach requires methods capable of synthesising evidence from both randomised and non-randomised studies.

Bayesian hierarchical modelling [5,8] has recently been proposed for synthesising evidence from randomised and non-randomised studies. Prevost et al. [5] applied their method to combine evidence relating to the relative risk for mortality from five randomised trials and five non-randomised studies evaluating mammographic screening. Other applications of Prevost's model include Grines et al. [9] and Sampath et al. [10].

As an extension to the model, Prevost et al. [5] proposed the inclusion of study covariates to explain differences in mean effects at the study type level. Although this is important, the authors did not model differences between study arms, which may be a limitation of this approach when dealing with non-randomised studies due to potential differences in baseline characteristics. Adjustment made using aggregate values will not account for potential imbalances between study arms resulting from the lack of randomisation. Another extension proposed by Prevost made use of a prior constraint, reflecting the assumption that evidence from non-randomised studies, having been derived from study designs with potential weaknesses [4], may be more biased than evidence from randomised studies. The effect of the prior constraint is to downweight the evidence from the non-randomised studies. This approach has been criticized as it may not eliminate bias [11].

The objective of this paper was to extend the Bayesian three-level hierarchical model developed by Prevost et al. [5] in order to accommodate the greater potential for bias among the non-randomised studies by adjusting study estimates for potential confounders using differences in patient characteristics between study arms. Modeling differences between study arms is important in order to correct for potential imbalances within studies which could bias the results. We applied this new model to a subset of studies from a systematic review of endovascular (EVAR) and open surgical repair (OSR) in the treatment of abdominal aortic aneurysms (AAAs) [12]. The results were compared with those using covariates representing aggregate values for patient characteristics (e.g., mean age) within studies, as in Prevost et al. [5] and Sampath et al. [10], and with two approaches for downweighting biased evidence. Prevost's prior constraint to downweight the non-randomised studies was considered as well as an additional approach that combined a prior distribution based on the non-randomised studies with data from the randomised studies [8].

Methods

Prevost's original Bayesian three-level hierarchical model

The three-level Bayesian hierarchical model proposed by Prevost et al. [5] extends the standard two-level random-effects meta-analysis [13] to include an extra level to allow for variability in effect sizes between different types of evidence (e.g., randomised versus non-randomised study designs). In addition to variability between study estimates within each study type, this model has the capacity to deal with any added uncertainty due to study design [14]. The three levels allow for inferences to be made at the study, study type, and population levels. Although the model can accomodate more than two types of study designs, the application presented by Prevost et al. [5] combined evidence from two study types, randomised and non-randomised.

This model can be written as follows:

$$y_{ij} \sim \text{Normal}(\psi_{ij}, s_{ij}^2) \qquad \text{(eq.1)}$$

$$\psi_{ij} \sim \text{Normal}(\theta_i, \sigma_i^2) \qquad \text{(eq.2)}$$

$$\theta_i \sim \text{Normal}(\mu, \tau^2) \qquad \text{(eq.3)}$$

$$(i = 1 \text{ or } 2 \text{ for the 2 study types; } j = 1,\ldots,k_i \text{ studies}).$$

At the first level of the model (eq.1), y_{ij} is the estimated log relative risk in the jth study of type i, which is normally distributed with mean ψ_{ij} and variance s_{ij}^2. The ψ_{ij} represent the underlying effect, on the log relative risk scale, in the jth study of type i. At the second level of the model (eq.2), the ψ_{ij} are distributed about an overall effect for the ith type of study, θ_i, with σ_i^2 representing the between-study variability for studies of type i. At the third level of the model (eq.3) the study-type effects are distributed about an overall population effect, μ, with τ^2 representing the between-study-type variability.

To try to explain between study heterogeneity, Prevost et al. [5] extended their model to include a covariate for age at the study type level. This is shown in the equation below.

$$\psi_{ij} \sim \text{Normal}(\theta_i + (\alpha \times x_{ij}),\ \sigma_i^2) \qquad \text{(eq.4)}$$

In equation 4, x_{ij} took the values of 0 and 1 for studies of women aged less than 50 years and studies of women 50 years and over, respectively. The same approach was used by Sampath et al. [10] to adjust for study covariates representing continuous variables such as average age and proportion of males in each study. Grines et al. [9] did not conduct covariate adjustment but rather used funnel plots to assess heterogeneity among individual study estimates.

Extension of Prevost's model to adjust for imbalances between study arms

While heterogeneity refers to unexplained variation, bias refers to systematic deviations from the true underlying effect due, for example, to imbalances between study arms [2]. One potential source of bias is confounding [15], where an extraneous factor is associated with both the exposure under study (e.g., treatment) and the outcome of interest, but is not affected by the exposure or outcome [16]. Only when the groups being compared are balanced in all factors, both those that can be measured and those that cannot, that are associated with exposure and that affect the outcome (other than treatment) will it be certain that any observed differences between the groups are due to treatment and not the result of the confounding effects of extraneous variables. Randomisation increases the likelihood that the groups will be balanced not only in terms of the variables that we recognize and can measure but also in terms of variables that we may not recognize and may not be able to measure (i.e., unknowns) but that nevertheless may affect the outcome [3]. In contrast, the greater likelihood of imbalances within the non-randomised studies could have implications especially when combining both types of study designs. In order to deal with this problem, we extended Prevost's three-level model to adjust for differences within studies rather than adjusting for aggregate values at the study type level as in equation 4. The proposed approach uses the variation in imbalances across studies to adjust for differences in patient characteristics between treatment arms within studies. As with RCTs, the resulting balance in patient characteristics within studies should avoid the influence of confounding.

The following presents an extension of Prevost's model based on odds ratios, but could be extended to relative risk. This analysis was undertaken using a binomial model in which the odds of the event (e.g., death) are calculated for each study and study arm level information is incorporated in the model. The model can be written as follows:

$$\begin{aligned} &r_{Cij} \sim \text{Binomial}(p_{Cij}, n_{Cij}) \text{ and} \\ &r_{Tij} \sim \text{Binomial}(p_{Tij}, n_{Tij}) \end{aligned} \qquad \text{(eq.5)}$$

$$\log \text{odds}(p_{Cij}) = \gamma_{ij} \text{ and}$$
$$\log \text{odds}(p_{Tij}) = \gamma_{ij} + \psi_{ij} \qquad \text{(eq.6)}$$

$$\psi_{ij} \sim \text{Normal}(\theta_i + \sum_{m=1}^{M} \alpha_m (x_{mT_{ij}} - x_{mC_{ij}}), \sigma_i^2) \quad \text{(eq.7)}$$

$$\theta_i \sim \text{Normal}(\mu, \tau^2) \qquad \text{(eq.8)}$$

(i = 1 or 2 for the 2 study types;

j = 1,...,k_i studies, m = 1,..,M confounders).

It is assumed that the number of events in each arm in the jth study of type i (i.e., r_{Cij} and r_{Tij} for control (C) and treatment (T), respectively) follows a binomial distribution defined by the proportion of patients who experience the event in each arm in the jth study of type i (i.e., p_{Cij} and p_{Tij}) and the total number of patients in each arm in the jth study of type i (i.e., n_{Cij} and n_{Tij}), as shown in equation 5. Equation 6 describes the log odds for the event in the control (γ_{ij}) and treatment ($\gamma_{ij} + \psi_{ij}$) arms of each of the k_i studies.

This model assumes that the log odds ratio, ψ_{ij}, follows a normal distribution with a mean which is the sum of θ_i (i.e., the overall intervention effect in the ith type of studies) and a study specific bias adjustment, $\alpha_m(x_{mTij}-x_{mCij})$, that is proportional to the relative differences between the study arms in each of the studies (eq.7). In this expression, x_{mTij} and x_{mCij} are the values of the m-th potential confounder in each of the study arms (i.e., treatment and control) in the jth study of type i while α_m represents the mean bias for the m-th potential confounding variable, across all the studies. The remaining variables were defined as before.

Prior distributions for the unknown parameters were intended to be vague. Normal priors with mean zero and variance 0.26 truncated to be positive, were specified for both random-effects standard deviations (σ_i,τ). The priors for σ_i and τ corresponded to the priors used in Grines et al. [9] as they represented what may be considered reasonable priors in many situations [13]. These priors support equality between studies while discounting substantial heterogeneity. A Normal prior with mean zero and variance ten was used for the overall population effect (μ). Vague Normal priors with mean zero and variance 1000 were assigned to the log odds (γ_{ij}'s). These priors were applied to generate results both adjusted and unadjusted for potential confounders. In addition to these priors, the adjusted model also required priors for the bias coefficients (α_m) for each of the m-th potential confounders. These were also given vague Normal prior distributions with mean zero and variance 1000.

Alternative methods for potentially biased evidence

For comparison purposes, we also considered two approaches proposed to downweight the evidence from non-randomised studies. This is generally done by increasing the variance. The first method considered was the prior constraint used by Prevost et al. [5] to assess the influence of the assumption that the randomised studies were less biased than the non-randomised studies, and hence that $|\mu - \theta_1| < |\mu - \theta_2|$. This approach increased the relative proportion of the between-study-type variance (τ^2) associated with the non-randomised studies compared to the randomised studies. In so doing the interpretation of μ is altered. Since the constraint gives more weight to the randomised studies, μ no longer represents the total population studied. The overall effects in the randomised and non-randomised studies are represented by θ_1 and θ_2, respectively. The second approach was the informative prior distribution used by Sutton et al. [8] which included the evidence from the non-randomised studies via the prior for the treatment effect and combined this with a likelihood based only on the data from the randomised studies. Sutton et al. [8] centred their informative prior for the population mean on the non-randomised pooled estimate but used a variance four times larger than that of the randomised studies. The same approach was used for the current analysis, hence an informative Normal(-0.5619,0.8179) prior distribution was specified for μ. The same prior distributions as previously specified were used for the other unknown parameters.

Analyses

All of the analyses were conducted using MCMC simulation implemented in WinBUGS 1.4.3 software [17]. A ‘burn-in’ of 100 000 iterations was followed by a further 100 000 iterations during which the generated parameter values were monitored and summary statistics such as the median and 95% credible interval of the complete samples were obtained. History plots, autocorrelation plots, and various diagnostics available in the package Bayesian Output Analysis [18], performed on two chains, were used to assess convergence. See additional file 1: Appendix for WinBUGS codes. The data are available from the author upon request.

Illustration

Data

Data from a previously published systematic literature review evaluating EVAR and OSR in the treatment of AAAs [12] were used to illustrate the impact of adjusting for imbalances between study arms when combining evidence from randomised and non-randomised studies. The review identified 79 comparative studies of which four were randomised and 75 were non-randomised. One of the primary outcomes was 30-day mortality reported as an odds ratio.

Evidence of the relative imbalances within the randomised and non-randomised studies, together with information on the predictors of perioperative mortality in patients undergoing OSR, from several risk scoring methods (e.g., Leiden score) [19], were used to inform the choice of covariates for adjustment in both the base case scenario and sensitivity analyses. No adjustment was made for imbalances in the original study [12]. The extent to which some covariate data were missing was also considered in an additional sensitivity analysis, in which values for the missing covariates were imputed.

Base case scenario

In the base case analysis, the results were adjusted for imbalances in age, gender, and cardiac disease. For all three covariates imbalances were greater among the non-randomised studies. The three covariates were available in a total of 44 studies, four randomised and 40 non-randomised. A description of the data is given in Table 1.

Sensitivity analyses

Priors

A sensitivity analysis was conducted to assess the impact of using different prior distributions for the between-study (σ_i) and between-study-type (τ) standard deviations. The vague priors used in the base case analysis (σ_i, τ ~ half-normal $(0,0.51^2)$) were compared to the more informative yet "fairly unrestrictive" priors used by Prevost et al. [5] (σ_i ~ half-normal$(0,0.36^2)$, τ ~ half-normal$(0,0.18^2)$) and to a set of less informative priors. The latter involved Normal truncated to be positive priors with mean zero and variance one for the between-study standard deviation for the randomised studies (σ_1) and the between-study-type standard deviation (τ). A Uniform prior over the range (0,10) was specified for the between-study standard deviation for the non-randomised studies (σ_2). The prior distributions for the other unknown parameters remained unchanged from the base case analysis.

Imputation for missing data

A second sensitivity analysis was conducted to use all the studies providing comparative information (i.e., 79 studies including four randomised) rather than a subset of studies (i.e., 44 studies including four randomised) and to adjust for additional covariates which could affect the 30-day mortality risk. Among the other risk factors used to predict mortality following AAA surgery, the Leiden and modified Leiden scores both included pulmonary and renal disease [19]. These may be particularly relevant in the current context, as imbalances in pulmonary and renal disease were found to be greater among the randomised studies than among the non-randomised studies [12].

Since all five covariates were present together in less than one third of all studies (i.e., two randomised and 23 non-randomised studies), missing covariate values were imputed. Multiple imputation was conducted using R 2.9.2 software [20] assuming that the covariates were missing completely at random.

This approach implemented the bootstrap method to first impute values for each missing variable by randomly selecting from the observed outcomes of that variable and then generated multiple imputations (three datasets) using iterative regression imputation, looping through until approximate convergence. The data are described in Table 1. The result was a single imputed dataset of 79 studies (four randomised and 75 non-randomised) which was then analysed, in WinBUGS, adjusting for imbalances in age, gender, cardiac disease, pulmonary disease, and renal disease. Results were generated using all three types of priors described in the sensitivity analysis.

Results

Base case scenario

Unadjusted for potential confounders

The four randomised and 40 non-randomised studies were first analysed separately without adjusting for differences in study arms using a standard Bayesian two-level hierarchical model [13] together with a Normal(0,10) prior distribution for the population mean and a Normal(0,0.26) truncated to be positive prior distribution for between-study standard deviation. This produced estimates of the pooled median odds ratio for the randomised studies alone of 0.32 (95% credible interval (CrI) 0.13,0.76) and for the non-randomised studies alone of 0.57 (95% CrI 0.41,0.82).

In comparison, the Bayesian three-level hierarchical model estimated the pooled median odds ratio for the randomised studies to be 0.43 (95% CrI 0.19,0.75) and for the non-randomised studies to be 0.54 (95% CrI 0.40,0.76). When randomised and non-randomised evidence was combined, the overall median odds ratio from the three-level model was 0.49 (95% CrI 0.21, 0.98). This comparison illustrates the effect of the three-level hierarchical model allowing the cross contribution of evidence between the randomised and non-randomised studies. As a result, the estimated odds ratios for the study types were drawn towards one another and the uncertainty associated with them was reduced. The relative discrepancy in the number of randomised and non-randomised studies resulted in the pooled estimate for the randomised studies being greatly influenced by the non-randomised studies' estimate. The odds ratio in the non-randomised studies however, was drawn in by a much smaller amount.

Adjusted for differences in age, gender and cardiac disease between study arms

Upon synthesising the randomised and non-randomised evidence, the three-level hierarchical model adjusting for imbalances between study arms in terms of age, gender and cardiac disease (eq.7) was applied to the data. Important differences were observed compared to the unadjusted analysis. Posterior median odds ratios were 0.35 (95% CrI 0.17,0.63) for the randomised studies and 0.39 (95% CrI 0.25,0.61) for the non-randomised studies. The overall estimated odds ratio was 0.37 (95% CrI 0.17,0.77).

'Naive' adjustments made using the mean age, proportion of males and proportion of patients with cardiac disease in each study, as in Prevost and Sampath [5,10], produced estimates of 0.57 (95% CrI 0.27,1.03) for the randomised studies and 0.62 (95% CrI 0.44,0.87) for the non-randomised studies. The estimated population odds ratio was 0.60 (95% CrI 0.28,1.20).

Alternative methods for potentially biased evidence

The prior constraint resulted in estimated posterior median odds ratios of 0.44 (95% CrI 0.20,0.76) and 0.54 (95% CrI 0.40,0.76), respectively for the randomised and non-randomised studies and an overall estimate of 0.43 (95% CrI 0.18,0.89). An informed prior distribution centred on the pooled estimate from the analysis of the non-randomised studies alone with a variance four times that of the randomised studies generated a single overall estimate of 0.35 (95% CrI 0.16,0.76).

Figure 1 compares the estimated odds ratios obtained from separate analyses of each type of study design using a two-level Bayesian hierarchical model with a three-level Bayesian hierarchical model synthesising evidence from both types of designs. In addition the estimates obtained when adjusting for differences in age, gender and cardiac disease between study arms or using aggregate study values are also presented. Estimates resulting from approaches downweighting the non-randomised evidence are displayed as well. All odds ratios are described in terms of the numerically approximated (via MCMC) median value of their posterior distribution and the associated 95% Bayesian CrI.

Sensitivity analysis

Priors

As shown in Table 2, all three sets of priors produced similar values for the study type effects θ_1 (randomised), θ_2 (non-randomised) and for the overall odds ratio μ, though the precision of the credible intervals varied. Our vaguest priors produced an overall estimate which was not statistically significant.

Imputed dataset

Adjustment for imbalances in pulmonary and renal disease in addition to age, gender and cardiac disease increased the estimated posterior median odds ratios for each of the study types and for the overall estimated odds ratio, though the inferences remained the same (Table 2).

Discussion

We expanded the methods initially proposed by Prevost et al. [5] to take into account differences in patient characteristics between study arms. Comparison of the estimated odds ratios between the unadjusted three-level model, dominated by the 40 non-randomised studies, and the model adjusted using study arm differences revealed an overall odds ratio that had moved closer to the pooled estimate from the four randomised studies alone. The estimate was more precise than the randomised studies' estimate, reflecting the additional information from the adjusted non-randomised studies. 'Naive' adjustments made using aggregate values in each study (centred about their respective mean values across all the studies) resulted in estimated odds ratios that were relatively closer to the pooled estimate from the non-randomised studies alone. The prior constraint proposed by Prevost et al. [5] did not alter the type level estimates to any noticeable extent. It did however change the contribution that each made towards the population level estimate. Relative to the unadjusted model, the introduction of the constraint resulted in a population level estimate which had moved towards the randomised studies' estimate both in terms of its location and precision. However, the shift and the precision of the credible interval were both less than when the model was adjusted for study arm differences. Sutton et al.'s [8] informative prior approach resulted in an overall odds ratio that was slightly closer to the randomised estimate than the model adjusted for imbalances. Its estimate was also slightly more precise.

All of the methods, with the exception of the model using aggregate study values for adjustment, produced population level estimates that had moved towards the randomised studies' estimate. While this lends credence to the ability of the extended model to adjust for potential confounders, this new model, in its current form, has some potential limitations. Because the imbalanced studies are adjusted, but not downweighted the credible intervals do not reflect the uncertainty due to this source of bias [15]. While downweighting itself may not eliminate bias, in conjunction with adjustment, it would give the biased studies less weight in the analysis. Ideally, this would be achieved by inflating the variances in such a way that, like the study specific bias adjustments, the downweighting was proportional to the relative differences between the study arms. Also, in its current form the proposed model does not address the extent to which variation in age, gender, and cardiac disease across studies may explain variation in study estimates. Rather, the objective of this study was to propose a method to adjust for differences in patient characteristics within studies, as a way of controlling for potential confounders. A practical limitation, as evidenced by this example, is the availability of arm level data from the primary papers. Any analysis could only be based on a subset of studies for which information on potential confounding variables happened to be available. This could bias the results if the observations were not missing at random [21]. Assuming that the covariates were missing completely at random the current analysis attempted to impute the missing values, though admittedly the two-stage nature of the current approach may appear inelegant (i.e., using R to impute the data and then analysing the new data in WinBUGS). A more natural solution would be to include the unobserved covariate values along with the unobserved parameters inside the MCMC, although this may add an additional layer of complexity. Due to the focus of the paper being Bayesian hierarchical models for combining randomised and non-randomised studies rather than methods to impute missing data, and for convenience, we decided to generate the missing values using R. Finally, adjustment cannot address the problem of unknown potential confounders [21].

Despite these limitations, we believe that the approach presented in this paper provides a systematic way of incorporating potentially biased evidence, relying on bias adjustment rather than arbitrarily downweighting the evidence. Prevost’s and Sutton’s approaches to downweighting assume the evidence from non-randomised studies is uniformly more biased, which, if there are well balanced non-randomised studies, may not necessarily be the case. Future research would be required to assess the generalisability of the proposed model beyond this single applied example. In particular, simulation studies would be necessary to ascertain its broader applicability. Part of the justification for combining evidence from both randomised and non-randomised studies rests on an all available evidence approach to health care decision making. The extent of missing covariate data in the current example suggests authors should be encouraged to better report the main characteristics of their study populations. The current example also illustrates the impact of different prior distributions on the precision of the results. The choice of prior could have implications in terms of informing health care decision making and may be particularly important in situations in which the data are not very informative [22].

Conclusion

Synthesising evidence from both randomised and non-randomised studies requires methods for incorporating potential biases. In this paper, we propose a new approach to deal with bias due to confounding when combining randomised and non-randomised studies. This approach uses differences in patient characteristics to adjust for imbalances between study arms. Including aggregate study values for patient level covariates does not account for imbalances and downweighting may not eliminate bias. Within the context of a Bayesian hierarchical model the proposed approach could facilitate the use of all available evidence to inform health policy decisions.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CEM conceived of the study, developed the model, analysed and interpreted the data, and drafted the manuscript. EMP conceived of the study, helped with statistical analysis, and critically reviewed the manuscript. LT conceived of the study, helped with statistical analysis, and critically reviewed the manuscript. RG conceived of the study, and critically reviewed the manuscript. JET conceived of the study, acquired the data, and helped with interpretation and drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

CEM is funded through a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada. JET holds a 2007 Career Scientist Award, Ontario Ministry of Health and Long-Term Care.

References

1. **International Network of Agencies for Health Technology Assessment** [http://www.inahta.org/HTA]
2. Centre for Reviews and Dissemination: *Systematic Reviews: CRD's guidance for undertaking reviews in health care*. York: University of York; 2009.
3. Gordis L: *Epidemiology*. Philadelphia: Elsevier Inc.; 2004.

4. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, Lu G: **Bayesian methods for evidence synthesis in cost-effectiveness analysis**. *Pharmacoeconomics* 2006, **24**(1):1-19.
5. Prevost TC, Abrams KR, Jones DR: **Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening**. *Stat Med* 2000, **19**:3359-3376.
6. Sculpher MJ, Claxton K, Drummond M, McCabe C: **Whither trial-based economic evaluation for health care decision making**. *Health Econ* 2006, **15**:677-687.
7. Sutton AJ, Cooper NJ, Jones DR: **Evidence synthesis as the key to more coherent and efficient research**. *BMC Med Res Methodol* 2009, **9**:29.
8. Sutton AJ, Abrams KR: **Bayesian methods in meta-analysis and evidence synthesis**. *Stat Methods Med Res* 2001, **10**(4): 277-303.
9. Grines CL, Nelson TR, Safian RD, Hanzel G, Goldstein JA, Dixon S: **A Bayesian meta-analysis comparing AngioJet thrombectomy to percutaneous coronary intervention alone in acute myocardial infarction**. *J Interv Cardiol* 2008, **21**:459-482.
10. Sampath S, Moran JL, Graham PL, Rockliff S, Bersten AD, Abrams KR: **The efficacy of loop diuretics in acute renal failure: assessment using Bayesian evidence synthesis techniques**. *Crit Care Med* 2007, **35**(11): 2516-2524.

11. Eddy DM, Hasselblad V, Shachter R: **An introduction to a Bayesian method for meta-analysis: the confidence profile method**. *Med Decis Making* 1990, **10**:15-23.
12. Hopkins R, Bowen J, Campbell K, Blackhouse G, De Rose G, Novick T, O'Reilly D, Goeree R, Tarride JE: **Effects of study design and trends for EVAR versus OSR**. *Vasc Health Risk Manag* 2008, **4**(5): 1011-1022.
13. Spiegelhalter DJ, Abrams KR, Myles JP: *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, West Sussex: John Wiley & Sons Ltd; 2004.
14. Ades AE, Sutton AJ: **Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches**. *J R Stat Soc Ser A* 2006, **169**:5-35.
15. Greenland S: **Multiple-bias modelling for analysis of observational data**. *J R Stat Soc Ser A* 2005, **168**(Part 2):267-306.
16. Rothman KJ, Greenland S, Lash TL: *Modern Epidemiology Third Edition*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
17. Lunn DJ, Thomas A, Best N, Spiegelhalter D: **WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility**. *Stat Comput* 2000, **10**:325-337.
18. **Bayesian Output Analysis Program (BOA) Version 1.1 User's Manual 2005** [www.public-health.uiowa.edu/boa/BOA.pdf]

19. Nesi F, Leo E, Biancari F, Bartolucci R, Rainio P, Satta J, Rabitti G, Juvonen T: **Preoperative risk stratification in patients undergoing elective infrarenal aortic aneurysm surgery: evaluation of five risk scoring methods**. *Eur J Vasc Endovasc Surg* 2004, **28**:52-58.
20. **The R Project for Statistical Computing** [http://www.r-project.org/]
21. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG: **Evaluating non-randomised intervention studies**. *Health Technol Assess* 2003, **7**(27):1-173.
22. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR: **How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS**. *Stat Med* 2005, **24**:2401-2428.

Figure 1 - Estimated overall (μ) and study type (θ_1, θ_2) odds ratios from Bayesian hierarchical models

Perioperative mortality in studies of EVAR and OSR for the treatment of abdominal aortic aneurysms (four randomised controlled trials and 40 non-randomised studies)

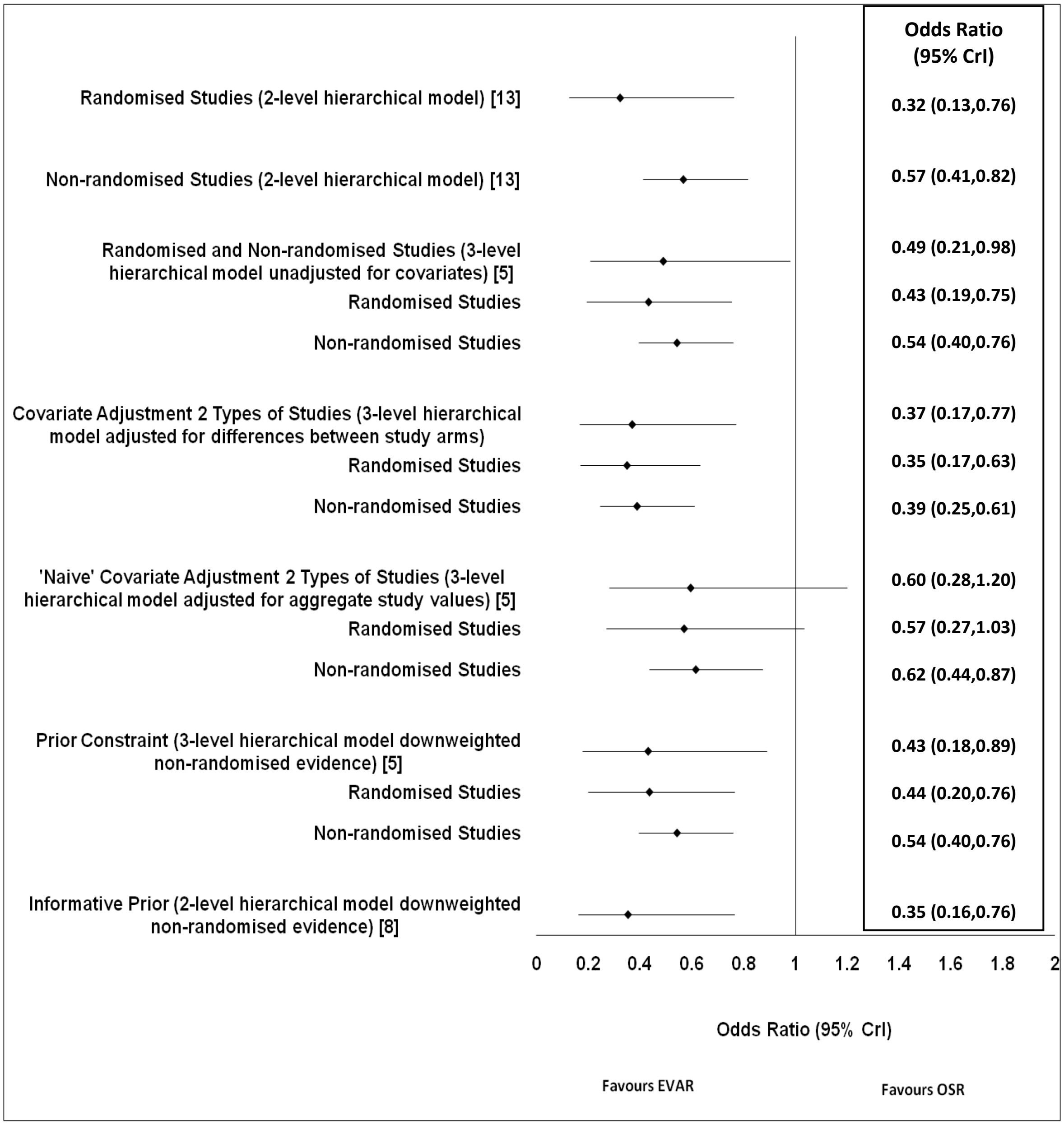


Table 1. Covariate Data: Average Imbalance between Study Arms

	Base Case	**Imputed data**
	3 Covariates[a]	**5 Covariates**[b]
	(k=4 randomised and 40 non-randomised)	**(k=4 randomised and 75 non-randomised)**
Study Type		
Non-randomised	**Average Difference (EVAR-OSR)**	**Average Difference (EVAR-OSR)**
Male (proportion)	0.09	0.10
Age (years)	2.40	2.53
Cardiac disease (proportion)	0.12	0.14
Pulmonary disease (proportion)	not considered as missing in 43% of the 75 non-randomised studies	0.10
Renal disease (proportion)	not considered as missing in 54% of the 75 non-randomised studies	0.05
Randomised		
Male (proportion)	0.05	0.05
Age (years)	0.82	0.82
Cardiac disease (proportion)	0.05	0.05
Pulmonary disease (proportion)	not considered as missing in 25% of the 4 randomised studies	0.13
Renal disease (proportion)	not considered as missing in 50% of the 4 randomised studies	0.07

a.male, age, cardiac disease, b.male, age, cardiac disease, pulmonary disease, renal disease

Table 2. Adjustment for Differences in Patient Characteristics between Study Arms: Sensitivity to Prior Distributions

		Type of Prior		
Dataset	**Posterior Estimate** Median OR (95% credible interval)	**Base Case Analysis: "Reasonably Vague" (Grines)**	**Sensitivity Analysis: "Fairly Unrestrictive" (Prevost)**	**Sensitivity Analysis: "Vaguest"**
Base Case **3 Covariates[a] (k=44)**	**Overall (μ)**	0.37 (0.17,0.77)	0.37 (0.23,0.60)	0.37 (0.18,1.25)
	Randomised (θ_1)	0.35 (0.17,0.63)	0.36 (0.21,0.59)	0.34 (0.13,0.74)
	Non-Randomised (θ_2)	0.39 (0.25,0.61)	0.38 (0.25,0.57)	0.40 (0.23,0.68)
Imputed **5 Covariates[b] (k=79)**	**Overall (μ)**	0.45 (0.20,0.95)	0.47 (0.28,0.74)	0.44 (0.13,1.31)
	Randomised (θ_1)	0.42 (0.18,0.78)	0.46 (0.25,0.73)	0.39 (0.14,0.87)
	Non-Randomised (θ_2)	0.49 (0.33,0.72)	0.49 (0.33,0.71)	0.49 (0.32,0.74)

a.male, age, cardiac disease, b.male, age, cardiac disease, pulmonary disease, renal disease

Appendix to Chapter 2: WinBUGS Code

```
#Model
model{
for (j in 1:k){
#Likelihood for within-type model
rOSR[j] ~ dbin(pOSR[j],nOSR[j])
rEVAR[j] ~ dbin(pEVAR[j],nEVAR[j])
logit(pEVAR[j]) <- gamma[j] + psi[j]
logit(pOSR[j]) <- gamma[j]
gamma[j] ~ dnorm(0,0.001)
#Covariate Adjustment
psi[j] <- theta[type[j]] + alpha1*(cadEVAR[j]-cadOSR[j]) + alpha2*(maleEVAR[j]-maleOSR[j]) + alpha3*
(ageEVAR[j]-ageOSR[j]) + (sigma[type[j]]*z[j])        #Differences between study arms
psi[j] <- theta[type[j]] + alpha1*(cad[j]-cad.bar) + alpha2*(male[j]-male.bar) + alpha3* (age[j]-age.bar) +
(sigma[type[j]]*z[j])                                  #Aggregate study values
z[j] ~ dnorm(0,1)}
#Likelihood for between-type model
for (i in 1:2){
theta[i] <- mu + (tau*epsilon[i])
#Prior for base case
sigma[i] ~ dnorm(0,4)I(0,)                             #Reasonably vague
#Priors for sensitivity analysis
sigma[i] ~ dnorm(0,8)I(0,)                             #Fairly unrestrictive
sigma[1] ~ dnorm(0,1)I(0,)                             #Vaguest
sigma[2] ~ dunif(0,10)
epsilon[i] ~ dnorm(0,1)
OR.type[i] <- exp(theta[i])}
#Prior constraint
for (i in 1:1){
low.epsilon[1] <- min(epsilon[2],-epsilon[2])
up.epsilon[1] <- max(epsilon[2],-epsilon[2])
epsilon[1] ~ dnorm(0,1)I(low.epsilon[1],up.epsilon[1])}
for (i in 2:2){
low.epsilon[2] <- max(epsilon[1],-epsilon[1])
mod.epsilon2 ~ dnorm(0,1)I(low.epsilon[2],)
sign ~ dbern(0.5)
epsilon[2] <- (mod.epsilon2*sign) - (mod.epsilon2*(1-sign))}
alpha1 ~ dnorm(0,0.001)
alpha2 ~ dnorm(0,0.001)
alpha3 ~ dnorm(0,0.001)
cad.bar <- mean(cad[])
age.bar <- mean(age[])
male.bar <- mean(male[])
mu ~ dnorm(0,0.1)
#Informative prior
mu ~ dnorm(-0.5619,1.2226)
#Prior for base case
tau ~ dnorm(0,4)I(0,)                                  #Reasonably vague
#Priors for sensitivity analysis
tau ~ dnorm(0,30)I(0,)                                 #Fairly unrestrictive
tau ~ dnorm(0,1)I(0,)                                  #Vaguest
OR.overall <- exp(mu)}
```

CHAPTER 3

Bayesian Hierarchical Models Combining Different Study Types and Adjusting for Covariate Imbalances: A Simulation Study to Assess Model Performance

C Elizabeth McCarron[a,b,§], Eleanor M Pullenayegum[a,c], Lehana Thabane[a,c], Ron Goeree[a,b], Jean-Eric Tarride[a,b]

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

[b]Programs for Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[c]Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

§ Corresponding author

Citation: McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *Accepted September 8, 2011 for publication in PLoS ONE*.

Abstract

Background: Bayesian hierarchical models have been proposed to combine evidence from different types of study designs. However, when combining evidence from randomised and non-randomised controlled studies, imbalances in patient characteristics between study arms may bias the results. The objective of this study was to assess the performance of a proposed Bayesian approach to adjust for imbalances in patient level covariates when combining evidence from both types of study designs.

Methodology/Principal Findings: Simulation techniques, in which the truth is known, were used to generate sets of data for randomised and non-randomised studies. Covariate imbalances between study arms were introduced in the non-randomised studies. The performance of the Bayesian hierarchical model adjusted for imbalances was assessed in terms of bias. The data were also modelled using three other Bayesian approaches for synthesising evidence from randomised and non-randomised studies. The simulations considered six scenarios aimed at assessing the sensitivity of the results to changes in the impact of the imbalances and the relative number and size of studies of each type. For all six scenarios considered, the Bayesian hierarchical model adjusted for differences within studies gave results that were unbiased and closest to the true value compared to the other models.

Conclusions/Significance: Where informed health care decision making requires the synthesis of evidence from randomised and non-randomised study designs, the proposed hierarchical Bayesian method adjusted for differences in patient characteristics between

study arms may facilitate the optimal use of all available evidence leading to unbiased results compared to unadjusted analyses.

Introduction

Evidence of the effects of interventions is a critical component of health care decision making as it contributes to the comparison of alternative interventions in terms of their relative costs and effects. Such comparisons form the basis of decisions regarding the economically efficient allocation of scarce resources. An all available evidence approach to informing these decisions may require the synthesis of evidence from different types of study designs (e.g., randomised controlled trials (RCTs) and comparative non-randomised or observational studies). Recently, Bayesian hierarchical models have been proposed to combine evidence from different types of study designs such as randomised and non-randomised studies [1,2].

Due to their inherent design, RCTs are more likely to be balanced in terms of patient characteristics between study arms than non-randomised studies, but they are subject to strict inclusion and exclusion criteria which may limit their generalisability. Despite the greater generalisability associated with non-randomised or observational studies, the increased likelihood of imbalances among the study arms compared to RCTs suggests the results may be more subject to the potential confounding effects of extraneous variables. Although other sources of bias, both internal (e.g., performance, attrition) and external (e.g., population, intervention) [3], may exist, it is the increased likelihood of imbalances among the non-randomised studies that constitutes the principal difference between randomised and non-randomised studies [4]. When these imbalances exist in factors that are also related to the outcome, bias may be introduced.

In order to address the problem of bias due to imbalances between study arms in non-randomised studies, we proposed [5] an extension to the Bayesian three-level hierarchical model, initially developed by Prevost et al. [1], and applied it to a case study. The proposed approach involved adjusting study estimates for potential imbalances using differences in patient characteristics between study arms. Results from the case study indicated a shift in the estimates for the model adjusted for differences towards the estimate for the randomised studies alone [5]. While this shift lends credence to the proposed model's ability to adjust for imbalances, these results pertain only to a single applied example.

Given the importance of using all available evidence for decision making and the increased use of Bayesian hierarchical models to combine evidence from different study types [6,7], the objective of this paper was to assess the performance of the proposed Bayesian approach to synthesise evidence from randomised and non-randomised studies and adjust for imbalances in patient characteristics within studies. To meet the study objective, we conducted a simulation study to generate sets of randomised and non-randomised studies in which bias that could be explained by covariate imbalances was introduced in the non-randomised studies. The data were also modelled using three other

Bayesian approaches: 1) results unadjusted for potential imbalances [1], 2) results adjusted for aggregate study values (e.g., mean age) [1] and 3) downweighting the potentially biased non-randomised studies [2]. The sensitivity of the results to changes in the impact of the imbalances and the relative number and size of studies of each type was also assessed.

Methods

The following presents the four models being compared, the scenarios considered, and the methods used to conduct the simulation study.

2.1 Bayesian methods to combine evidence from randomised and non-randomised studies

2.1.1 Unadjusted for potential imbalances (model I)

The first model presented is the Bayesian three-level hierarchical model unadjusted for potential imbalances. We undertook this analysis using a binomial model in which the odds of the event were calculated for each study and study arm level information was incorporated into the model. We assumed that for each study type (indexed by i) there were k_i studies (indexed by j), which allows for a different number of studies for each study type (i.e., randomised and non-randomised).

The model can be written as follows:

$$r_{Cij} \sim \text{Binomial}(p_{Cij}, n_{Cij}) \text{ and } r_{Tij} \sim \text{Binomial}(p_{Tij}, n_{Tij}) \qquad \text{(eq. 1)}$$

$$\text{log odds}(p_{Cij}) = \gamma_{ij} \text{ and } \text{log odds}(p_{Tij}) = \gamma_{ij} + \psi_{ij} \qquad \text{(eq. 2)}$$

$$\psi_{ij} \sim \text{Normal}(\theta_i, \sigma_i^2) \qquad \text{(eq. 3)}$$

$$\theta_i \sim \text{Normal}(\mu, \tau^2) \qquad \text{(eq. 4)}$$

(i = 1 or 2 for the 2 study types; j = 1,...,k_i studies).

At the first level of the model, represented by equations one and two, it was assumed that the number of events in each arm in the jth study of type i (i.e., r_{Cij} and r_{Tij} for control (C) and treatment (T), respectively) followed a binomial distribution defined by the proportion of patients who experienced the event in each arm in the jth study of type i (i.e., p_{Cij} and p_{Tij}) and the total number of patients in each arm in the jth study of type i (i.e., n_{Cij} and n_{Tij}). Equation two described the log odds for the event in the control (γ_{ij}) and treatment ($\gamma_{ij} + \psi_{ij}$) arms of each of the k_i studies.

The second level of the model, represented by equation three, assumed that the log odds ratio comparing treatment and control, ψ_{ij}, followed a normal distribution with a mean of θ_i (i.e., the overall intervention effect in the ith type of studies). The within-study-type variability for studies of type i was represented by σ_i^2. At the third level of the model, represented by equation four, the study-type effects were distributed about an overall population effect, μ, with τ^2 representing the between-study-type variability.

Prior distributions for the unknown model parameters were intended to be vague. Normal priors with mean zero and variance 0.26 truncated to be positive, were specified for the random-effects standard deviations (σ_i,τ). These priors support equality between studies while discounting substantial heterogeneity and represent what may be considered reasonable priors in many situations [8]. In keeping with Prevost et al. [1], Normal priors with mean zero and variance ten were used for the overall population effect (μ). Vague Normal priors with mean zero and variance 1000 were assigned to the log odds (γ_{ij}'s).

2.1.2 Adjustment using study arm differences (model II)

The following presents the extension of the Bayesian three-level hierarchical model (I) proposed by McCarron et al. [5]. The model was specified as before except equation three was replaced by equation five.

$$\psi_{ij} \sim \text{Normal}(\theta_i + \sum_{m=1}^{M} \alpha_m (x_{mT_{ij}} - x_{mC_{ij}}), \sigma_i^2) \qquad \text{(eq. 5)}$$

(i = 1 or 2 for the 2 study types; j = 1,...,k_i studies; m = 1,..,M confounders).

As shown in equation five, this model assumed that the log odds ratio, ψ_{ij}, followed a normal distribution with a mean which was the sum of θ_i (i.e., the overall intervention effect in the ith type of studies) and a study specific bias adjustment, $\sum_{m=1}^{M} \alpha_m (x_{mT_{ij}} - x_{mC_{ij}})$, that was proportional to the relative differences between the study arms in each of the studies. In this expression, x_{mTij} and x_{mCij} were the values of the m-th potential confounder in each of the study arms (i.e., treatment and control) in the jth study of type i while α_m represented the coefficient for the m-th potential confounding variable across all the studies. This variable described the impact of the imbalances on the study specific log odds ratios.

The priors for the unknown parameters were the same as for model I with the addition of a vague Normal prior with mean zero and variance 1000 for the coefficient (α_m) for the m-th potential confounder.

2.1.3 Adjustment using aggregate study values (model III)

This approach was initially proposed by Prevost et al. [1] to try to explain between study heterogeneity. The model was specified in the same way as in section 2.1.1, except equation three was replaced by equation six:

$$\psi_{ij} \sim \text{Normal}(\theta_i + (\alpha_m \times x_{mij}), \sigma_i^2) \qquad \text{(eq. 6)}$$

(i = 1 or 2 for the 2 study types; j = 1,...,k_i studies; m = 1,..,M confounders).

In this approach, x_{mij} represented the value of the m-th potential confounder aggregated across study arms (i.e., treatment and control) in the jth study of type i. This is in contrast to the previous approach which adjusted using the difference in the m-th potential confounder between the study arms. The prior distributions were the same as in the previous models.

2.1.4 Downweighting using an informative prior (model IV)

The informative prior approach used by Sutton and Abrams [2] included the evidence from the non-randomised studies via the prior for the treatment effect and combined this with a likelihood based only on the data from the randomised studies.

As in Sutton and Abrams [2], we centred the informative prior for the population mean (μ) on the non-randomised pooled estimate but used a variance four times larger than that of the RCTs. Such a prior reflects scepticism regarding the non-randomised evidence and would be appropriate in situations where a researcher believes that although the non-randomised evidence provides some information, concern that serious biases may exist (e.g., as a result of imbalances in study arms) means that it should be treated with caution. The pooled estimate for the non-randomised studies was generated using a two-level Bayesian hierarchical model (simple Bayesian random-effects model). We chose to use a variance that was four times as large as that for the RCTs, because this was the variance inflation factor used by Sutton and Abrams [2]. Other choices are possible, however. The more the variance from the non-randomised studies is inflated, the more their evidence is downweighted.

2.2 Assessment framework

The effect of these models is to produce a weighted average of the evidence from the randomised and non-randomised studies, where the weights are determined either implicitly, as in the Bayesian three-level hierarchical models (I,II and III), or explicitly, as in the informative prior approach (model IV) [9]. The results for each of the four models were simulated under different scenarios which varied as a function of the impact of the imbalances in the non-randomised studies, and the relative number and size of studies of each type. These factors were selected as they were deemed to be the most important in terms of calculating a weighted average of the evidence from the randomised and non-randomised studies. For the purpose of this simulation study, imbalance in a single continuous covariate (i.e., age) was considered, but the analysis could be extended to adjust for other covariates [5]. Imbalances in age between study arms were only assumed for the non-randomised studies, in keeping with the general assumption that due to their design RCTs are more likely to be balanced.

Table 1 presents the parameters used in the six scenarios considered. Two different values were investigated for the impact of the imbalances in the non-randomised studies (α_m). Log scale values of 0.10 and 0.50 were chosen as they represent lower and upper estimates of what may appear reasonable in terms of variation in the between-study log odds ratios [8]. A magnitude of 0.10 would indicate that there is not much systematic variation in the study specific log odds ratios while a magnitude of 0.50 would result in much more systematic variation. This means that, all else being equal, every one unit increase in the difference in age between study arms would result in an increase in the study specific log odds ratio of 0.10 or 0.50. For example, the impact of going from no imbalances to a one year difference in mean age between study arms would increase the study specific log odds ratio from a true value of -0.20 to values of -0.10 and 0.30 respectively.

The impact of the precision and quantity of information contained in each of the two types of studies (i.e., randomised and non-randomised) was examined by comparing study sizes of 100 to 500 patients per arm and 500 to 1000 patients per arm for the randomised and non-randomised studies respectively and four randomised studies with 40 non-randomised studies. These values reflect the fact that non-randomised studies tend to be larger than randomised studies [2]. Also, the number of studies in a meta-analysis of RCTs in medicine tends to be small and it is common to see meta-analysis performed on five or fewer studies [10]. These values were also based on the systematic literature review comparing endovascular aneurysm repair (EVAR) and open surgical repair (OSR) [11] that informed the results of the previous case study [5]. For the six scenarios presented in Table 1, it was assumed that the true log odds ratio was -0.20, which corresponds roughly to an odds ratio of 0.82. Although this represented a much more modest treatment effect than was observed for 30-day mortality in either the randomised or non-randomised studies in the EVAR case study [5], this odds ratio may better reflect the magnitude of relative treatment effects seen in practice for other conditions.

2.3 Simulation study

As the truth is known, simulation studies allow one to assess model performance relative to this known truth [12]. This is in contrast to a case study, like the one in which we initially proposed model II, where the truth is not known. In order to demonstrate empirically whether model II is able to adjust for imbalances we have conducted a simulation study. The simulation study was concerned with synthesising evidence from randomised and non-randomised studies and adjusting for bias due to imbalances in the non-randomised studies, consequently we have simulated data under a model that produces imbalances in the non-randomised studies (see supporting Figure S1).

Each simulated data set consisted of a number of hypothetical randomised (i.e., four) and non-randomised studies (i.e., four or 40) comparing treatment and control groups. The outcome was defined as a dichotomous variable indicating the occurrence or not of the event of interest (i.e., death). Each data set for each of the two study types was generated by the following model:

$$r_{Cij} \sim \text{Binomial}(p_{Cij}, n_{Cij}) \text{ and } r_{Tij} \sim \text{Binomial}(p_{Tij}, n_{Tij}) \qquad \text{(eq. 7)}$$

$$\text{log odds}(p_{Cij}) = \gamma_{ij} \text{ and } \text{log odds}(p_{Tij}) = \gamma_{ij} + \psi_{ij} \qquad \text{(eq. 8)}$$

$$\psi_{ij} = \theta_i + \alpha_{age}(x_{ageTij} - x_{ageCij}) \qquad \text{(eq. 9)}$$

The number of subjects in the control (n_{Cij}) and treatment (n_{Tij}) groups in the jth study of type i were assumed to be equal and were sampled from a uniform distribution of either 100 to 500 or 500 to 1000 patients. Based on the data for perioperative mortality from the previous systematic literature review [11] the probability for the event (i.e., death) in the control group (p_{Cij}) in each of the k_i studies was drawn from a beta distribution with mean 0.04 and variance 0.001. For scenarios 1- 6, the true log-odds ratio (θ_i) was -0.20 for both the randomised and non-randomised studies. A possible explanation for the effect of treatment on mortality in our simulation study was assumed to be differences in age between study arms (x_{ageTij} - x_{ageCij}), as shown in equation nine. Age is related to mortality and α_{age} addresses the relationship between differences in age and mortality. The variables x_{ageTij} and x_{ageCij} are both sampled from uniform distributions based on the age distribution observed in the previous systematic literature review (i.e., $x_{ageTij} \sim$ uniform(75,90), $x_{ageCij} \sim$ uniform(70,85)) [11]. As randomisation will likely minimize differences between study groups, x_{ageTij} and x_{ageCij} were assumed to be equal in the randomised studies. Simulated values were generated for the number of events and subjects as well as for the age in the control and treatment groups given the impact of the imbalances (α_{age}), the number of randomised and non-randomised studies, and the study size being considered.

In order to justify the number of simulations (i.e., 100), we calculated the difference in mean treatment effects for each of the models (I,III,IV) relative to the difference model (II) and compared these to the standard errors of the differences in treatment effects. This was repeated across 100 simulations for each of two seeds (starting values for the simulation). The results across both seeds suggested that 100 simulations were sufficient to average out the sampling variation. For scenario 1, for example, the differences in mean treatment effects relative to model II were 0.27 for model I, 0.28 for model III and 0.10 for model IV. The standard errors of the differences were 0.02, 0.03 and 0.02 respectively for the three comparisons. For the second seed the mean differences were 0.27, 0.28, and 0.09 respectively and the standard errors were approximately 0.02 across all three comparisons, illustrating that sampling variation was small compared to the size of the differences that were detected.

Markov chain Monte Carlo (MCMC) simulation using the Gibbs sampling technique was used to assess the models. The Brooks, Gelman & Rubin, Geweke and Heidelberger and Welch diagnostics available in the package Bayesian Output Analysis [13], performed on two chains, were used to assess convergence. To provide a sense of the convergence diagnostics we give the Brooks, Gelman & Rubin diagnostics for the overall log odds ratio (μ) for each of the models in scenario 4: the estimated values for the ratio of total variability to within-chain variability were approximately 1.01, 1, 1.01, and 1for models I through IV respectively, suggesting little between-chain variability. Based on these and the results from the other diagnostics, we decided to use a burn-in of 50 000 iterations for every model for each simulated data set except for the unadjusted and aggregate models in scenarios 2 and 5, which required a longer burn-in of 100 000 iterations to converge. After discarding the burn-in iterations, we sampled from a further 10 000 iterations with a thin rate of 20, for each of the two chains, such that summary statistics for the parameter values were based on thinned samples of 1000 iterations.

The simulated data sets were generated in R 2.9.2 [14]. The Bayesian hierarchical models (I,II,III,IV) were fitted to each generated data set in WinBUGS 1.4 [15] using the R 2.9.2 package R2WinBUGS. To validate the simulation model the mean value for α_{age} was calculated across all 100 simulations for model II and compared to the true value. The results for the six scenarios were 0.10, 0.10, 0.10, 0.52, 0.51, and 0.51 respectively. These correspond to true values of 0.10 for scenarios 1-3 and 0.50 for scenarios 4-6.

2.4 Criteria for assessing model performance

The median value of the overall log odds ratio (μ) was calculated for each simulated data set. The four different models for the six scenarios were then evaluated relative to the true value using the criterion of bias under repeated sampling. The estimated bias in the log odds ratio was defined as the mean value of the median log odds ratios across the simulated samples minus the true value [12]. As the results may be subject to sampling variation, we also reported the bias divided by its standard error, which is equal to the standard error of the mean of the median log odds ratios and would be expected to follow a standard normal distribution. If an estimation technique is unbiased, we would expect the observed bias divided by its standard error (Z-statistic) to lie between -1.96 and +1.96 ninety-five percent of the time. Formulas for the various calculations are given in supporting Figure S1.

Results

Table 2 shows the point estimates for the mean of the median log odds ratios, and the associated standard errors of the mean median log odds ratios as well as the estimated bias and Z-statistics for each of the four models in the six scenarios. As shown in this table, the estimates of the pooled effect size appear to be unbiased for the model adjusted for differences (model II) across all six scenarios. The informative prior approach appears to give less biased results than the model adjusted for aggregate study values while bias is roughly equal for both the model adjusted for aggregate values and the unadjusted model. An increase in the study arm size for the non-randomised studies relative to the randomised studies tends to increase the precision of the estimates for all of the models. However, combining evidence from four randomised studies and 40 non-randomised studies seems to increase the precision of the estimates the most compared to the other scenarios. In general, as might be expected, there is more variability in the model estimates when the assumed value of the impact of imbalances in age across all of the studies (α_{age}) is greater. The most extreme cases of bias appear to occur with the aggregate and unadjusted models in scenario five, when the value of α_{age} is 0.50 and there are four randomised and 40 non-randomised studies, and scenario two, when α_{age} is 0.10 and there are 40 non-randomised studies. However, as shown in table 2, the extent of the bias is more pronounced in scenario five compared to scenario two, where the magnitude of the impact of the imbalances is relatively smaller.

Figure 1 presents the point estimates and the confidence intervals for the overall log odds ratio (μ) for each of the models in the six scenarios. Comparing the point estimates to a log odds ratio of zero (i.e., no effect) indicates that among the aggregate and unadjusted models and even the informative prior, for scenarios 4-6, the impact of the imbalances is such that it alters the estimate as to whether or not the treatment is effective, thus deviating from the truth.

Discussion

This simulation study demonstrated that when bias in the non-randomised studies can be explained by covariate imbalances between study arms, the proposed Bayesian three-level hierarchical model adjusted for differences in patient characteristics within studies can handle this problem. Using simulation techniques, wherein the truth is known, we have been able to produce empirical evidence that this is the case. Failure to take into account these imbalances could bias the results.

Specifically, six scenarios incorporating different aspects of the impact of the imbalances and the relative numbers and sizes of each study type were considered. The results from the model adjusted for differences in patient characteristics within studies were, in every scenario, unbiased and closest to the true value compared to the results from the other models. This trend was robust to changes in the magnitude of the impact of the imbalances across studies as well as to both the relative number and size of studies being combined. Results also showed that none of the previously proposed Bayesian approaches could handle the issue of bias due to covariate imbalances. In certain instances, the bias observed among the other models was such that it changed the treatment estimate from one of benefit to one of harm. This could have implications in terms of health care decision making.

A practical limitation of the study concerns the number of simulations. There are no exact standards for the number of simulations necessary to average out sampling variation. We had initially considered performing 1000 simulations, but given the breadth of the study in terms of the number of scenarios considered and the associated run times, which ranged from five to 40 hours (1.83 GHz processor) for the 100 simulations, depending on the scenario, we determined that this would not be feasible. All of the parameters in each of the models were sampled and none were marginalised. This was done to ensure that the appropriate probabilistic dependence between the unknown parameters was propagated through the model. This could be particularly important when propagating inferences which are likely to be strongly correlated. For example, the current study considers both baseline levels and treatment differences estimated from the same studies [8]. In addition, the study is based on the assumption that there is also some association with imbalances in patient characteristics. As such it was important to sample all parameters in our simulation study. In other cases, perhaps, some parameters could be marginalised which could potentially improve the speed of the algorithm. Concerns regarding the number of simulations conducted were mitigated by comparing the effect sizes relative to the standard errors for each of two data sets. The results of these comparisons suggested we could be reasonably confident that the number of simulations was adequate.

Another potential limitation is that we assumed that the only source of variation between study estimates was due to imbalances between treatment arms in a single patient characteristic. As such, the underlying study type effects in both the randomised and non-randomised studies were assumed to be the same, which may not always be true. In practice, there may be other unexplained reasons why the estimates may differ. For example, patients enrolled in RCTs may be comparatively younger than those enrolled in non-randomised studies. The result of incorporating different values for the study type treatment effects is that there is no longer one true underlying effect, as there was in the six scenarios we considered. As the objective of the simulation study was to evaluate the performance of the proposed model in terms of adjusting for bias due to covariate imbalances, we did not address this issue in our study. Such an analysis would likely require a separate simulation study in which each scenario considered would involve its own base case assuming no imbalances. This would allow one to distinguish between the borrowing of strength across study types that is part of Bayesian hierarchical modelling and the appropriate adjustment for imbalances. This is left for future research. Future research could also assess the practical implications of these results within a decision analytic model. Another potential area of research could be the choice of prior distribution for the random-effects standard deviations (σ_i,τ). In contrast to the half-normal priors used in the current analysis, other suggestions include an inverse gamma distribution such as $1/\sigma_i^2 \sim$ Gamma[0.001,0.001]. Though, because such a distribution gives a high weight near zero for the standard deviation, the true variability may be underestimated [8,10]. As the current analysis relies on the existence of within-study-type and between-study-type variability, such a prior could be problematic, especially in those scenarios with only four non-randomised studies.

Despite potential limitations, we believe the results of this simulation study demonstrate the ability of the Bayesian three-level hierarchical model adjusted for differences to account for imbalances in patient characteristics within non-randomised studies that could bias the results. Such an approach does, however, rely on authors reporting the main characteristics of their study populations. This is important as the unadjusted model performed poorly in the presence of imbalances between study arms, as shown in our simulations. Unfortunately, few studies report all relevant covariates [4]. For example, in the initial case study, over half of the non-randomised studies were missing information on at least one covariate. Based on the results of our study, and the performance of the proposed approach, authors should be encouraged to improve the reporting of covariate information as this would facilitate adjustment for future evidence synthesis. The performance of the informative prior approach depends on how well one anticipates the impact of the imbalances on the results and downweights the evidence accordingly. Though the factor we used to inflate the variance and downweight the non-randomised studies was based on Sutton and Abrams [2], this value was somewhat arbitrary. In practice the selection of an appropriate discount factor would require a careful consideration of the relative weight and information each study type should contribute to the analysis. Nonetheless, the factor of four used for model IV in the current study means that in calculating a weighted average of both study types, the randomised studies would contribute the majority of the information. This reflects the existence of scepticism regarding the evidence generated by the non-randomised studies, but assumes there is still some value in combining these studies with the randomised studies. As has been demonstrated, by holding constant the amount by which the non-randomised studies were downweighted, downweighting is not an automatic procedure, nor does it explicitly address the potential for imbalances in patient characteristics within individual studies. Only one of the methods for downweighting used in the case study was considered in this analysis. The number of failures that occurred when simulating values for the prior constraint method [1] suggested that it could not be used reliably in the situations being investigated. However, the results of the case study suggest it is unlikely that this method would be able to handle the covariate imbalances, especially in those scenarios where the relative number or size of the non-randomised studies was greater compared to the randomised studies. Adjustment using aggregate study values attempts to explain heterogeneity across studies by adjusting for variation in study level characteristics. However, the absence of variation in mean age across studies does not preclude the presence of imbalances in age within studies. This will not be adjusted for using aggregate study values.

Based on the six scenarios considered, covariate adjustment using differences in patient characteristics between study arms (i.e., model II) provides a way of adjusting for imbalances that is robust to changes in the magnitude of the impact of the imbalances and the relative number and size of the studies of each type (i.e., randomised or non-randomised studies). This is important as this new methodology provides a way to synthesise randomised and non-randomised studies by adjusting for bias in non-randomised studies that is due to imbalances between treatment arms. Where informed health care decision making requires the synthesis of evidence from randomised and non-randomised study designs, such Bayesian hierarchical models adjusting for covariate imbalances could facilitate the optimal use of all available evidence.

References

1. Prevost TC, Abrams KR, Jones DR (2000) Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. Stat Med 19: 3359-76.

2. Sutton AJ, Abrams KR (2001) Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res 10(4): 277-303.

3. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG (2009) Bias modeling in evidence synthesis. J R Stat Soc Ser A 172: 21-47.

4. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, et al. (2003) Evaluating non-randomised intervention studies. Health Technol Assess **7**(27).

5. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE (2010) The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. BMC Med Res Methodol 10:64.

6. Sampath S, Moran JL, Graham PL, Rockliff S, Bersten AD, et al. (2007) The efficacy of loop diuretics in acute renal failure: assessment using Bayesian evidence synthesis techniques. Crit Care Med 35(11): 2516-24.

7. Grines CL, Nelson TR, Safian RD, Hanzel G, Goldstein JA, et al. (2008) A Bayesian meta-analysis comparing AngioJet thrombectomy to percutaneous coronary intervention alone in acute myocardial infarction. J Interv Cardiol 21:459-82.

8. Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, West Sussex: John Wiley & Sons Ltd. 391 p.

9. Ades AE, Sutton AJ (2006) Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. J R Stat Soc Ser A 169: 5-35.

10. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR (2005) How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 24: 2401-28.

11. Hopkins R, Bowen J, Campbell K, Blackhouse G, De Rose G, et al. (2008) Effects of study design and trends for EVAR versus OSR. Vasc Health Risk Manag 4(5): 1011-22.

12. Burton A, Altman DG, Royston P, Holder RL (2006) The design of simulation studies in medical statistics. Stat Med 25:4279-92.

13. Bayesian Output Analysis Program (BOA) Version 1.1 User's Manual 2005. Accessed at www.public-health.uiowa.edu/boa/BOA.pdf

14. The R Project for Statistical Computing. Accessed at http://www.r-project.org/

15. Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10: 325-37.

Figure 1. Overall log odds ratios for Bayesian hierarchical models scenarios 1-6

The overall log odds ratios (μ) and associated 95% confidence intervals (CIs) from the simulations are presented for scenarios 1-6. A solid line intersects the x axis at the true overall log odds ratio (i.e., -0.20). A dashed line intersects the x axis at no effect (i.e., 0).

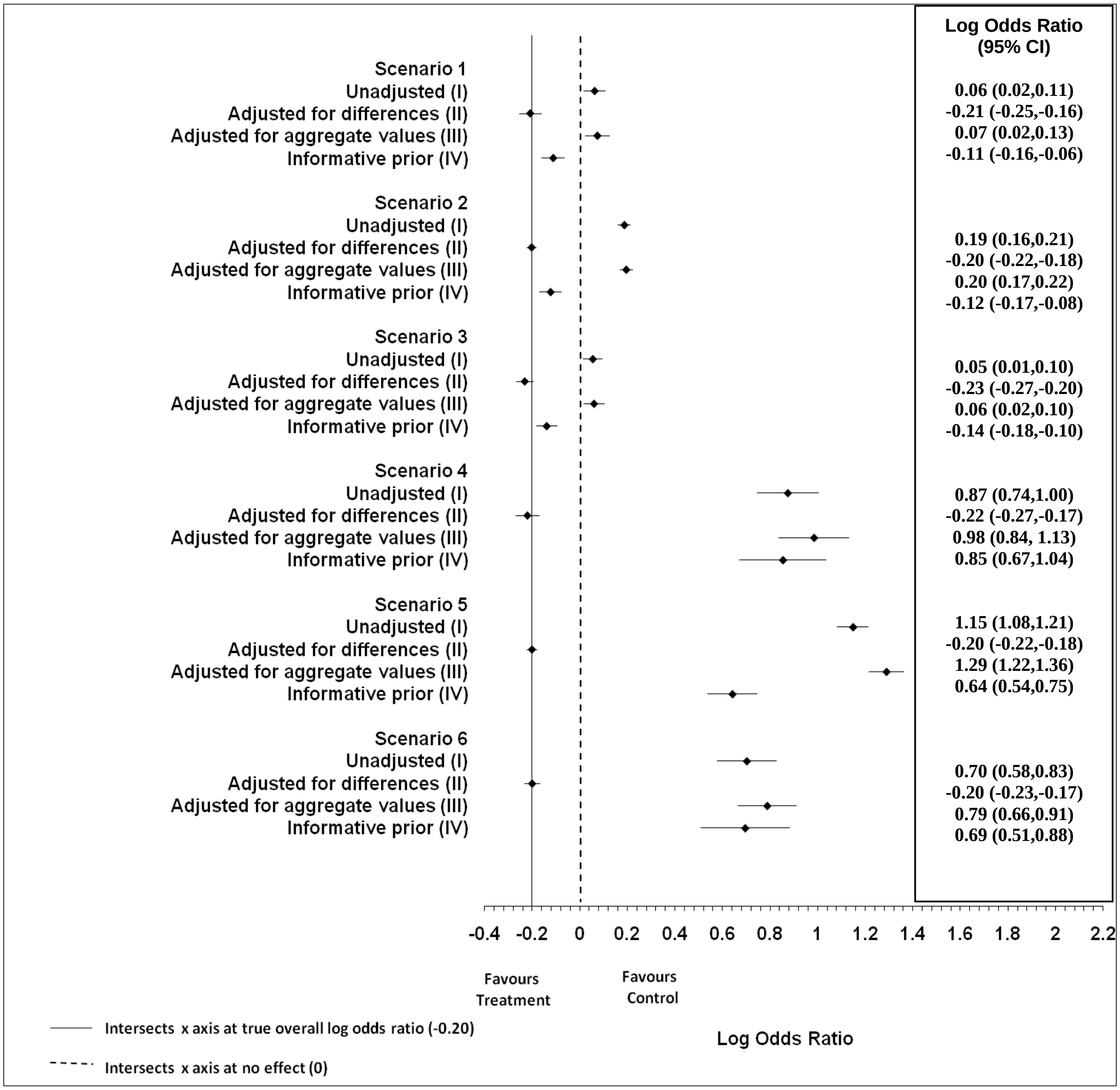


Table 1. Simulation parameters for scenarios 1-6

	Criteria					
Scenario	**Impact of imbalances in non-randomised studies**[a]	**Number of randomised studies**	**Number of non-randomised studies**	**Study arm size randomised studies**[b]	**Study arm size non-randomised studies**[b]	**True overall log odds ratio**
1	0.10	4	4	100-500	100-500	-0.20
2	0.10	4	40	100-500	100-500	-0.20
3	0.10	4	4	100-500	500-1000	-0.20
4	0.50	4	4	100-500	100-500	-0.20
5	0.50	4	40	100-500	100-500	-0.20
6	0.50	4	4	100-500	500-1000	-0.20

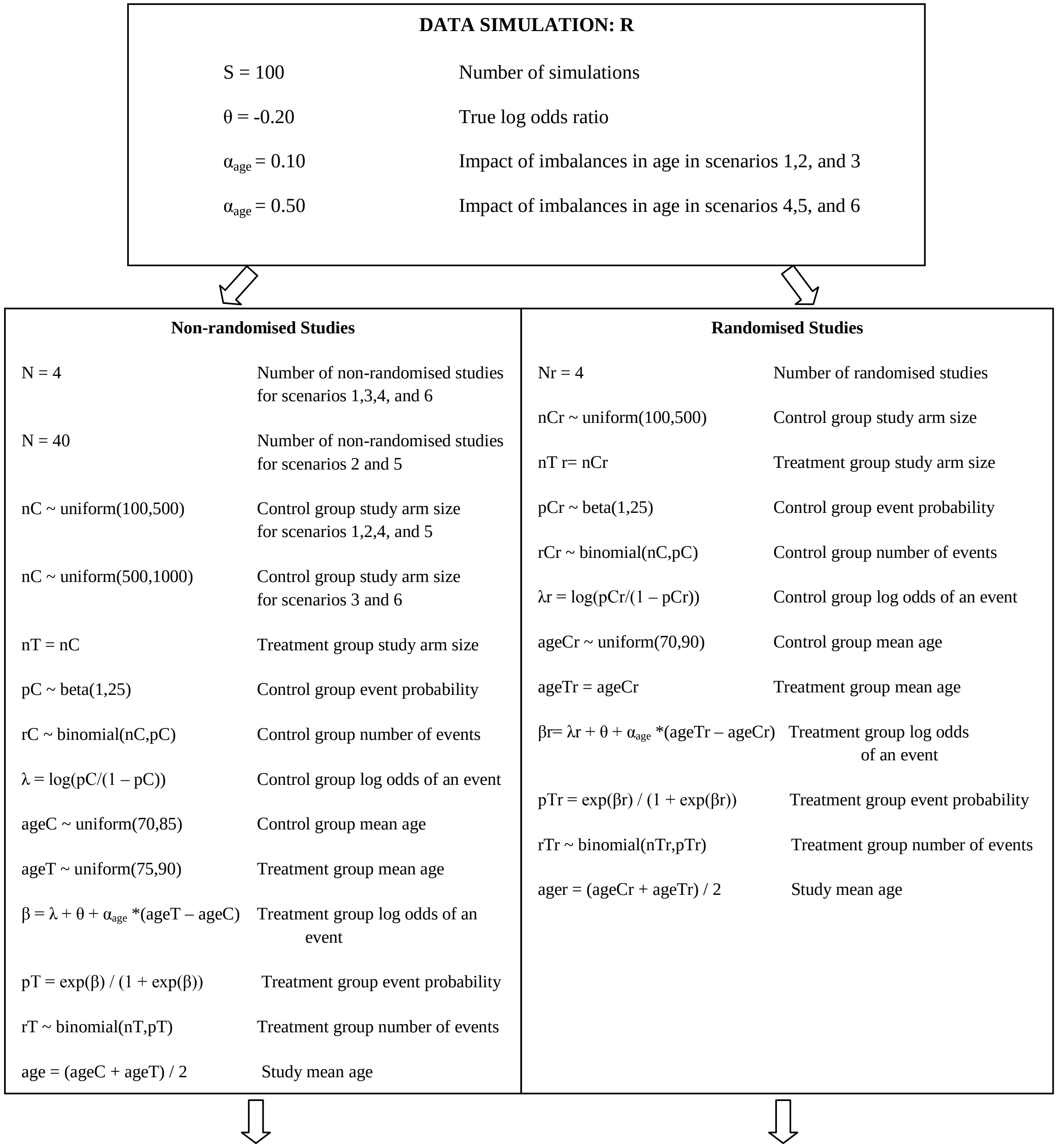
[a] α_{age} measured on the log scale , [b] sampled from a uniform distribution

Table 2: Simulation results comparing Bayesian hierarchical models for scenarios 1-6

Scenario	Model	Mean median log odds ratio	Standard error mean median log odds ratio	Bias	Z-statistic
1	Unadjusted (I)	0.06253	0.02268	0.26253	11.57665
	Adjusted for differences (II)	-0.20836	0.02374	-0.00836	-0.35207
	Adjusted for aggregate values (III)	0.07407	0.02622	0.27407	10.45383
	Informative prior (IV)	-0.11156	0.02437	0.08844	3.62828
2	Unadjusted (I)	0.18750	0.01330	0.38750	29.12960
	Adjusted for differences (II)	-0.20216	0.01010	-0.00216	-0.21398
	Adjusted for aggregate values (III)	0.19520	0.01355	0.39520	29.17138
	Informative prior (IV)	-0.12240	0.02385	0.07760	3.25356
3	Unadjusted (I)	0.05473	0.02079	0.25473	12.25142
	Adjusted for differences (II)	-0.23125	0.01816	-0.03125	-1.72104
	Adjusted for aggregate values (III)	0.05979	0.02189	0.25979	11.86562
	Informative prior (IV)	-0.13908	0.02235	0.06092	2.72561
4	Unadjusted (I)	0.87357	0.06602	1.07357	16.26034
	Adjusted for differences (II)	-0.22000	0.02535	-0.02000	-0.78904
	Adjusted for aggregate values (III)	0.98388	0.07572	1.18388	15.63405
	Informative prior (IV)	0.85343	0.09327	1.05343	11.29504
5	Unadjusted (I)	1.14790	0.03313	1.34790	40.67943
	Adjusted for differences (II)	-0.20083	0.01146	-0.00083	-0.07268
	Adjusted for aggregate values (III)	1.28827	0.03734	1.48827	39.85580
	Informative prior (IV)	0.64133	0.05340	0.84133	15.75488
6	Unadjusted (I)	0.70170	0.06319	0.90170	14.27030
	Adjusted for differences (II)	-0.19981	0.01721	0.00019	0.01117
	Adjusted for aggregate values (III)	0.78753	0.06303	0.98753	15.66646
	Informative prior (IV)	0.69489	0.09509	0.89489	9.41122

Figure S1. Flow chart depicting data simulation, analysis and output for scenarios 1-6

The flow chart depicts the simulation of the data in R, the analysis of the simulated data in WinBUGS and the statistics used to assess the performance of the four models.



DATA SIMULATION: R

S = 100	Number of simulations
$\theta = -0.20$	True log odds ratio
$\alpha_{age} = 0.10$	Impact of imbalances in age in scenarios 1,2, and 3
$\alpha_{age} = 0.50$	Impact of imbalances in age in scenarios 4,5, and 6

Non-randomised Studies

N = 4	Number of non-randomised studies for scenarios 1,3,4, and 6
N = 40	Number of non-randomised studies for scenarios 2 and 5
nC ~ uniform(100,500)	Control group study arm size for scenarios 1,2,4, and 5
nC ~ uniform(500,1000)	Control group study arm size for scenarios 3 and 6
nT = nC	Treatment group study arm size
pC ~ beta(1,25)	Control group event probability
rC ~ binomial(nC,pC)	Control group number of events
$\lambda = \log(pC/(1 - pC))$	Control group log odds of an event
ageC ~ uniform(70,85)	Control group mean age
ageT ~ uniform(75,90)	Treatment group mean age
$\beta = \lambda + \theta + \alpha_{age} *(ageT - ageC)$	Treatment group log odds of an event
$pT = \exp(\beta) / (1 + \exp(\beta))$	Treatment group event probability
rT ~ binomial(nT,pT)	Treatment group number of events
age = (ageC + ageT) / 2	Study mean age

Randomised Studies

Nr = 4	Number of randomised studies
nCr ~ uniform(100,500)	Control group study arm size
nT r= nCr	Treatment group study arm size
pCr ~ beta(1,25)	Control group event probability
rCr ~ binomial(nC,pC)	Control group number of events
$\lambda r = \log(pCr/(1 - pCr))$	Control group log odds of an event
ageCr ~ uniform(70,90)	Control group mean age
ageTr = ageCr	Treatment group mean age
$\beta r= \lambda r + \theta + \alpha_{age} *(ageTr - ageCr)$	Treatment group log odds of an event
$pTr = \exp(\beta r) / (1 + \exp(\beta r))$	Treatment group event probability
rTr ~ binomial(nTr,pTr)	Treatment group number of events
ager = (ageCr + ageTr) / 2	Study mean age

ANALYSIS: WinBUGS

N = 4	Number of non-randomised studies for scenarios 1,3,4, and 6
N = 40	Number of non-randomised studies for scenarios 2 and 5
Nr = 4	Number of randomised studies
i = 1,.......,N	For non-randomised studies
j = 1,.......,Nr	For randomised studies
nCi	Control group study arm size non-randomised study i
nTi	Treatment group study arm size non-randomised study i
rCi	Control group number of events non-randomised study i
rTi	Treatment group number of events non-randomised study i
nCrj	Control group study arm size randomised study j
nTrj	Treatment group study arm size randomised study j
rCrj	Control group number of events randomised study j
rTrj	Treatment group number of events randomised study j
ageCi	Control group mean age non-randomised study i
ageTi	Treatment group mean age non-randomised study i
ageCrj	Control group mean age randomised study j
ageTrj	Treatment group mean age randomised study j
agei	Study mean age non-randomised study i
agerj	Study mean age randomised study j

⇩

OUTPUT

θ_k = median log odds ratio for simulation *k*

θ_{bar} = Mean median log odds ratio = $\sum_{k=1}^{S} \theta_k \;/\; S$

SD = Standard deviation = $\sqrt{[1/(S-1)]\sum_{k=1}^{S}(\theta_k - \theta_{bar})^2}$

SE = Standard error mean median log odds ratio = $SD/\sqrt{S}$

Bias = $\theta_{bar} - \theta$

Z-statistic = Bias / SE

CHAPTER 4

The impact of using informative priors in a Bayesian cost-effectiveness analysis: an application of endovascular versus open surgical repair for abdominal aortic aneurysms in high risk patients

C Elizabeth McCarron MA, MSc[1,2***], Eleanor M Pullenayegum PhD[1,3], Lehana Thabane PhD[1,3], Ron Goeree MA[1,2], Jean-Eric Tarride PhD[1,2]

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

[2]Programs for Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[3]Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[***]Corresponding author

Citation: McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The impact of using informative priors in a Bayesian cost-effectiveness analysis: an application of endovascular versus open surgical repair for abdominal aortic aneurysms in high risk patients. *Submitted October 3, 2011 to Medical Decision Making*.

Abstract:

Background:

Bayesian methods have been proposed as a way of synthesising all available evidence to inform decision making. However, few practical applications of the use of Bayesian methods for combining patient level data (i.e., trial) with additional evidence (e.g., literature) exist in the cost-effectiveness literature. To address the lack of such applied examples, the objectives of this study were to compare Bayesian and non-Bayesian methods to assess the impact of incorporating additional information into a cost-effectiveness analysis.

Methods:

Patient level data from a previously published non-randomised study were first analysed using traditional bootstrap techniques and then compared using bivariate Normal Bayesian models with vague and informative priors. Two different types of informative priors were considered to reflect different valuations of the additional evidence relative to the patient level data (i.e., 'face value' and 'sceptical'). Models were compared in terms of estimates for expected costs and effects and cost-effectiveness acceptability curves (CEACs).

Results:

The bootstrapping and Bayesian analysis using vague priors provided similar results in terms of both cost and effect estimates and CEACs. The most pronounced impact of incorporating the informative priors was the increase in estimated life years in the control arm relative to what was observed in the patient level data alone. In our example, the incremental difference in life years originally observed in the patient level data was reduced and the CEACs shifted accordingly.

Conclusion:

The results of this study demonstrate the potential impact and importance of incorporating additional information into an analysis of patient level data. As the results suggest this could alter the decision as to whether or not a treatment is cost-effective and should be adopted.

1. Introduction

Economic evaluations are an important tool for informing health care decision making. As they compare the relative costs and effects of alternative interventions, economic evaluations provide decision makers with information necessary to make rational decisions regarding the allocation of scarce resources. Critical to this process are the sources of evidence from which estimates of the relative costs and effects are derived.

In the case of an economic evaluation conducted alongside a clinical trial (i.e., a patient level analysis), cost and effect data would be determined for each patient in the study. These sample data could then be used to generate estimates for the mean costs and effects for patients under each of the treatments being compared. As these values represent estimates for the true mean costs and effects, uncertainty around these sample values is often incorporated using the non-parametric bootstrap method [1]. The bootstrap propagates uncertainty using only the information contained in the data, effectively completely discounting all other sources of evidence external to the trial (e.g., literature). In contrast, in a Bayesian approach the trial data as well as any external evidence can be taken into account through the combination of the prior distributions (i.e., external evidence) and the likelihood function (i.e., the data from the trial) [2], thus allowing for a more comprehensive approach to the incorporation of uncertainty.

Despite the importance of incorporating all available evidence to inform decision making [3-6], a recent review [7] of sixteen trial based Bayesian cost-effectiveness studies reported that 50% of the studies used non-informative or vague priors only in their analyses. This provides little guidance to policy makers on the potential of Bayesian methods to integrate all available evidence to capture the uncertainty inherent in decision making [7]. Non-informative or vague priors are appropriate in those situations where there is a genuine lack of additional (i.e., prior or new external) information. However, in those situations where prior information exists, or new information becomes available either during the course of a trial or after its completion, failure to take this into account could impact the results. Through the use of the prior distribution the Bayesian approach provides a mechanism by which this additional information can be incorporated into a trial based cost-effectiveness analysis. At the very least it would be useful to have a sense of what impact this external evidence might have on the trial results.

Six of the sixteen studies in the review examined the impact the information contained in the priors had on the cost-effectiveness results [8-13]. The sources used to inform the prior distributions included trials, Medicare claims data, and informal reasoning. Of the six studies, three reported [8-10] that the more informative priors led to higher probabilities of cost-effectiveness for the respective interventions and one study [11] reported that the more informative priors were associated with lower probabilities of cost-effectiveness. Two studies [12,13] indicated that the results were insensitive to changes in the prior distributions. In addition to the potential impact on the results of using informative priors based on external evidence, there are also issues to consider concerning the relative value of this information compared to the patient level data from the trial. For instance, one study in the review examined the potential impact of discounting the additional information. This study [8] compared three different prior distributions: 1) a non-informative prior disregarding all information from an additional trial, 2) a prior that used 50% of the information from the additional trial and 3) a prior using all of the information from the additional trial. The authors concluded that the different prior distributions, and the strength of information contained therein, may lead to different decisions. For example, given a willingness to pay of 30,000 Netherland guilders per event-free survivor, the probability of cost-effectiveness was 0.65 for the non-informative prior, 0.80 for the prior based on 50% of the information from the first trial, and 0.90 for the prior based on all of the information from the first trial.

These results speak to the potential impact on decision making that additional evidence could have when combined with a patient level economic evaluation. To the best of our knowledge, no other study has attempted to evaluate the impact on the cost-effectiveness results of using different methods to incorporate the external evidence. To generalise these findings, the objective of the current analysis is to compare the results of a traditional frequentist analysis (i.e., non-parametric bootstrap) that relies only on the information contained in the patient level data to a Bayesian approach incorporating evidence from both the patient level data as well as other sources (i.e., published trials). Contrary to the studies from the review which tended to rely on a single source of information (e.g., a trial) to inform their priors, our analyses combined the results of published studies, available at the time of the original analysis, in a meta-analysis. The paper also makes use of two different types of informative prior distributions to reflect different potential valuations of the additional information (i.e., 'face value' and 'sceptical'). These prior distributions are then used to combine the additional information with the patient level data from a published trial based economic evaluation comparing endovascular aneurysm repair (EVAR) with open surgical repair (OSR) [14].

2. Case study

A previous trial based economic evaluation comparing elective EVAR and OSR for the treatment of abdominal aortic aneurysms for patients at high surgical risk provides the patient level data for the current analysis [14]. These data were based on a one year non-randomised study conducted at a single site in Ontario, Canada. Total costs expressed in 2006 Canadian dollars and life years at one year were reported for 140 EVAR patients (treatment group) and 52 OSR patients (control group). Despite the non-randomised nature of the trial, the two groups were matched in terms of clinical characteristics [14]. The estimated mean costs indicated that EVAR ($34,147) was slightly less expensive than OSR ($34,170) and estimated mean life years indicated EVAR (0.96) was more effective than OSR (0.85). Thus, on the basis of point estimates only, EVAR dominated

OSR in terms of incremental cost per life year gained. However, when sampling uncertainty in the trial data was incorporated using the non-parametric bootstrap, the differences in costs were not statistically significant at the 5% level (mean difference of -$24; 95% confidence interval (CI) of -$11,582 to $9,165). EVAR was still more effective than OSR in terms of life years at one year (mean difference of 0.11; 95% CI of 0.022 to 0.213). Cost-effectiveness acceptability curves (CEACs) were used to represent decision uncertainty and the results indicated that at a willingness to pay of $50,000 per life year gained (LYG) the probability of EVAR being cost-effective was 0.76. As part of this health technology assessment, the authors conducted a systematic review of published studies comparing EVAR and OSR [15]. However the results of the review were not incorporated into the analysis of the trial data.

3. Methods

The following describes the methods being compared and introduces the sources of evidence used to illustrate the potential impact of incorporating informative priors into a Bayesian trial based economic evaluation.

3.1 The bootstrapping method

The bootstrapping method is non-parametric by nature, meaning it makes no assumption about the parametric distribution of the data. The method re-samples with replacement from the original sample data to build an empirical estimate of the sampling distribution for the statistic of interest [1]. Although non-parametric bootstrapping does not assume any particular form of distribution, the choice of statistic used implicitly does. For example, if the sample mean is the statistic chosen to be monitored in the repeated samples, the results will be similar to those based on a parametric assumption of normality [16], provided the sample size is large enough.

One thousand bootstrap replicates were generated to estimate the sampling distribution for the sample mean costs and effects for both the EVAR and OSR groups as well as for the incremental costs and effects of EVAR compared to OSR. Using the percentile method, the limits of the 95% CIs around the sample means were calculated based on the 25th and 976th ordered values.

3.2 Bayesian analysis

The basis for making inferences from a Bayesian perspective is Bayes' theorem. In essence Bayes' theorem describes the combination of information from two sources, the likelihood and the prior [2]. The likelihood function summarizes all of the information that is contained in the data (e.g., a trial). In the current analysis this refers to the patient level data comparing EVAR and OSR [14]. The prior distribution represents information that is available in addition to the data. In this analysis the prior describes the information from the literature available at the time of the original analysis [15]. In the absence of additional information, vague or uninformative prior distributions can be used. The less informative the prior, the more weight is given to the data in the analysis. The priors are combined with the data to generate the posterior distribution which represents what is now known about the unknown quantity (e.g., mean effects) given the prior information and the data. The posterior is proportional to the product of the likelihood (i.e., the data) and the prior [2].

3.2.1 Bivariate Normal likelihood

In order to allow for the potential correlation between costs and life years, the cost and effect data were modelled using correlated Normal distributions [16] where the 140 EVAR and 52 OSR patients were indexed by i and the two study arms were indexed by j (i.e., j=1 for EVAR and 2 for OSR):

$$C_{ij} \sim \text{Normal}(\mu_{Cj}, \sigma_{Cj}^{2}) \qquad \text{(eq.1)}$$

$$E_{ij} \sim \text{Normal}(\mu_{Eij}, \sigma_{Ej}^{2}) \qquad \text{(eq.2)}$$

$$\mu_{Eij} = \mu_{Ej} + \beta_j (C_{ij} - \mu_{Cj}) \qquad \text{(eq.3).}$$

Here the costs have a Normal distribution with mean μ_{Cj} and standard deviation σ_{Cj} (eq.1). The effects have a Normal distribution with mean μ_{Eij} and standard deviation σ_{Ej} (eq.2). As seen in equation three, the mean of E_{ij} depends, through the parameter β_j, on how much the cost C_{ij} is above the mean cost μ_{Cj}. The subtraction of μ_{Cj} ensures that μ_{Ej}

remains interpretable as the overall mean effect in the jth arm of the study. As implied by the regression in equation three, effects have been made a function of costs. The model allows the correlation between costs and life years to be different in the two study groups, through the separate respective β_j parameters [16].

3.2.2 Vague priors

In addition to the likelihood function, a Bayesian analysis requires prior distributions for the unknown population parameters. In the initial analysis vague priors were used so that the resulting inferences essentially depended only on the data. In that regard, we would expect the results from the Bayesian analysis to be similar to those from the non-parametric bootstrap approach [16].

3.2.3 Informative priors

In order to incorporate all available evidence in the Bayesian cost-effectiveness analysis the results from a published systematic review [15], which identified eight non-randomised studies conducted in high risk patients, were combined with the trial data. The review provided estimates of 30 day post-operative mortality for the eight studies of high risk patients. Two of the high risk studies also provided estimates of longer-term mortality, but not at one year (i.e., mean follow-up of 26.8 months for EVAR and 27.6 months for OSR in one study [17] and 15.6 months for EVAR and 19.8 months for OSR in the other study [18]). In addition to the body of evidence in high risk patients, the review also contained information from another eight non-randomised studies that were not restricted to high risk patients, but measured mortality at 30 days and one year post treatment in a mixed population of low to high risk patients.

To estimate the one year mortality rate in a high risk population, the 30 day mortality rates observed in the eight high risk studies were combined with conditional probabilities measuring the probability of being dead at one year given you are alive at 30 days. These conditional probabilities were calculated from two sets of evidence. First, the two high risk studies reporting mortality data at around two years were used [17,18], assuming that the one year and two year probabilities of being dead conditional on being alive at 30 days were similar. The second set of evidence consisted of the eight studies which measured mortality at 30 days and at one year in a mixed risk population [15].

Table 1 presents details of the studies used for the informative priors, including the mortality rates at 30 days for the eight high risk studies and the mortality rates at one year conditional on being alive at 30 days for the two high risk and eight mixed risk studies. For EVAR, the mortality rates for both 30 days and conditional on being alive at 30 days for the studies were fairly consistent with those from the trial (i.e., 3% vs. 1% for 30 day mortality and 4% high risk, 5% mixed risk vs. 6% for conditional rates). In contrast, the studies reported, on average, lower mortality rates for OSR compared to the trial (i.e., 6%

vs. 10% for 30 day mortality and 3% high risk, 3% mixed risk vs. 9% for conditional rates). In order to estimate the one year mortality associated with EVAR and OSR in high risk patients, for each of the two sets of evidence, binomial models were constructed in WinBUGS [19] to combine information on 30 day mortality and longer-term mortality. The details of these calculations are provided in the Appendix. In the absence of additional information on costs, the informative priors were limited to effects.

For each of the two sets of data used to estimate the one year mortality to inform the prior on effects (i.e., two high risk studies and eight mixed risk studies), two different types of informative prior distributions for μ_{Ej} were examined [20]. Table 2 presents these details along with the vague priors. As evidenced by the informative prior distributions given in Table 2 for the mean life years, the lower mortality rates for OSR reported in the literature translated into higher estimates for mean life years in the OSR group (i.e., 0.93) relative to the patient level data (i.e., 0.85). The results for EVAR were roughly the same for both the informative priors (i.e., 0.95) and the patient level data (i.e., 0.96). The first informative prior used in the analysis was labelled a ‘face value’ prior since the additional information was taken at face value and no concession, beyond that due to between study heterogeneity, was made for any potential differences between the additional information and the patient level data. This could reflect a belief that the evidence from the literature was of high quality and as reliable as the patient level data. The second informative prior was labelled ‘sceptical’ since caution was being expressed due to concerns about the additional information. Specifically, there could be issues concerning the risk level of the patients in the studies, the time periods over which mortality was measured or the quality of the evidence. In that case, the additional information could be explicitly given less weighting than the patient level data. To downweight the external evidence relative to the trial data, we used a prior variance for mean life years that was four times the variance of the patient level data. This was based on a previous study by Sutton and Abrams [20]. To get a better understanding of the impact of using different inflation factors to downweight the additional information, a sensitivity analysis was conducted (e.g., inflating the variance by two rather than by four).

3.2.4 Estimations

All posterior distributions of quantities of interest for both the informative priors and the Bayesian cost-effectiveness analyses were estimated in WinBUGS [19]. For all Bayesian analyses an initial burn-in of 100 000 iterations was discarded to ensure convergence. History plots, autocorrelation plots, and various diagnostics available in the package Bayesian Output Analysis [21], performed on two chains, were used to assess convergence. Posterior estimates were based on a subsequent sample of 100 000 iterations. These posterior distributions were summarized as posterior means and 95% credible intervals (CrIs). In contrast to a frequentist 95% confidence interval, a Bayesian 95% credible interval is an interval that has a 95% probability of containing the true parameter value. Other estimated quantities included the mean cost difference (ΔC) and the mean effect difference (ΔE) between the EVAR and OSR groups. Representation of decision uncertainty was done through the use of CEACs which show the probability of EVAR being cost-effective at different threshold values (e.g., $50,000 per LYG).

4. Results

4.1 The bootstrapping method

The estimated values for the mean and incremental costs and life years and their associated 95% CIs are presented in Table 3 for both the EVAR and OSR groups. These estimates are based on the patient level data from Tarride, Blackhouse, De Rose, Novick, Bowen, Hopkins et al. [14] and the 1000 bootstrap replicates that were used to estimate sampling uncertainty. The results closely correspond to those from the original study (i.e., ΔC = -$24(-$11582, $9165) and ΔE = 0.11(0.02, 0.21)) [14].

4.2 Bayesian analysis

4.2.1 Vague prior distributions

The bivariate Normal likelihood was used to accommodate the negative correlation observed between total costs and life years in both the EVAR (-0.20) and OSR (-0.31) groups. The posterior mean estimates and 95% CrIs obtained from the Bayesian analysis with vague priors were similar to the mean estimates and 95% CIs from the non-parametric bootstrap (Table 3). The CEACs were also quite similar (Figure 1). These results reinforce the vagueness of the prior distributions and suggest that most of the information in the analysis is coming from the patient level data.

4.2.2 Informative priors

In contrast to the results for the vague prior distributions, the incorporation of informative priors for mean life years increased the posterior estimates for mean life years in the OSR group from 0.85 LYG to between 0.87 LYG and 0.89 LYG depending on the type of prior used (e.g., 'face value'). Similarly the associated 95% CrIs shifted upwards and as a result of the added information became narrower. The posterior estimates for mean life years and the associated intervals were unchanged for EVAR. These results were consistent across both types of priors for both sets of evidence (Table 3).

The extent of the increase relative to the mean values observed in the non-parametric bootstrap and vague models reflected the weight of the information associated with each type of informative prior. In the current analysis the weight of the additional information relative to the data decreased as the priors moved from 'face value' to 'sceptical.' This was also apparent in the incremental estimates as the differences in mean life years between EVAR and OSR got progressively larger as the additional information was given less weight. In terms of the 95% CrIs associated with these incremental differences they all shifted downwards and became more precise compared to those based on vague priors. The shifts in the credible intervals were such that for the 'face value' priors the incremental differences in life years ceased to be statistically significant at the 5% level. They remained statistically significant, though barely, for the 'sceptical' priors in which the variance was inflated by a factor of four. A sensitivity analysis revealed that the additional evidence for the OSR group would have to be downweighted by between two and half and three times the variance of the patient level data in order for the results to remain statistically significant.

Due to modelling the correlation between costs and effects, the informative priors on life years also impacted the mean costs. As the mean life years in the OSR group increased, the mean costs in the OSR group decreased. Since the incremental costs and effects both directly contribute to the estimation of the CEACs, the combined impact of these changes on decision uncertainty can be seen in a comparison of these curves. Figure 1 shows the impact of the priors on the probability of EVAR being cost-effective compared to OSR at different threshold values. The curves indicate that the more informative the prior, the lower the probability that EVAR was cost-effective for a given willingness to pay. For example, at a willingness to pay of $50,000 per LYG the probability of EVAR being cost-effective was approximately 0.60 and 0.70 respectively for the 'face value' and ‘sceptical’ priors for both sets of evidence, which compared to a probability of about 0.80 for the vague model.

5. Discussion

By comparing the non-parametric bootstrap to a Bayesian approach with both vague and informative priors this study has sought to assess the potential impact of incorporating all available evidence into a trial based economic evaluation. While the non-parametric bootstrap and the Bayesian approach using vague priors produced similar results, our study has demonstrated the potential for informative priors to impact expectations and decision uncertainty. For instance, in the bivariate Normal model if the external evidence was taken at ‘face value,’ a decision maker would have to be willing to pay approximately $20,000 per LYG, in order for EVAR to have a greater probability of being cost-effective than OSR. Alternatively, when the external information was ignored and vague priors were used, EVAR was more likely to be cost-effective relative to OSR for all values of willingness to pay. Based on whether the additional information was incorporated into the analysis and depending on a decision maker’s willingness to pay for a LYG, this could result in very different funding decisions. These results also speak to the importance of incorporating the potential correlation between costs and effects, even when the correlation appears modest. The impact on decision uncertainty observed in this study as well as in the studies from the previous review [7] suggest the synthesis of evidence from different sources could also play a role in decisions about future research, ensuring that resources are used efficiently. This could be particularly important in those situations where the additional information suggests something different from the patient level data, as was observed in our case study.

In addition to exploring the potential impact of combining all available evidence, this study also considered how the additional information might be weighted or valued relative to the patient level data from the original cost-effectiveness analysis. As the objective is to combine all available evidence to inform decision makers, this study provides insight into how multiple sources of evidence may be combined together in the prior and used in addition to the trial data. Integral to this process is an understanding of how to value the additional information relative to the patient level data. Attempts were made to assess the impact on the cost-effectiveness results of different types of informative prior distributions. Specifically, two types of priors were examined (i.e., ‘face value’ and ‘sceptical’).

In terms of deciding how much the additional information should contribute to the analysis, a more thorough consideration would need to be given as to why the mortality rates for OSR reported in the literature differed from the patient level estimates. This could have implications both in terms of the weight ascribed to the additional information and to the potential need for future research. Unfortunately, none of the studies in the literature provided detailed information on all of the clinical characteristics necessary to evaluate the risk level of the patients. In addition, although both the trial based economic evaluation and the studies from the literature were non-randomised, the trial was well balanced in terms of patient characteristics, while there was evidence of covariate imbalance among the literature studies. Again, though, attempts to understand the potential impact of these imbalances are limited by the extent of missing covariate data.

In combination, these factors (i.e., real surgical risk level unknown in many studies and non-randomised evidence) suggest that we may be unlikely to take the evidence from the literature at 'face value.' As in our case study, this essentially gives the external evidence and the patient level data for the OSR group equal weighting. Rather some degree of downweighting would seem to be necessary. The results of the sensitivity analysis indicate that the additional information for the OSR group must be downweighted by at least 60% in order for the incremental differences in effects to remain statistically significant at the 5% level. Whether this represents a reasonable valuation of the evidence in the literature relative to the patient level data is not clear and likely would require additional research. Future research could also look at the feasibility of using models that elicit expert opinion concerning the rigour and relevance of the studies being combined [22]. The use of hierarchical modelling either instead of or in addition to the use of informative priors could also be explored as a way of combining additional evidence with patient level data. Specifically, methods have been proposed that use estimates from previously published meta-analyses to adjust and downweight studies [23]. Again, the limited availability of covariate data would likely make any assessments regarding adjustment and downweighting difficult in the current analysis.

This paper focussed on the use of the prior distribution to combine all available evidence in a Bayesian cost-effectiveness analysis. A possible concern was the absence of data for life years at one year post treatment for the eight high risk studies from the literature. This meant that these values had to be estimated. Though actual data would have been preferable, the similarity of the outcomes for both sets of information (i.e., two high risk and eight mixed risk studies) reinforced the results. Under ideal circumstances, additional information on total one year costs in EVAR and OSR patients would also have been available. Another possible limitation of the analysis is that the study has assumed Normal distributions for both costs and life years. Although the data may be skewed, and as a result, the costs and effects not normally distributed, the central limit theorem (CLT) states that for any population distribution of costs and effects the distributions of the sample means will converge to Normal distributions as the sample size increases [24]. The simulation results of Nixon, Wonderling and Grieve [24] suggest that the CLT can be invoked in the current study to justify the assumption of Normal distributions for both costs and life years. Their simulations were based on different scenarios for sample size and skewness and indicate that for moderate to large sample sizes (i.e., n>50) the CLT performs well [24]. However, other distributions which may fit the data better could lead to more efficient estimators for the population means [25]. The associated increase in precision could have implications in terms of decision uncertainty and consequently future research priorities. Gamma distributions, for example, have been used to deal with the potential skewness in cost data. Since the assumption of distributions other than normal can lead to misleading conclusions if the assumptions are incorrect [25] it may be safer to assume normality [26], provided the sample size is sufficient to justify the CLT.

Despite these limitations, this study has demonstrated the potential importance of using all available evidence to inform decision makers. Where cost-effectiveness analyses and economic evaluations are a critical input to health care policy making, it is paramount that these policy decisions be based on the available evidence. This study contributes to the literature an example of how this may be achieved using actual data from a previous patient level cost-effectiveness analysis and evidence available from the literature at the time of the original analysis. Future research could focus on further refinements and of course the approaches undertaken will likely vary depending on the context and availability of data.

6. Conclusion

This analysis indicates that ignoring specific sources of evidence could undermine cost-effectiveness results. Not only might it change inferences and possibly influence decisions regarding the cost-effectiveness of one intervention compared to another, but it could also impact decisions regarding the need for future research. Only when all available evidence is taken into consideration can we be confident of well informed health care decisions.

References

1. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. Oxford: Oxford University Press; 2005.
2. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, West Sussex: John Wiley & Sons Ltd; 2004.
3. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics. 2006; 24(1):1-19.
4. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. Stat Med. 2000; 19:3359-3376.
5. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making. Health Econ. 2006; 15:677-687.
6. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. BMC Med Res Methodol. 2009; 9:29.
7. McCarron CE, Pullenayegum EM, Marshall DA, Goeree R, Tarride JE. Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments. Int J Technol Assess Health Care. 2009; 25(4): 546-554.
8. Al MJ, Van Hout BA. A Bayesian approach to economic analyses of clinical trials: the case of stenting versus balloon angioplasty. Health Econ. 2000; 9(7):599-609.

9. O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. Stat Med. 2001; 20(5):733-753.

10. Shih YC, Bekele NB, Xu Y. Use of Bayesian net benefit regression model to examine the impact of generic drug entry on the cost effectiveness of selective serotonin reuptake inhibitors in elderly depressed patients. Pharmacoeconomics. 2007; 25(10):843-862.

11. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. Health Econ. 2001; 10(4):303-315.

12. Briggs AH. A Bayesian Approach to Stochastic Cost-Effectiveness Analysis: An Illustration and Application to Blood Pressure Control in Type 2 Diabetes. Int J Technol Assess Health Care. 2001; 17(1):69-82.

13. Heitjan DF, Li H. Bayesian estimation of cost-effectiveness: An importance-sampling approach. Health Econ. 2004; 13(2):191-198.

14. Tarride JE, Blackhouse G, De Rose G, Novick T, Bowen JM, Hopkins R, et al. Cost-effectiveness analysis of elective endovascular repair compared with open surgical repair of abdominal aortic aneurysms for patients at a high surgical risk: A 1-year patient level analysis conducted in Ontario, Canada. J Vasc Surg. 2008; 48(4):779-787.

15. Hopkins R, Bowen J, Campbell K, Blackhouse G, De Rose G, Novick T, et al. Effects of study design and trends for EVAR versus OSR. Vasc Health Risk Manag. 2008; 4(5): 1011-1022.

16. Thompson SG, Nixon RM. How sensitive are cost-effectiveness analyses to choice of parametric distributions? Med Decis Making. 2005; 25:416-423.

17. Mendonca CT, Moreira RCR, Ribas T Jr, Miyamotto M, Martins M, Stanischesk IC, et al. Comparison between open and endovascular treatment of abdominal aortic aneurysms in high surgical risk patients. Journal Vascular Brasileiro. 2005; 4:232-42.

18. Parmer SS, Fairman RM, Karmacharya J, Carpenter JP, Velazquez OC, Woo EY. A comparison of renal function between open and endovascular aneurysm repair in patients with baseline chronic renal insufficiency. J Vasc Surg. 2006; 44:706-11.

19. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput. 2000; 10:325-337.

20. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res. 2001; 10(4): 277-303.

21. Bayesian Output Analysis Program (BOA) Version 1.1 User's Manual 2005 [www.public-health.uiowa.edu/boa/BOA.pdf]

22. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modeling in evidence synthesis. J R Stat Soc Ser A. 2009; 172: 21-47.

23. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. J.R. Statist. Soc. A. 2009; 172:119-136.

24. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health Econ. 2010; 19:316-333.

25. Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence. Health Econ. 2005; 14:421-428.

26. Lambert PC, Billingham LJ, Cooper NJ, Sutton AJ, Abrams KR. Estimating the cost-effectivness of an intervention in a clinical trial when partial cost information is available: A Bayesian approach. Health Econ. 2008; 17:67-81.

Table 1. Mortality data for patient level and additional studies

Study	**Only high risk**	**Length of Follow-up (days)**	**30 day mortality[*]**				**Mortality at end of follow-up conditional on being alive at 30 days[†]**			
			EVAR		**OSR**		**EVAR**		**OSR**	
Patient level			**Patients**	**Deaths (%)**	**Patients**	**Deaths (%)**	**Patients**	**Deaths (%)**	**Patients**	**Deaths (%)**
Tarride 2008	Y	365	140	1 (1)	52	5 (10)	139	9 (6)	47	4 (9)
Additional										
High risk										
Du Toit 1998	Y	30	12	0 (0)	10	1 (10)				
Carpenter 2002	Y	30	174	7 (4)	163	7 (4)				
Forbes 2002	Y	30	7	0 (0)	31	0 (0)				
Patel 2003	Y	30	16	0 (0)	35	6 (17)				
Ianneli 2005	Y	30	34	0 (0)	28	1 (4)				
Mendonca 2005	Y	815 EVAR 840 OSR	18	1 (6)	31	2(6)	17	2 (12)	29	2 (7)
De Donato 2006	Y	30	19	1 (5)	8	1 (13)				
Parmer 2006	Y	475 EVA 602 OSR	52	0 (0)	46	0 (0)	52	0 (0)	46	0 (0)
Mixed risk										
Zarins 1999	N	365					185	8 (4)	60	2 (3)
Becquemin 2000	N	365					71	3 (4)	105	2 (2)
Cohnert 2000	N	365					35	3 (9)	37	0 (0)
Ting 2003	N	365					26	1 (4)	24	0 (0)
Ballard 2004	N	365					22	1 (5)	107	2 (2)
Elkouri 2004	N	365					94	0 (0)	258	0 (0)
Greenberg 2004	N	365					199	7 (4)	78	3 (4)
Bush 2006	N	365					695	40 (6)	1120	77 (7)

[*] Mortality data at 30 days for mixed risk studies were not used in calculations.

[†] Only applies to studies with mortality data beyond 30 days.

Table 2. Vague and informative prior distributions for the Bayesian models

Parameter	**Vague prior**	**Informative prior**[‡]	
		Face value prior	**Sceptical prior**
EVAR			
Mean costs (μ_{C1})	Normal(25000,1E11)	Not applicable	Not applicable
Precision costs ($1/\sigma_{C1}^2$)	Gamma(0.50,1E-07))	Not applicable	Not applicable
Mean life years (μ_{E1})[§]	Beta(1,1)	Normal(0.95,1.01E-03)	Normal(0.95,8.38E-04)
		Normal(0.95,4.73E-04)	Normal(0.95,8.38E-04)
Standard deviation life years (σ_{E1}^2)	Uniform(0,10)	Not applicable	Not applicable
Relationship between costs and life years (β_1)	Normal(0,10000)	Not applicable	Not applicable
OSR			
Mean costs (μ_{C2})	Normal(25000,1E11)	Not applicable	Not applicable
Precision costs ($1/\sigma_{C2}^2$)	Gamma(0.50,1E-07))	Not applicable	Not applicable
Mean life years (μ_{E2})[§]	Beta(1,1)	Normal(0.93,2.21E-03)	Normal(0.93,8.76E-03)
		Normal(0.93,2.12E-03)	Normal(0.93,8.76E-03)
Standard deviation life years (σ_{E2}^2)	Uniform(0,10)	Not applicable	Not applicable
Relationship between costs and life years (β_2)	Normal(0,10000)	Not applicable	Not applicable

[‡] Informative priors were only available for mean life years (μ_{Ej}) in the EVAR and OSR groups.

[§] Two informative priors for each type derived from two sets of evidence (i.e., two high risk and eight mixed risk studies).

Table 3. Results for non-parametric bootstrap and Bayesian models with vague and informative priors

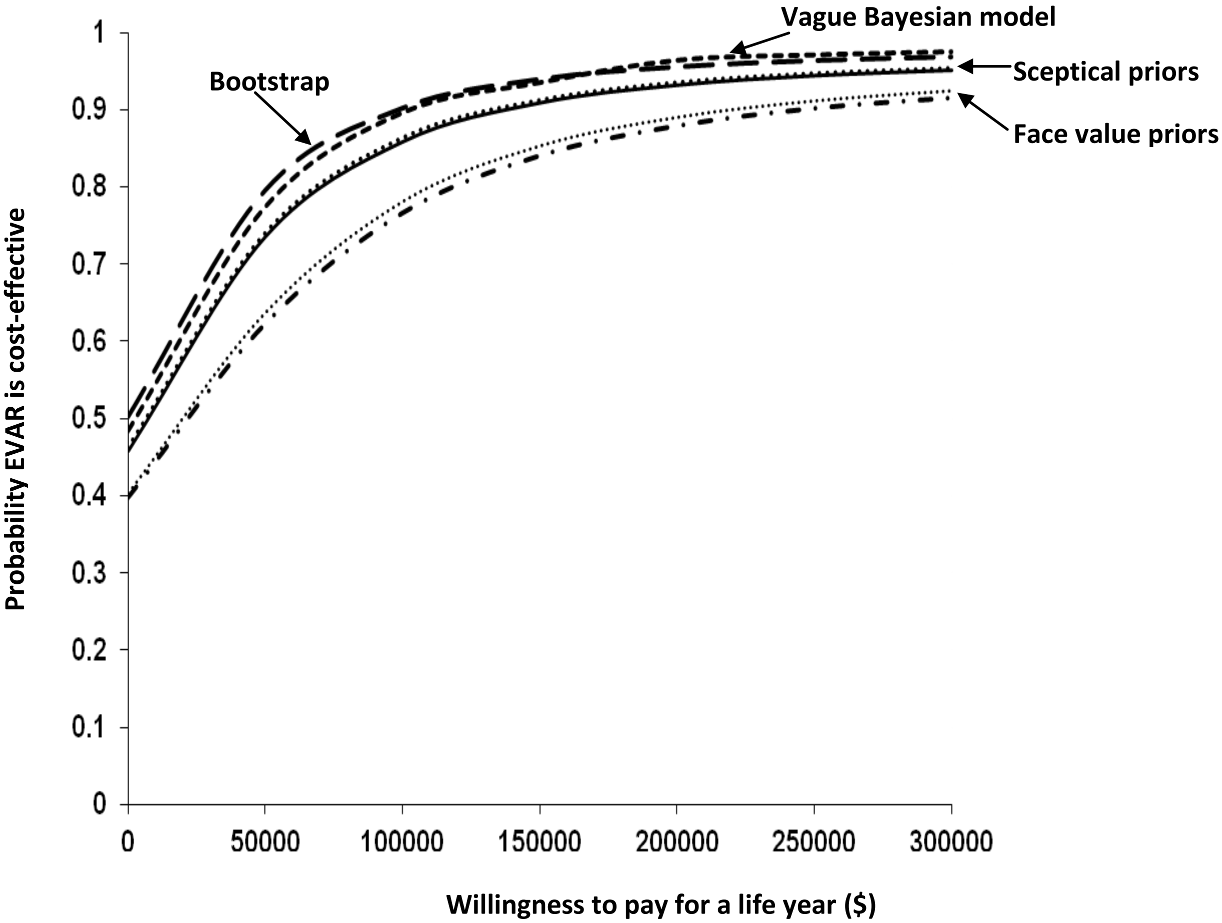
Model	**Annual mean costs ($)**[‖] **($\mu_{Cj}$)**		**Incremental($)**[‖] **($\Delta C$)**	**Mean life years**[‖] **(μ_{Ej})**		**Incremental**[‖] **(ΔE)**
	EVAR	**OSR**	**EVAR-OSR**	**EVAR**	**OSR**	**EVAR-OSR**
Bootstrap	34147 (32485,36356)	34170 (24331,45606)	-23 (-11586,9692)	0.96 (0.93,0.98)	0.85 (0.76,0.93)	0.11 (0.02,0.20)
Bayesian models						
Vague priors	34150 (32240,36050)	34170 (23620,44750)	-18 (-10740,10660)	0.96 (0.93,0.99)	0.85 (0.75,0.94)	0.11 (0.01,0.21)
Informative priors:						
Two high risk studies[¶]						
Face value	34150 (32250,36060)	32840 (22460,43110)	1318 (-9116,11870)	0.96 (0.93,0.98)	0.89 (0.82,0.96)	0.07 (-0.005,0.14)
Sceptical	34160 (32270,36050)	33630 (23180,44030)	531 (-10050,11160)	0.96 (0.93,0.98)	0.87 (0.78,0.95)	0.09 (0.002,0.18)
Eight mixed risk studies[**]						
Face value	34180 (32290,36070)	32790 (22420,43030)	1387 (-9012,11900)	0.96 (0.93,0.98)	0.89 (0.83,0.96)	0.06 (-0.007,0.13)
Sceptical	34170 (32270,36070)	33600 (23130,44010)	565 (-9986,11190)	0.96 (0.93,0.98)	0.87 (0.78,0.95)	0.09 (0.00002,0.18)

[‖] Values in parentheses represent 95% confidence intervals for non-parametric bootstrap and 95% credible intervals for Bayesian models.

[¶] High risk studies reporting mortality after 30 days used in the calculation of the informative priors.

[**] Mixed risk studies reporting mortality at one year used in the calculation of the informative priors.

Figure 1. Cost-effectiveness acceptability curves (CEACs) for non-parametric bootstrap and Bayesian models with vague and informative priors



Legend: Two CEACs are reported for each type of informative prior (i.e., face value and sceptical). Each of the two CEACs refers to the set of evidence on longer-term mortality used in the calculation of the informative priors (i.e., two high risk or eight mixed risk studies).

Appendix to Chapter 4: Calculation of Informative Priors

The following presents the methods used to combine the studies from the literature and generate the informative prior distributions for the cost-effectiveness analysis comparing EVAR and OSR in high risk patients. Eight studies presenting 30 day mortality in high risk patients were found in the literature [15]. The studies did not present information on life years, nor did they present mortality data at one year. Therefore, mortality at one year in high risk patients was estimated by combining 30 day mortality rates from the eight studies in high risk patients with probabilities of being dead at one year conditional on being alive at 30 days from two sources of evidence. The two sets of evidence were as follows: 1) two high risk studies presenting mortality data at around two years; 2) eight studies conducted in a mixed risk population and reporting mortality data at one year. Table 1 of the manuscript presents the details of the studies.

For each of the two sets of data the binomial model given below was used to generate estimates for 30 day mortality and one year mortality conditional on being alive at 30 days:

$$\text{deaths}_{\text{EVARmn}} \sim \text{Binomial}(\text{pdead}_{\text{EVARmn}},\text{patients}_{\text{EVARmn}}) \text{ and}$$
$$\text{deaths}_{\text{OSRmn}} \sim \text{Binomial}(\text{pdead}_{\text{OSRmn}},\text{patients}_{\text{OSRmn}}) \qquad \text{(eq.4)}$$
$$\text{log odds}(\text{pdead}_{\text{EVARmn}}) = \psi_{\text{mn}} \text{ and log odds}(\text{pdead}_{\text{OSRmn}}) = \gamma_{\text{mn}} \qquad \text{(eq.5)}$$
$$\psi_{\text{mn}} \sim \text{Normal}(\theta_{\text{m}}, \sigma_{\text{m}}^2) \qquad \text{(eq.6)}$$
$$\gamma_{\text{mn}} \sim \text{Normal}(\alpha_{\text{m}}, \tau_{\text{m}}^2) \qquad \text{(eq.7)}$$
$$\psi_{\text{m}}.\text{new} \sim \text{Normal}(\theta_{\text{m}}, \sigma_{\text{m}}^2) \qquad \text{(eq.8)}$$
$$\gamma_{\text{m}}.\text{new} \sim \text{Normal}(\alpha_{\text{m}}, \tau_{\text{m}}^2) \qquad \text{(eq.9)}$$

(m = 30 for deaths occurring 0 to 30 days after treatment or 1 for deaths occurring after 30 and up to 365 days post treatment; n = 1,…..,x_m studies).

As shown in equation four, this model assumed that the number of events in each arm of the nth study of time m (i.e., $\text{deaths}_{\text{EVARmn}}$ and $\text{deaths}_{\text{OSRmn}}$ for the treatment and control groups, respectively) followed a binomial distribution defined by the proportion of patients who died in each arm in the nth study of time m (i.e., p_{EVARmn} and p_{OSRmn}) and the total number of patients alive in each arm in the nth study at time zero and 30 days post treatment (i.e., $\text{patients}_{\text{EVARmn}}$ and $\text{patients}_{\text{OSRmn}}$). Equation five describes the log odds for death in the treatment (ψ_{mn}) and control (γ_{mn}) arms of each of the x_m studies. For each of the two time periods, the log odds of dying for both the treatment and control groups were assumed to follow normal distributions with means of θ_m and α_m, respectively. Between-study variability for studies at time m was represented by σ_m^2 for the EVAR group and τ_m^2 for the OSR group. Predictions for the log odds of dying in the patient level trial are provided in equations eight and nine for EVAR and OSR respectively. These distributions incorporate all of the uncertainty associated with θ_m and α_m, the pooled effects in the mth time periods, and σ_m^2 and τ_m^2, the between study variability for the mth time periods.

Prior distributions for the unknown parameters θ_m, σ_m^2, α_m, τ_m^2 were intended to be vague. Normal priors with means of zero and standard deviations of 100 were specified for the mean log odds θ_m and α_m (i.e., 30 day and one year mortality conditional on being alive at 30 days for EVAR and OSR respectively). Normal prior distributions with means of zero and standard deviations of 0.50 truncated to be positive were used for the between-study standard deviations (σ_m, τ_m).

Summary of the posterior distribution and posterior predictive distribution for the log odds of dying for EVAR and OSR from the Bayesian meta-analysis

Variable	Parameter	Mean	Standard deviation
30 day mortality			
EVAR	θ_{30}	-3.83	0.4852
	ψ_{30}.new	-3.83	0.7058
OSR	α_{30}	-3.00	0.3678
	γ_{30}.new	-3.00	0.722
One year mortality conditional on being alive at 30 days			
Two high risk studies			
EVAR	θ_1	-3.73	0.9317
	ψ_1.new	-3.73	1.113
OSR	α_1	-3.90	0.9121
	γ_1.new	-3.90	1.066
Eight mixed risk studies			
EVAR	θ_1	-3.14	0.2386
	ψ_1.new	-3.14	0.467
OSR	α_1	-3.98	0.4433
	γ_1.new	-3.98	0.9845

After combining the studies to generate estimates for the mean log odds in both the EVAR and OSR groups for 30 day mortality (i.e., ψ_{30}.new and γ_{30}.new) and one year mortality conditional on being alive at 30 days (i.e., ψ_1.new and γ_1.new) , the corresponding probabilities were derived by exponentiating the results. The resulting values were used to estimate the probabilities of dying for EVAR and OSR between zero and 30 days post-operative (i.e., $p_{deadEVAR30}$ and $p_{deadOSR30}$) and the conditional probabilities used to estimate mortality after 30 days and up to one year (i.e., $p_{deadEVAR1|aliveEVAR30}$ and $p_{deadOSR1|aliveOSR30}$).

In order to convert these probabilities into life years, we assumed death was equally likely to occur at any time within the respective time periods. As a result we assumed mean life years of approximately 0.04 (i.e., 15/365) for patients who died between zero and 30 days and mean life years of approximately 0.46 (i.e., 168/365) for patients who died after 30 days and up to one year after treatment. Life years of one were applied to those patients still alive at one year. The probability of being alive at one year for EVAR ($p_{aliveEVAR1}$) and OSR ($p_{aliveOSR1}$) was calculated as one minus the respective probabilities of being dead by one year in each of the groups. That is,

$p_{aliveEVAR1} = 1 - p_{deadEVAR30} - (p_{deadEVAR1|aliveEVAR30} \times (1-p_{deadEVAR30}))$ and

$p_{aliveOSR1} = 1 - p_{deadOSR30} - (p_{deadOSR1|aliveOSR30} \times (1-p_{deadOSR30}))$.

Based on these assumptions the following equations were used to estimate mean life years in the trial at one year for EVAR and OSR respectively:

$\mu_{E1} = (p_{aliveEVAR1} \times 1) + (p_{deadEVAR30} \times (15/365)) + (p_{deadEVAR1|aliveEVAR30} \times (1-p_{deadEVAR30}) \times (168/365))$

$\mu_{E2} = (p_{aliveOSR1} \times 1) + (p_{deadOSR30} \times (15/365)) + (p_{deadOSR1|aliveOSR30} \times (1-p_{deadOSR30}) \times (168/365))$.

Summary of the posterior predictive distribution for mean life years for EVAR and OSR from the Bayesian meta-analysis

Variable	Parameter	Mean	Standard deviation
Two high risk studies			
EVAR	μ_{E1}	0.9548	0.03179
OSR	μ_{E2}	0.9281	0.04704
Eight mixed risk studies			
EVAR	μ_{E1}	0.951	0.02176
OSR	μ_{E2}	0.9301	0.04601

The mean and standard deviation of μ_{Ej}, the predictive value of mean life years, are used as the parameters of the Normal ‘face value’ prior distributions for mean life years in the trial. The mean of the posterior predictive distribution, $\mu_{Ej,}$ and a variance four times the variance of the patient level data are used for the ‘sceptical’ prior. The standard deviations for mean life years in the patient level data were 0.01447 and 0.04678 for EVAR and OSR respectively.

CHAPTER 5

CONCLUSIONS

The economic evaluation of health technologies plays an important role in informed health care decision making. Similarly, the associated methodological issues and challenges offer important opportunities to advance knowledge in the field of HTA by providing new insights and approaches. This thesis has addressed issues related to synthesising evidence on treatment effects from different sources of information. Specifically, we have conducted research on combining evidence from randomised and non-randomised studies and combining patient level trial data with additional evidence from the literature. This final chapter offers a summary of the findings of the thesis as well as identifying potential areas for future research. The implications and contributions of the thesis research are also discussed.

In Chapter 2 [13], we proposed a new approach for combining evidence from randomised and non-randomised studies and adjusting for covariate imbalances that could bias the results. This approach involved Bayesian hierarchical modelling and adjustment using differences in patient characteristics between study arms. The analysis compared the proposed approach to four other Bayesian hierarchical methods in the context of a case study comparing EVAR and OSR for the treatment of abdominal aortic aneurysms. As the case study provided evidence of greater imbalance among the non-randomised studies relative to the randomised studies, the methods were assessed in terms of whether they moved the estimated odds ratios from the less balanced non-randomised studies towards the more balanced randomised studies [9]. The closer the estimated overall odds ratio was to the odds ratio for the randomised studies alone, the better the model was deemed to have accounted for covariate imbalances. Based on the data in this case study, we concluded that the proposed Bayesian hierarchical model adjusted for differences in patient characteristics between study arms was capable of accounting for imbalances that could otherwise bias the results. As the synthesis of evidence from randomised and non-randomised studies likely necessitates some concession as to the possibility of imbalances, this research offers a potential approach that would allow both randomised and non-randomised studies to contribute usefully to the estimation of treatment effects. Though in order to do this, the work also points to the need for researchers to better report covariate information so as to facilitate adjustment for future evidence synthesis. As the results were based on a single case study, however, it was unclear how they might be affected by changes in factors such as the impact of the imbalances or the relative number or size of the randomised and non-randomised studies.

To assess the influence of these factors on the performance of the proposed Bayesian method adjusted for differences in patient characteristics between study arms, we conducted a simulation study in Chapter 3 [14]. Six scenarios involving changes in the impact of the imbalances and the relative number and size of studies of each type were examined. Unlike in the case study in Chapter 2 [13], because the truth is known in a simulation study, model performance was assessed relative to this known truth [15]. Across all six scenarios the Bayesian hierarchical model adjusted for differences within studies gave results that were closest to the true value compared to the other models. These results reinforced those observed in the single applied case study of EVAR and OSR in Chapter 2 [13], where the new method moved the estimated odds ratios from the less balanced non-randomised studies towards the more balanced randomised studies. Furthermore, this simulation study demonstrated that when bias can be explained by covariate imbalances between study arms, the proposed Bayesian approach handles this problem in a way that appears robust to changes in the underlying characteristics of the studies. In so doing the model adjusted for differences outperformed the other models. This research adds support to the proposed approach in terms of its ability to adjust for covariate imbalances when combining randomised and non-randomised studies, thus strengthening the assertion that non-randomised studies can contribute usefully to the estimation of treatment effects.

The proposed method for synthesising evidence on effects from randomised and non-randomised studies, that was the focus of Chapters 2 [13] and 3 [14], was based on taking weighted averages of the two sources of information, where the weights were determined implicitly. In Chapter 4 [16], we investigated the informative prior approach [12]. This approach is based on an explicit weighting of the evidence sources. As seen in the preceding chapters this allows for the downweighting of one source of evidence relative to the other source. The amount by which to downweight the evidence is not an automatic process, but rather requires careful consideration. In Chapter 4 [16], we explored the use of the informative prior approach to combine evidence on effects from an existing patient level economic evaluation comparing EVAR and OSR in high risk patients [8] with additional evidence from the literature [7]. Combining the additional evidence with the patient level data resulted in higher estimated mean life years in the OSR group compared to those based on the patient level data alone while the estimates for EVAR were unchanged. Consequently, the incremental estimates decreased and uncertainty regarding the cost-effectiveness of EVAR increased. The more the evidence from the literature was downweighted the closer the results were to those from the original economic evaluation. This research provides an applied example of the potential importance of synthesising evidence from all available sources. Using an actual economic evaluation that was used to inform decision making regarding reimbursement of EVAR in high risk patients in the province of Ontario, this chapter outlines how additional evidence might have been incorporated. However, a key limitation, especially from a policy making perspective, is the choice of inflation factor used to downweight the evidence from the literature. A more thorough consideration of the relative differences between the literature studies and the patient level data would be required when conducting such an analysis in practice. Again, however, these assessments would rely on the necessary data being sufficiently reported.

Underpinning the research in this dissertation is the question, how do the sources of evidence differ? Chapters 2 [13], 3 [14], and 4 [16] have each addressed this question in trying to combine randomised studies with non-randomised studies and patient level data with study level evidence from the literature. By examining how sources of evidence differ we can gain insight into how best to combine them together. Future research could consider issues related to further refinements and practical applications of the work conducted in this thesis. For example, how to downweight as well as adjust for imbalanced studies within the Bayesian hierarchical model proposed in Chapter 2 [13]. While we have used an informative prior approach for combining patient level data with additional evidence in Chapter 4 [16], hierarchical modelling could also be explored. Within such an approach, and with appropriate data, adjustment for imbalances might also be possible. Methods that explicitly downweight sources of evidence also pose interesting challenges for future research. In particular, how to quantify differences between sources of evidence.

The research presented in this thesis has important implications for the future of health care decision making. The establishment of the PRUFE framework [6], whereby information from both systematic reviews and patient level field evaluations can be combined, attests to the importance of a comprehensive approach to evidence based decision making. This thesis focussed on the development and use of methods capable of combining evidence on effects from different sources, an essential element of any comprehensive approach. Despite limitations, this thesis research provides insights and ideas as well as practical examples of how to address some of the challenges faced in evidence synthesis. Their potential to combine all available evidence makes Bayesian methods ideal for the kind of comprehensive approach envisioned by the PRUFE framework. As health care is of such vital importance both individually and collectively, the evidence upon which decisions are based must be carefully considered.

Taking advantage of all available evidence has the potential either to reinforce belief in the effect of one intervention compared to another, or to introduce a healthy questioning of that belief. The result of which could be to expedite decisions, or, where necessary, to allow for a careful reconsidering of the evidence. Regardless of the outcome, the consequence would be a more prudent approach to health care decision making, defined by a comprehensive evidence base. Evidence synthesis could also contribute to informing decisions about the need for additional research, thus helping to ensure research dollars are used efficiently. If new ideas are required to address the challenges facing, for example, the Canadian health care system [17], then let comprehensive evidence based decision making be at the core of these new initiatives. Let decision makers know what is working and what is not and let them have the flexibility to respond accordingly.

Health care decision making is not static; it must be responsive and flexible. As has been demonstrated by the research conducted in this thesis, the potential exists within a Bayesian approach for these goals to be achieved as well as allowing for the necessary comprehensiveness. By addressing important issues such as how to deal with possible imbalances when combining randomised and non-randomised studies and how to integrate external evidence when analysing patient level economic data, we have contributed to the further development and application of Bayesian methods in HTA. In helping to move the methodology forward, the preceding chapters of the thesis form a coherent and substantial body of work that contributes significantly to the advancement of knowledge in the field of HTA.

REFERENCES

1. International Network of Agencies for Health Technology Assessment [http://www.inahta.org/HTA]

2. Health Technology Assessment International [http://www.htai.org/index.php?id=420]

3. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes*. 3rd ed. Oxford: Oxford University Press, 2005.

4. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, Lu G. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics*. 2006; 24(1):1-19.

5. Ashby D. Bayesian Statistics in Medicine: A 25 Year Review. *Stat Med.* 2006; 25: 3589-3631.

6. Goeree R, Levin L. Building bridges between academic research and policy formulation the PRUFE framework- and integral part of Ontario's evidence based HTPA process. *Pharmacoeconomics*. 2006; 24(11):1143-1156.

7. Hopkins R, Bowen J, Campbell K, Blackhouse G, De Rose G, Novick T, O'Reilly D, Goeree R, Tarride JE. Effects of study design and trends for EVAR versus OSR. *Vasc Health Risk Manag*. 2008; 4(5): 1011-1010.

8. Tarride JE, Blackhouse G, De Rose G, Novick T, Bowen JM, Hopkins R, O'Reilly D, Goeree R. Cost-effectiveness analysis of elective endovascular repair compared with open surgical repair of abdominal aortic aneurysms for patients at a high surgical risk: A

1- year patient level analysis conducted in Ontario, Canada. *J Vasc Surg*. 2008; 48(4):779-787.

9. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003; **7**(27):1-173.

10. McCarron CE, Pullenayegum EM, Marshall DA, Goeree R, Tarride JE. Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments. *Int J Technol Assess Health Care*. 2009; 25(4): 546-554.

11. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Stat Med*. 2000; 19:3359-3376.

12. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res*. 2001; 10(4): 277-303.

13. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Med Res Methodol*. 2010; 10:64.

14. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *Accepted September 8, 2011 for publication in PLoS ONE*.

15. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006; 25: 4279-92.

16. McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride JE. The impact of using informative priors in a Bayesian cost-effectiveness analysis: an application of endovascular versus open surgical repair for abdominal aortic aneurysms in high risk patients. *Submitted October 3, 2011 to Medical Decision Making*.

17. Smith J. New ideas on health needed, doctors warn. *Toronto Star*, Monday, August 22, 2011.

APPENDIX

Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments

C Elizabeth McCarron[a,b,§], Eleanor M Pullenayegum[a,c], Deborah A Marshall[d,e], Ron Goeree[a,b], Jean-Eric Tarride[a,b]

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

[b]Programs for Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[c]Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[d]University of Calgary, Calgary, Alberta, Canada

[e]Alberta Bone Joint Health Institute, Calgary, Alberta, Canada

[§]Corresponding author

Citation: McCarron CE, Pullenayegum EM, Marshall DA, Goeree R, Tarride JE. Handling uncertainty in economic evaluations of patient level data: A review of the use of Bayesian methods to inform health technology assessments. *International Journal of Technology Assessment in Health Care*. 2009; 25(4): 546-554.

ABSTRACT AND KEYWORDS

Objectives: Due to potential advantages (e.g., using all available evidence), Bayesian methods have been proposed to assist health care decision making. This review provides a detailed description of how Bayesian methods have been applied to economic evaluations of patient level data. The results serve both as a reference and as a means by which to examine the appropriate application of Bayesian methods to inform decision making.

Methods: MEDLINE, EMBASE, and Cochrane Economic Evaluation databases were searched to identify studies, published up to November 2007, meeting three inclusion criteria: 1) the study conducted an economic evaluation; 2) sampling uncertainty was incorporated using Bayesian methods; 3) the likelihood function was informed by patient level data from a single source. Data were collected on key study characteristics (e.g., prior distribution, likelihood function, presentation of uncertainty).

Results: The search identified 366 potentially relevant studies, from which 103 studies underwent full-text review. Sixteen studies met the inclusion criteria. Half of the studies used uninformative priors; most studies incorporated the potential dependence between costs and effects, and presented cost-effectiveness acceptability curves. Results were sensitive to changes in the priors and likelihoods.

Conclusions: Limited use of informative priors, among the included studies, gives policy makers little guidance on one of the main benefits of Bayesian methods, the ability to integrate all available evidence to capture the uncertainty inherent in decision making.

Keywords: Bayes Theorem, Cost-Benefit Analysis, Uncertainty

INTRODUCTION

Economic evaluation in health care can be defined as the comparison of alternative options in terms of their costs and consequences.(6) The purpose being to inform the efficient allocation of scarce resources.(5) Two main approaches exist, those using patient level data and those using decision analytic modelling.

When patient level data are used, economic outcomes are the result of a single sample drawn from the population.(19) However, decisions are made at the population level. Consequently, uncertainty arises from using limited samples to estimate the true (population) value of costs and effects. This source of uncertainty can be referred to as sampling variation.(13) Two methods have been used regularly in the applied literature to incorporate sampling variation: the nonparametric bootstrap method, and Fieller's method.(6) Both methods propagate uncertainty using only the information contained in the original data.

The last 25 years have seen an increase in the prevalence of Bayesian statistics.(2) In particular, the Bayesian Initiative in Health Economics & Outcomes Research was established, "to explore the extent to which formal Bayesian statistical analysis can and should be incorporated into the field of health economics and outcomes research for the purpose of assisting rational health care decision making."(15)

Under a Bayesian interpretation, parameters of interest are ascribed a distribution reflecting uncertainty concerning the true value of the parameter.(4) A Bayesian analysis synthesises two sources of information about the unknown parameters of interest. One source is the prior distribution, which represents information that is available prior to (or, more generally, in addition to) the data (e.g., previous trials, literature, expert opinion). In the absence of prior information, vague or uninformative prior distributions can be used. The less informative the prior, the more weight is given to the data in the analysis. The other source of information is the data, which contribute to the analysis through the likelihood function.(23) The likelihood summarises all of the information about the unknown parameters that is contained in the data.(22) These two sources of information are combined through the use of Bayes' theorem. Bayes' theorem updates the prior information by taking into account, via the likelihood, the newly observed data. The result is a posterior distribution that represents what is now known about the unknown parameters based both on the data and the prior information.(23) Posterior distributions can be generated using simulation techniques such as Markov Chain Monte Carlo.

The inferential outputs from a Bayesian analysis and the ability to make direct probability statements regarding unknown quantities provide a natural way of informing policy makers. The ability to take into account all available evidence, through the combination of the prior and the likelihood, speaks to another potential advantage. By focussing on the vital question: how does this new piece of evidence change what we currently believe? Bayesian methods present a more iterative approach to evaluation because prior beliefs can be updated as new evidence becomes available.(22) Another potential advantage is the use of a likelihood function to model the underlying distribution of the data. (17)

In the wake of a renewed interest in Bayesian statistics, the primary objective of this review is to describe how Bayesian methods have been used to handle uncertainty due to sampling variation in patient level economic evaluations. The results serve as a reference, detailing how these methods have been used to evaluate health care interventions. Specifically, the review focuses on describing the priors, the likelihoods, the presentation of uncertainty, and sensitivity analyses in these studies. Concentrating on these aspects gives a sense of how Bayesian methods have been used to incorporate additional information, accurately model the data, communicate the impact of uncertainty, and assess the robustness of the results. Findings and implications from the review are discussed.

METHODS

Data Sources and Search Strategy

We conducted a comprehensive search strategy to identify all relevant published Bayesian analyses (to the second week of November 2007). We developed the search strategy in MEDLINE and modified it for other databases. Only articles in English were considered. Ovid MEDLINE In-Process & Other Non-Indexed Citations and Ovid

MEDLINE (1950 to Present), EMBASE (1980 to 2007 Week 45), and Cochrane Library NHS Economic Evaluation (Issue 4, 2007) databases were searched. In addition, we searched the reference sections of relevant papers for potentially eligible studies.

Search terms were derived based on mapping keywords for Bayesian analysis (e.g., Bayesian, WinBUGS) and economic evaluation (e.g., cost, economic) to indexed subject headings within the respective databases. Terms were also derived based on investigator-nominated terms and keywords from the titles and abstracts of potentially relevant studies. Relevant keywords and subject headings were then combined allowing for alternative spellings and suffixes. Operators denoting the proximity of various search terms in relation to others were also used in order to derive a comprehensive retrieval strategy. The search strategy is provided in Supplementary Table 1 (www.journals.cambridge.org/thc).

Study Selection

We screened citation records in two stages. In the first stage, the titles and abstracts of retrieved articles were screened for potential inclusion or exclusion. In the second stage, those records not excluded at the first stage underwent a full-text review. Included studies met the following criteria: 1) the study conducted an economic evaluation comparing two or more health care interventions; 2) the impact of uncertainty (sampling variation) on the results of the economic evaluation was incorporated using Bayesian methods; and 3) the likelihood function was informed by patient level data from a single source (trial, study, etc.).

Excluded studies involved only patient level costs or only patient level effects, incorporated any sort of decision analytic modeling, or used Bayesian methods for purposes other than the incorporation and assessment of sampling variation (e.g., evidence synthesis, value of information analysis, heterogeneity). In both stages, a single reviewer (CEM) selected articles for inclusion.

Data Synthesis and Analysis

In the context of the current analysis a descriptive synthesis of the included studies was undertaken. An abstraction form was developed to collect information on key study characteristics. To get a sense of how Bayesian methods were used to combine additional information with the data, the type of prior distribution was recorded. To illustrate how the underlying data were modeled and whether these distributions allowed for issues such as the potential dependence between costs and effects or skewness in costs, information was collected on the likelihood functions. To understand how Bayesian methods were used to inform decision makers, the presentation of uncertainty was documented. Attention was also given to whether the studies explored the sensitivity of the results to changes in the priors and the likelihoods, as this could have implications for the results. The data were then synthesized to provide an overall description of the use of Bayesian methods to handle uncertainty in economic evaluations of patient level data.

RESULTS

Literature Review

The literature search yielded 366 potentially relevant bibliographic records. From the 366 citations, 103 articles were retrieved for relevance assessment. The selection of included studies is presented in the QUORUM diagram given in Figure 1. Sixteen studies met the final inclusion criteria.(1,3,4,7,9-12,14,16-18,20,26-28) Thirteen of these studies were classified as methodological papers with applications (1,3,4,9-12,14,16-18,27,28) and three were classified as application papers (7,20,26). For the purpose of this review the former classification pertains to those papers that used applications merely for illustrative or pedagogic purposes. The latter refers to those papers whose primary objective was an economic evaluation, where Bayesian methods were used to incorporate sampling uncertainty. Supplementary Table 2 (www.journals.cambridge.org/thc) describes the included studies.

Assessment of Bayesian Methods

Prior Distributions

The most common type of prior used for either costs or effects was a vague or uninformative prior. Uninformative priors for costs were used in 14 studies (1,3,4,7,9,11,12,14,16-18,20,26,27) and for effects in 13 studies (1,3,4,7,9,12,14,16-18,20,26,27). These priors were incorporated either exclusively (3,7,9,12,14,17,26,27), or as part of a sensitivity analysis (1,4,11,16,18,20). Informative priors (empirical, subjective, or structural) were included in half of the studies (1,4,10,11,16,18,20,28). Priors based on empirical data were used for effects in five studies (1,4,10,11,20) and for costs in three studies (1,4,20). Data sources for the empirical priors included previous trials (1), pilot studies (10,11), the literature (4), and individual Medicare claims data (20).

Priors based on subjective opinion were applied equally to costs (1,10,11,28) and effects (1,16,18,28). Subjective priors most often reflected informal reasoning.(1,10,11,16,18) However, one study (28) referred to a process of eliciting expert opinion. Experts who participated in the study were asked about the mean and the probability interval to obtain the prior mean and variance of the parameters of interest. Structural priors, denoting the relative relationship between parameters as opposed to the actual numerical values, appeared in one of the studies. (16) In this case, the prior represented the belief that the variances of costs should not be too different between patient groups. The effect of this prior information was to moderate the influence of the extreme costs. Table 1 describes the prior distributions.

Likelihood Functions

Two of the applied studies (7,26) did not specify the distributional form of their likelihood functions, and one of the methodological papers (3) analyzed the individual level data using two different approaches. Therefore, there are 13 examples where costs and effects are modeled directly (1,3,4,9-12,14,16-18,27,28), and two examples using regression-based modeling of net benefits (3,20).

The majority of studies incorporated the potential dependence between the cost and effect data. This was achieved through the use of both multivariate normal distributions and regression analysis. Two of the studies (27,28) that applied regression analysis directly to costs and effects included covariates in their likelihood functions and assessed the resulting impact on uncertainty.

For three of the studies (1,4,12), the use of multivariate normal distributions was based on large sample approximations for the means of costs and effects. Where the likelihood functions allowed for a specific relationship between the cost and effect data, costs most often depended on effects: four studies (10,11,16,27) allowed costs to depend on effects whereas only one study (3) allowed effects to depend on costs. Different distributions were used for effects based on whether the outcome measure was a continuous or discrete random variable.

Six studies (3,9,10,11,14,16) incorporated the potential skewness in the cost data. Three of these studies used gamma distributions (3,10,11) and two (14,16) used lognormal distributions. One study (9) divided total cost into three components and applied distributions (e.g., lognormal) to each of the cost components. Table 2 describes the likelihood functions.

Presentation of Uncertainty

The predominant approaches to the presentation of uncertainty were Bayesian 95% credibility intervals (0.95 posterior probability that the true value lies in the interval), and cost-effectiveness acceptability curves (CEAC) (posterior probability that the intervention is cost-effective given the data and willingness to pay). Almost all of the studies presented cost-effectiveness acceptability curves.(1,3,4,7,9-11,16-18,20,26-28) Six studies (1,3,10,11,12,27) presented Bayesian 95% credibility intervals for the incremental cost-effectiveness ratio (ICER), two studies (3,10) presented credibility intervals for the incremental net monetary benefit (INMB), and one study (11) presented credibility intervals for the incremental net health benefit (INHB). Another study (14) that compared multiple treatment options and incorporated two measures of effectiveness proposed the cost-effectiveness acceptability plane frontier (CEAPF) as an alternative to the cost-effectiveness acceptability curve. Table 3 summarizes the presentation of uncertainty in each study.

Sensitivity Analysis

In addition to assessing the impact of sampling variation on the results, 10 studies (1,3,4,9,11,16,18,20,27,28) considered the sensitivity of the results to changes in the prior distributions and the likelihood functions. Of those studies, four (1,4,11,16) used different priors, four (3,9,27,28) used different likelihoods, and two (18,20) changed both the priors and the likelihoods. Table 3 describes the sensitivity analyses that were conducted. The following summarizes the findings of those studies.

Priors: The study by Al and Van Hout (1) assessed the sensitivity of the results to three different prior distributions for costs and effects: an uninformative prior disregarding all information from a previous trial, an empirical prior equal to the posterior of the previous trial, and a subjective prior that uses only 50% of the information from the previous trial. The authors concluded that different prior distributions may lead to different decisions. For example, given a specific willingness to pay, the probability of cost-effectiveness was 0.65 for the uninformative prior, 0.80 for the subjective prior, and 0.90 for the empirical prior.

Another study (16) that assessed the impact of different priors found that varying the prior information on effects made negligible difference to conclusions, since the data quite strongly indicated an improvement in effectiveness. However, the impact of the prior information on costs was much more substantial. For smaller willingness to pay values, where cost is a real consideration, the different priors produced quite different probabilities of cost-effectiveness. When weak prior information was used, the probability of cost-effectiveness never went below 0.70. When structural prior information was used, the probability of cost-effectiveness went from 0.45 to 0.65 for smaller willingness to pay values. The authors argued that this difference was primarily being driven by two outlying observations. The use of structural prior information, representing the belief that the variances of costs should not be too different between patient groups, effectively mitigated the impact of the outliers and resulted in a correspondingly lower cost-effectiveness acceptability curve for small willingness to pay values. In the remaining studies (4,11) the priors did not appear to be a source of sensitivity.

Likelihoods: Hahn and Whitehead (9) compared five different likelihood functions for the cost and effect data. For two of the likelihoods only one cost was considered, namely total cost. In the other three, the total cost was broken down into three components. Four of the likelihoods used normal or multivariate normal distributions for the cost and effect data. The remaining likelihood used other distributions (e.g., lognormal) for the cost components. The cost-effectiveness acceptability curve associated with this likelihood was different from those based on the other four likelihoods. In particular, the willingness to pay value for which the probability of cost-effectiveness is 0.50 was greater than that suggested when the other likelihoods were used.

In the regression framework presented in Vazquez-Polo, Negrin, Badia and Roset (28) the authors assessed the sensitivity of the results to the inclusion of covariates, first using a continuous outcome and then a binary outcome. When the continuous outcome was used the willingness to pay value at which the probability of cost-effectiveness is 0.50 was approximately 75% greater without covariates than when covariates were included. When the binary measure of effect was used the control treatment dominated the new treatment. Similar results were found in the other study by Vazquez-Polo, Hernandez and Lopez-Valcarcel (27) that used only a single continuous outcome. In this study the cost-effectiveness acceptability curve was also higher when covariates were included in the likelihood. The study by Bachmann, Fairall, Clark and Mugford (3) compared the joint modeling of costs and effects using a binomial-gamma likelihood and a regression-based model of net benefits. Both likelihood functions produced similar results;

however, the point estimate for the incremental cost-effectiveness ratio was approximately 20% higher for the binomial-gamma likelihood.

Priors and Likelihoods: One study (20) examined the cost-effectiveness impact of generic drug entry using two approaches to Bayesian net benefit regression analysis. One approach pooled the data from the pre and post-entry periods and used uninformative priors to estimate the regression parameters. The second approach proceeded in two steps. In the first step, the authors assumed uninformative priors for the regression parameters and updated these with data from the pre-entry period. In the second step, the authors used the posterior distributions generated in the first step as empirical priors for the regression parameters. Information from the post-entry period formed the likelihood data and was used to update the parameter values. At a willingness to pay of $US5000, the probabilities of cost-effectiveness for the four non-generic drugs were 96.7%, 77.6%, 96.3%, and 97.0%, respectively, in the pre-entry period in the pooled analysis. These probabilities reduced to 36.7%, 62.7%, 33.0%, and 60.1%, respectively, in the post-entry period. The probabilities became 94.1%, 71.9%, 89.1%, and 92.1% in the analysis using the pre-entry data as a prior to update the post-entry data.

In O'Hagan, Stevens and Montmartin (18), the only substantial aspect of prior information was in regard to the true mean effect. When an informative prior was used for the effect measure the cost-effectiveness acceptability curve was uniformly higher than when a weak prior was used. On the basis of the weak prior the uncertainty associated with the decision was much greater, although in both cases the probability of cost-effectiveness was greater than 0.50 for all willingness to pay values. To test the robustness of the conclusions to changes in the likelihood, the authors replaced the assumption of normally distributed costs with lognormal distributions. Quite substantial differences in the cost-effectiveness acceptability curve were observed, especially for small willingness to pay values. While the probability of cost-effectiveness still exceeded 0.70 for almost all willingness to pay values, it never went beyond the level of 0.90 that was reached when the informative prior was used.

The sensitivity of the results to changes in the priors and the likelihoods is discussed in terms of changes in the probability of cost-effectiveness, as represented by the cost-effectiveness acceptability curve. This measure is chosen based on the frequency of its use among the studies as well as the relevancy of the information it imparts to decision makers. However, when considering the impact of using a more informative prior distribution, estimates of the mean difference in costs and effects might be more revealing. In general you would expect the probability of cost-effectiveness to change when using an informative prior, even if the point estimates of the mean differences in costs and effects stayed exactly the same, because the probability of cost effectiveness is a function of both the point estimates and the uncertainties. Al and Van Hout (1) reported posterior mean differences in costs and effects of NLG 2149 and 0.098, NLG 2567 and 0.137, and NLG 2564 and 0.158 (NLG = Netherland guilders), for increasingly informative priors. The only other study to do so was O'Hagan, Stevens and Montmartin

(18) which presented posterior estimates of the mean differences in costs and effects of -£574 and 1.24 (weak prior), and -£626 and 2.03 (informative prior). From the perspective of a decision maker, the issue therefore becomes one of whether the primary impact of the more informative prior is to reduce uncertainty, or if it actually changes the estimated differences in costs and effects in such a way as to alter the relative cost-effectiveness.

DISCUSSION AND LIMITATIONS

The Bayesian approach allows for the ability to accurately model the data and to incorporate additional information, in the form of prior distributions. The use of priors based on previous data may be less susceptible to accusations of subjectivity than opinion based priors, but they may also fail to subscribe to the notion of a fully Bayesian analysis. Some Bayesians would argue that such an approach is not in fact Bayesian at all since no subjective beliefs are employed.(4) Despite the use of more informative priors among some of the included studies, the most common type of prior found in this review remains the vague or uninformative prior. This may reflect a deliberate attempt to give more weight to the data in the analysis. However, if prior information exists, the use of uninformative priors seemingly negates a fundamental feature of the Bayesian approach. The ability to incorporate genuine prior information in addition to the data in the final analysis is compromised when uninformative priors are used.(23)

The rationale for choosing certain likelihoods and priors reflects the need to accurately model the data and to include all relevant prior information in the analysis. Likelihood functions, chosen based on the need to accommodate specific characteristics of the data (e.g., skewness, dependence), together with prior distributions, are intended to represent the totality of available evidence. Where the studies gave a reason for using uninformative priors (1,11,14,16,18,20,26), most stated a lack of genuine prior information. However, one study (7) commented, "vague priors ensured that the trial results had a larger influence upon the analysis than the prior beliefs." Reasons for using informative priors included the presence of preceding trial or study results, which though the populations might differ, were viewed as being informative. The study by Shih, Bekele, and Xu (20) justified their use of prior information on the basis of preserving some of the original cost-effectiveness information in decision making.

In a Bayesian analysis of patient level data, one would assume that any estimate of the impact of sampling variation would be conditional on both the prior distribution and the likelihood function. The sensitivity of the results to changes in the priors and the likelihoods was considered in 10 of the included studies.(1,3,4,9,11,16,18,20,27,28) The results suggest that a failure to include sensitivity analysis could affect the estimated uncertainty and potentially lead to inappropriate inferences. Several authors (22,24,25) have recommended the use of sensitivity analysis when reporting the results of Bayesian analyses.

The results of this review are intended to provide, for the first time, a comprehensive description of the use of Bayesian methods to handle uncertainty due to sampling variation in patient level economic evaluations. The review was limited to published studies identified from three databases and relied on a single reviewer. However, the search strategy covered the largest databases and was designed in consultation with a trained research librarian. The review was limited to patient level economic evaluations using information from a single source and did not consider decision analytic models using several data sources (e.g., Fryback, Chinnis and Ulvila (8)).

Despite these limitations, we believe that this review serves as a reference to those engaged in, or considering Bayesian analysis of patient level data. The decision to use Bayesian methods, rather than more traditional approaches, requires consideration of the relative advantages and disadvantages, in terms of informing health care policy decisions.

Potential disadvantages of Bayesian methods in health care evaluation include a lack of expertise, difficulty specifying and potential subjectivity of priors, as well as the additional complexity.(22) Future research on the choice and elicitation of prior distributions in practical applications would seem critical to ensuring the ability of the Bayesian approach to synthesize all available evidence is fully exploited.

POLICY IMPLICATIONS

To the extent that important health policy decisions are informed by the results of economic evaluations, and that these results are subject to uncertainty, a comprehensive and robust approach is required. This would include the use of all relevant evidence to inform decision makers. The ability to combine informative priors with the data, as well as providing a natural way of handling uncertainty, suggests Bayesian methods may offer certain advantages over traditional methods.

REFERENCES

1. Al MJ, Van Hout BA. A Bayesian approach to economic analyses of clinical trials: the case of stenting versus balloon angioplasty. *Health Econ*. 2000;9(7):599-609.
2. Ashby D. Bayesian Statistics in Medicine: A 25 Year Review. *Stat Med.* 2006;25:3589-3631.
3. Bachmann MO, Fairall L, Clark A, Mugford M. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Eff Resour Alloc*. 2007;5:12.
4. Briggs AH. A Bayesian Approach to Stochastic Cost-Effectiveness Analysis: An Illustration and Application to Blood Pressure Control in Type 2 Diabetes. *Int J Technol Assess Health Care.* 2001;17(1):69-82.
5. Briggs AH. Handling Uncertainty in Cost-Effectiveness Models. *Pharmacoeconomics*. 2000;17(5): 479-500.
6. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes.* 3rd ed. Oxford: Oxford University Press; 2005.
7. Fenwick E, Wilson J, Sculpher M, Claxton K. Pre-operative optimisation employing dopexamine or adrenaline for patients undergoing major elective surgery: a cost-effectiveness analysis. *Intensive Care Med*. 2002;28(5):599-608.

8. Fryback DG, Chinnis JO, Ulvila JW. Bayesian Cost-Effectiveness Analysis: An Example Using the GUSTO Trial. *Int J Technol Assess Health Care.* 2001;17(1):83-97.

9. Hahn S, Whitehead A. An illustration of the modelling of cost and efficacy data from a clinical trial. *Stat Med*. 2003;22(6):1009-1024.

10. Heitjan DF, Kim CY, Li H. Bayesian estimation of cost-effectiveness from censored data. *Stat Med*. 2004;23(8):1297-1309.

11. Heitjan DF, Li H. Bayesian estimation of cost-effectiveness: An importance-sampling approach. *Health Econ*. 2004;13(2):191-198.

12. Heitjan DF, Moskowitz AJ, Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Econ*. 1999;8 (3):191-201.

13. Johnson-Masotti AP, Laud PW, Hoffmann RG, Hayat MJ, Pinkerton SD. Probabilistic Cost-Effectiveness Analysis of HIV Prevention Comparing a Bayesian Approach with Traditional Deterministic Sensitivity Analysis. *Eval Rev*. 2001; 25(4):474-502.

14. Negrin MA, Vazquez-Polo FJ. Bayesian cost-effectiveness analysis with two measures of effectiveness: The cost-effectiveness acceptability plane. *Health Econ*. 2006;15(4):363-372.

15. O'Hagan A, Luce BR. *A Primer on Bayesian Statistics in Health Economics and Outcomes Research*. Bayesian Initiative in Health Economics & Outcomes Research Centre for Bayesian Statistics in Health Economics; 2003.

16. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. *Health Econ*. 2001;10(4):303-315.
17. O'Hagan A, Stevens JW. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Stat Methods Med Res*. 2002;11(6):469-490.
18. O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. *Stat Med*. 2001;20(5):733-753.
19. Ramsey S, Willke R, Briggs A et al. Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report. *Value Health*. 2005;8(5):521-533.
20. Shih YC, Bekele NB, Xu Y. Use of Bayesian net benefit regression model to examine the impact of generic drug entry on the cost effectiveness of selective serotonin reuptake inhibitors in elderly depressed patients. *Pharmacoeconomics*. 2007;25(10):843-862.
21. Skrepnek GH. The Contrast and Convergence of Bayesian and Frequentist Statistical Approaches in Pharmacoeconomic Analysis. *Pharmacoeconomics.* 2007;25(8):649-664.
22. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, West Sussex: John Wiley & Sons Ltd; 2004.

23. Stevens JW, O'Hagan A. Incorporation of Genuine Prior Information in Cost-Effectiveness Analysis of Clinical Trial Data. *Int J Technol Assess Health Care*. 2002;18(4):782-790.

24. Sung L, Hayden J, Greenberg ML et al. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *J Clin Epidemiol*. 2005;58:261-268.

25. The BaSiS Group. *Bayesian standards in science (BaSiS)*; 2001. http://lib.stat.cmu.edu/bayesworkshop/2001/BaSis.html.

26. UK BEAM Trial Team. United Kingdom back pain exercise and manipulation (UKBEAM) randomised trial: cost effectiveness of physical treatments for back pain in primary care. *BMJ*. 2004;329:1381-1385.

27. Vazquez-Polo FJ, Hernandez MAN, Lopez-Valcarcel BG. Using covariates to reduce uncertainty in the economic evaluation of clinical trial data. *Health Econ*. 2005;14(6):545-557.

28. Vazquez-Polo FJ, Negrin M, Badia X, Roset M. Bayesian regression models for cost-effectiveness analysis. *Eur J Health Econ*. 2005;6(1):45-52.

Table 1. Description of Priors for Effects and Costs

Author(s)	Type of prior(s)[1,2,3,4]	
	Effects	Costs
Methodological Papers with Applications		
Heitjan 1999	Uninformative	Uninformative
Al 2000	Uninformative Empirical Subjective	Uninformative Empirical Subjective
Briggs 2001	Uninformative Empirical	Uninformative Empirical
O'Hagan 2001a	Subjective Uninformative	Uninformative
O'Hagan 2001b	Uninformative Subjective	Uninformative Structural
O'Hagan 2002	Uninformative	Uninformative
Hahn 2003	Uninformative	Uninformative
Heitjan 2004a	Empirical	Subjective Uninformative
Heitjan 2004b	Empirical	Subjective
Vazquez-Polo 2005a	Uninformative	Uninformative
Vazquez-Polo 2005b	Subjective	Subjective
Negrin 2006	Uninformative	Uninformative
Bachmann 2007	Uninformative	Uninformative
Application Papers		
Fenwick 2002	Uninformative	Uninformative
UK BEAM Trial Team 2004	Uninformative	Uninformative
Shih 2007	Uninformative Empirical	Uninformative Empirical

1. Uninformative: no information
2. Empirical: data based
3. Subjective: opinion based
4. Structural: relationship based

Total number of priors for effects = 22 [uninformative = 13(59%), empirical = 5(23%), subjective = 4(18%)]. Total number of priors for costs = 22 [uninformative = 14(64%), empirical = 3(14%), subjective = 4(18%), structural = 1(5%)]. Percentages rounded to nearest whole number.

Table 2. Description of Likelihoods

Author(s)	Distributional form of likelihood[1,2,3]
Methodological Papers with Applications	
Heitjan 1999	(meanEffects, meanCosts)~Multivariate normal Large sample approximation
Al 2000	(meanEffects, meanCosts)~Multivariate normal Large sample approximation
Briggs 2001	(meanEffects, meanCosts)~Multivariate normal Large sample approximation
O'Hagan 2001a	(Effects, Costs)~Multivariate normal
O'Hagan 2001b	Effects~Binomial Costs|Effects~Lognormal
O'Hagan 2002	Effects~Weibull Costs~nonparametric
Hahn 2003	i.Effects~Normal Costs~Normal ii.Effects~Normal Cost components~Normal iii.Effects~Normal Cost components~other distributions iv.(Effects, Costs)~Multivariate normal v.(Effects, Costs components)~Multivariate normal
Heitjan 2004a	Effects~Binomial Costs|Effects~Gamma
Heitjan 2004b	Effects~Weibull Cost|Effects~Gamma
Vazquez-Polo 2005a	Effects~Normal Costs|Effects~Normal
Vazquez-Polo 2005b	i.(Effects, Costs)~Multivariate normal with probit model for effects ii.(Effects,Costs)~Multivariate normal
Negrin 2006	(Effects,logCosts)~Multivariate normal
Bachmann 2007	i.Effects|Costs~Binomial Costs~Gamma ii. Net benefit~Normal
Application Papers	
Fenwick 2002	Not Specified
UK BEAM Trial Team 2004	Not Specified
Shih 2007	Net benefit~Normal

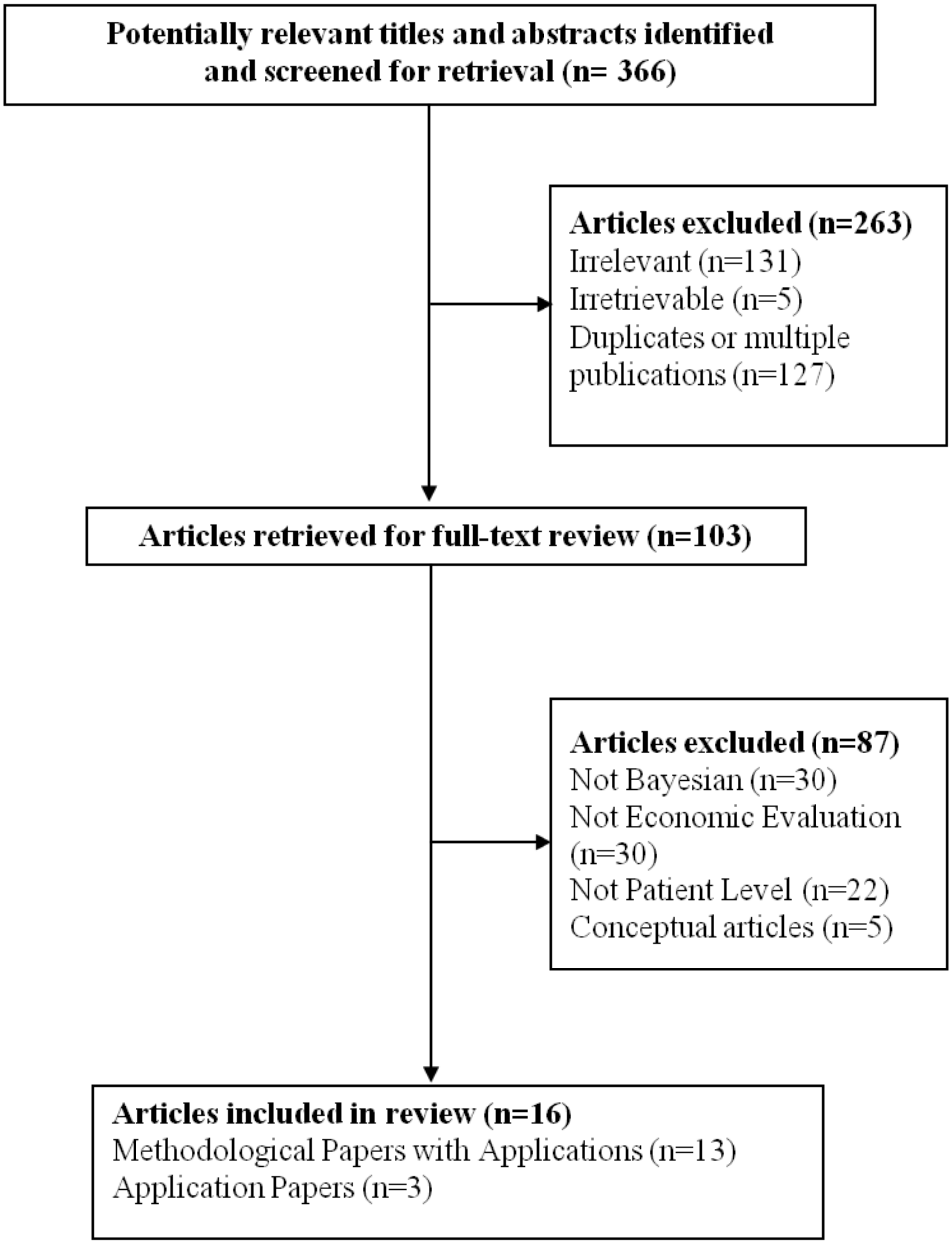
1. (effects,costs): effects and costs determined simultaneously, 2. cost|effects: costs depend on effects,
3. effects|costs: effects depend on costs. Total number of distributions = 29 [multivariate normal = 9(31%), normal = 9(31%), binomial = 3(10%), gamma = 3(10%), weibull = 2(7%), lognormal = 1(3%), other = 1(3%), nonparametric = 1(3%)]. Percentages rounded to nearest whole number.

Table 3. Presentation of Uncertainty and Sensitivity Analysis

Author(s)	Presentation of Uncertainty	Sensitivity Analysis[2]
Methodological Papers with Applications		
Heitjan 1999	95% Credibility Interval ICER	No
Al 2000	95% Credibility Interval ICER CEAC	Prior
Briggs 2001	CEAC	Prior
O'Hagan 2001a	CEAC	Prior Likelihood
O'Hagan 2001b	CEAC	Prior
O'Hagan 2002	CEAC	No
Hahn 2003	CEAC	Likelihood
Heitjan 2004a	95% Credibility Interval ICER 95% Credibility Interval INHB CEAC	Prior
Heitjan 2004b	95% Credibility Interval ICER 95% Credibility Interval INMB CEAC	No
Vazquez-Polo 2005a	95% Credibility Interval ICER CEAC	Likelihood
Vazquez-Polo 2005b	CEAC	Likelihood
Negrin 2006	CEAPF[1]	No
Bachmann 2007	95% Credibility Interval ICER 95% Credibility Interval INMB CEAC	Likelihood
Application Papers		
Fenwick 2002	CEAC	No
UK BEAM Trial Team 2004	CEAC	No
Shih 2007	CEAC	Prior Likelihood

1. proposed as an alternative to the cost-effectiveness acceptability curve when considering more than one measure of effect. 2. refer to paper for description of priors and likelihoods. Number of presentations of uncertainty = 24 [cost-effectiveness acceptability curves (CEAC) = 14 (58%), 95% credibility interval for incremental cost-effectiveness ratio (ICER) = 6 (25%), 95% credibility interval for incremental net monetary benefit (INMB) = 2(8%), 95% credibility interval for incremental net health benefit (INHB) = 1(4%), cost-effectiveness acceptability plane frontier (CEAPF) = 1(4%)]. Number of sensitivity analyses = 10 [Prior sensitivity = 4(40%), Likelihood sensitivity = 4(40%), Prior and Likelihood sensitivity = 2(20%)]. Percentages rounded to nearest whole number.

Figure 1. QUORUM Diagram of Studies Considered for Inclusion


Potentially relevant titles and abstracts identified and screened for retrieval (n= 366)
Articles excluded (n=263)
Irrelevant (n=131)
Irretrievable (n=5)
Duplicates or multiple publications (n=127)
Articles retrieved for full-text review (n=103)
Articles excluded (n=87)
Not Bayesian (n=30)
Not Economic Evaluation (n=30)
Not Patient Level (n=22)
Conceptual articles (n=5)
Articles included in review (n=16)
Methodological Papers with Applications (n=13)
Application Papers (n=3)

Supplementary Table 1. Search Strategy

DATABASE	SEARCH STRATEGY
Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1950 to Present>	1 Bayes Theorem/ (7604) 2 (Bayesian or WinBUGS).ti,ab. (7744) 3 (Baye$ adj3 (analy##s or decision$ or estimat$ or forecast$ or method$ or predict$ or theor$ or uncertai$ or simulat$ or net or sensiti$ or probabil$)).ti,ab. (5091) 4 or/1-3 (11069) 5 *economics/ (9551) 6 exp "Costs and Cost Analysis"/ (133465) 7 (cost$ or budget$ or economic or pharmacoeconomic$ or pharmaco economic$ or price$).ti. (78603) 8 (cost$ adj2 (benefit$ or effective$ or minimi#ation or utilit$)).ti,ab. (51677) 9 (econom$ adj5 (analy##s or evaluat$ or impact$)).ti,ab. (11481) 10 or/5-9 (202891) 11 (trial$ or stud$).ti,ab. (4692913) 12 4 and 10 and 11 (160) 13 limit 12 to english language (153) 14 from 13 keep 1-153 (153)
EMBASE <1980 to 2007 Week 45>	1 Bayes Theorem/ (5666) 2 (Bayesian or WinBUGS).ti,ab. (5495) 3 (Baye$ adj3 (analy##s or decision$ or estimat$ or forecast$ or method$ or predict$ or theor$ or net or simulat$ or sensitiv$ or uncertai$ or probabil$)).ti,ab. (3686) 4 or/1-3 (7542) 5 exp Health Economics/ (205516) 6 Economic Aspect/ (68420) 7 (cost$ or budget$ or economic or pharmacoeconomic$ or pharmaco economic$ or price$).ti. (49966) 8 (cost$ adj2 (benefit$ or effective$ or minimi#ation or utilit$)).ti,ab. (45739) 9 (econom$ adj5 (analy##s or evaluat$ or impact$)).ti,ab. (10396) 10 or/5-9 (280087) 11 (trial$ or stud$).ti,ab. (3655381) 12 4 and 10 and 11 (192) 13 limit 12 to english language (186) 14 from 13 keep 1-186 (186)
Cochrane Library NHS Economic Evaluation Database <Issue 4, 2007>	1 (baye*) 2 (baye* theorem) 3 (baye* near/3 (analy**s or decision* or estimat* or forecast* or method* or predict* or theor* or net or simulat* or uncertai* or probabil* or sensiti*)):ti,ab,kw 4 (Bayesian or WinBUGS):ti,ab,kw 5 (#1 OR #2 OR #3 OR #4) 6 (trial* or stud*):ti,ab 7 (#5 AND #6)

Supplementary Table 2. Description of Included Studies

Author(s)	Source of patient data	Number of Interventions	Effects	Costs	Type of Economic Evaluation
Methodological Papers with Applications					
Heitjan 1999	Trial	2	Proportion	Direct	CEA
Al 2000	Trial	2	Proportion	Direct	CEA
Briggs 2001	Trial	2	Life years	Direct	CEA
O’Hagan 2001a	Trial	2	Natural units	Direct	CEA
O’Hagan 2001b	Trial	2	Proportion	Direct	CEA
O’Hagan 2002	Trial	2	Life years	Direct	CEA
Hahn 2003	Trial	2	Natural units	Direct	CEA
Heitjan 2004a	Trial	2	Proportion	Direct	CEA
Heitjan 2004b	Trial	2	Life years	Direct	CEA
Vazquez-Polo 2005a	Simulated data	2	Quality adjusted life weeks	Direct	CUA
Vazquez-Polo 2005b	Trial	2	Proportion	Direct	CEA
			Change in visual analog score		CUA
Negrin 2006	Trial	3,4	Incorporated both a proportion and a change in quality of life into a single analysis	Direct	CUA
Bachmann 2007	Trial	2	Proportion	Direct	CEA
Application Papers					
Fenwick 2002	Trial	3	Life years	Direct	CEA
UK BEAM Trial Team 2004	Trial	4	Quality adjusted life years	Direct	CUA
Shih 2007	Individual Medicare claims data	5	Proportion	Direct	CEA

Abbreviations: Cost-Effectiveness Analysis (CEA), Cost-Utility Analysis (CUA)
Note: *Methodological Papers with Applications* refers to papers that used applications merely for illustrative or pedagogic purposes; *Application Papers* refers to papers whose primary objective was an economic evaluation, where Bayesian methods were used to incorporate sampling uncertainty.