Charlotte Rietbergen
Utrecht University, Utrecht

Quantitative Evidence Synthesis with Power Priors

Rietbergen, Charlotte
Quantitative Evidence Synthesis with Power Priors
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 978-90-6464-971-4
Druk: GVO drukkers & vormgevers B.V. — Ponsen & Looijen
Cover designed by Charlotte Rietbergen

Quantitative Evidence Synthesis with Power Priors

KWANTITATIEVE SYNTHESE VAN BEWIJS MET POWER PRIORVERDELINGEN

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht op gezag van
de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op vrijdag 12 februari 2016
des ochtends te 10.30 uur

door

Charlotte Rietbergen

geboren op 17 februari 1983
te Eindhoven

Promotoren: Prof. dr. H.J.A. Hoijtink
Prof. dr. K.G.M. Moons
Prof. dr. I. Klugkist

Copromotor: Dr. R.H.H. Groenwold

Contents

1 Introduction 1

1.1 An example 1

1.2 Techniques for quantitative evidence synthesis 2

1.3 Selection and weighing of useful information 3

1.4 Quantification of quality and relevance 4

1.5 Expert judgments 5

1.6 Power priors and aim of the thesis 5

1.7 Outline of the thesis 7

2 Incorporation of Historical Data in the Analysis of Randomized Therapeutic Trials 9

2.1 Introduction 9

2.2 Conceptual Framework 11

2.2.1 Short description of the data 11

2.2.2 The value of historical data. 11

2.3 The Power Prior Distribution for a Single Historical Study 12

2.3.1 The conditional power prior distribution 12

2.3.2 Empirical example: The power prior distribution in practice . . 14

2.4 The Power Prior Distribution for Multiple Historical Studies 16

2.4.1 Fixed or random study weights? 16

2.4.2 Eliciting study-ranking from experts 17

2.5 Discussion 19

2.6 Acknowledgements 21

2.7 Appendix Chapter 2 22

2.7.1 Elaborate Example of the Conditional Power Prior Distribution 23

3 Expert Elicitation of Study Weights for Bayesian Analysis and Meta-Analysis 25

3.1 Introduction 25

3.2 Methods 27

3.2.1 Motivating Example . . . 27
3.2.2 The Expert Panel . . . 28
3.2.3 Description of the Delphi Rounds . . . 29
3.3 Results . . . 30
3.3.1 Main Results: Relevant Study Characteristics, Ranking and Weights . . . 30
3.3.2 Report of Rounds . . . 31
3.4 Acknowledgements . . . 36
3.5 Appendix Chapter 3 . . . 37

4 Power prior distributions and sampling variability . . . 41
4.1 Introduction . . . 41
4.2 Types of Power prior distributions . . . 43
4.2.1 Power prior distributions with fixed weights . . . 43
4.2.2 Frequentist properties of the posterior estimates for normal data 43
4.2.3 Power prior distributions with random weights . . . 45
4.3 Sampling variability and the estimation of a_0 . . . 45
4.3.1 An example with normal data . . . 47
4.3.2 Real data example: the safety of rosiglitazone . . . 49
4.4 Discussion . . . 51

5 Evidence synthesis in drug safety assessment: the example of rosiglitazone . . . 53
5.1 Background . . . 54
5.2 Methods . . . 55
5.2.1 Stage I . . . 55
5.2.2 Stage II: Inclusion of studies per research question . . . 56
5.2.3 Stage III: Quality assessment . . . 60
5.2.4 Stage IV: Data extraction . . . 61
5.2.5 Stage V: Data analysis . . . 61
5.3 Results . . . 63
5.4 Discussion . . . 65
5.5 Conclusion . . . 66
5.6 Appendix Chapter 5 . . . 67

6 Reporting of Bayesian methods in epidemiological and medical research: a systematic review . . . 73
6.1 Background . . . 74
6.2 Methods . . . 74
6.2.1 Search strategy . . . 74
6.2.2 Methodological assessment . . . 77
6.3 Results . . . 77
6.3.1 Bayesian statistics in medical and epidemiological journals . . . 77
6.4 Discussion . . . 80

6.5 Conclusion 82
6.6 Acknowledgement 82
6.7 Appendix Chapter 6 83

7 General Discussion 85
7.1 The power prior distribution with random study weights 85
7.2 Elicitation of fixed study weights 86
7.3 Power priors for multiple historical studies 87
7.4 Sensitivity analyses and guidelines on reporting Bayesian data analyses 88
7.5 Concluding remarks 89

References 91

Studies included in the meta-analysis in Chapter 5 97

Studies included in the systematic review in Chapter 6 103

Summary in Dutch / Samenvatting 113

Acknowledgements 119

1

Introduction

Irrespective of the field of research scientific studies are never performed in a vacuum. In social sciences and, to a greater extent, medical sciences researchers tend to perform systematic reviews at the start of a new research in order to summarize what is already known about a research topic. The knowledge brought together in these reviews offer, besides the state of affairs at that time, a rationale for designing and conducting new research. Relevant findings from previous studies can be a motivation for further research and the required sample sizes of new studies can be calculated on the basis of effect sizes found previously. Researchers can even take it one step further by actually including the data or results from previous research into the analysis of the data obtained in new studies.

1.1 An example

Although in research practice the explicit quantitative synthesis of existing and new data is rare, researchers have an intuitive tendency to do this implicitly. Take, for example, the discussion on the effect of the intake of vitamin C on the prevention of complex regional pain syndrome (CRPS) (van der Graaf, 2013). In this discussion the a priori idea that a doctor or researcher has regarding the potential effectiveness of this vitamin plays an important role, since this expectation affects the interpretation of new research results on the relationship between vitamin C and CRPS. Based on what is known from research on the effect of vitamin C and the extent to which additional intake of this vitamin can prevent a common cold, the a priori expectation about its effectiveness in the prevention of CRPS is limited. After all, extensive research into the relationship between vitamin C and common cold has failed to demonstrate any positive effect of vitamin C. In the light of this knowledge, new research results which

This introduction is based on Rietbergen, C., Klugkist, I., & Groenwold, R.H.H. (2014). Kwantitatieve synthese van kennis. *STATOR*, *15*, 8:12.

seem to actually demonstrate a protective effect of vitamin C in the development of CRPS, will not easily convince the research community. However, the small number of previous studies that actually focus on the specific effect of vitamin C on the prevention of CRPS have found a small beneficial effect. When this knowledge would be taken into account, the a priori expectation with respect to a positive effect of vitamin C in the new study will be much larger. The results obtained in the new study will be considered convincing evidence much quicker.

The above illustrates how the results of previous research, in other words historical data, may affect the expectations about and credibility of new research outcomes and color the interpretation of new research findings. In addition, the example shows that different researchers might judge the suitability of sources of knowledge differently: as one person considers the information in a large collection of research on vitamin C and common cold to be very relevant, another one focusses on the small number of studies that evaluated the specific relationship between vitamin C and CRPS. In their informal attempt to synthesize the historical and new data, the researchers may disagree on the appropriateness of the historical data for the subject in question, and therefore the interpretation of the new results.

1.2 Techniques for quantitative evidence synthesis

Several formal procedures for the explicit synthesis of quantitative research data from historical and new sources exist. Two main analytical techniques for summarizing quantitative research can be distinguished: (cumulative) meta-analysis and Bayesian data analysis with informative prior distributions.

In a meta-analysis the results obtained in all studies on a particular topic are summarized as such that an overall estimate of an effect of interest is obtained. A cumulative meta-analysis is a series of meta-analyses in which each time the overall effect estimate is updated with results from a new study. Usually, the precision of each individual study determines the weight of that study in the meta-analysis. Studies in which larger numbers of participants are included, are often more precise in estimating the effect size of interest, and are therefore more heavily weighted in the meta-analysis than smaller studies.

In Bayesian statistics the historical information on the effect of interest is quantitatively summarized in a so-called prior probability distribution for that effect. This distribution reflects the available knowledge about the effect a researcher wants to estimate and the uncertainty associated with that knowledge. Consider for example the situation in which one examines the effectiveness of a new vaccine in human subjects for the first time (first-in-man study), and only information obtained through in vitro or animal studies is available. In this case a priori little or nothing is known about the effect of interest in humans and the prior distribution will contain little to no information. This means that a priori every result of the new study is considered equally plausible. With the data obtained in the new study the prior distribution can be updated. According to this updated or a posteriori distribution research results closer to the ones found in the new study are considered more plausible than outcomes farther away from these results. The more data obtained in the study, i.e. the larger the size of sample used in the study, the more the a posteriori distribution will be shaped by the new data. The a posteriori distribution can in turn serve as a prior distribution for a new study. This process shows great similarities with the above-mentioned cumulative meta-analysis.

1.3 Selection and weighing of useful information

The use of results from previous research in the analysis and interpretation of new data is neither simple nor uncontroversial. The aforementioned discussion on vitamin C shows that experts may differ in their assessment of the suitability of the historical information. This suitability depends on the quality of the information on the one hand and the relevance of the information on the other hand. With quality we mean the methodological quality of the research with which the data have been obtained and thus the extent to which the researchers have succeeded to limit the various forms of bias that can threaten the reliability and validity of research results. Relevance refers to the similarities between historical and new studies regarding for example the evaluated treatment or the formulated inclusion and exclusion criteria. In addition to the sample size, both the quality and relevance should determine the extent to which information from previous research is included in a quantitative synthesis. Not only large studies, but especially high quality and relevant historical studies should receive a higher weight in the meta-analysis such that these studies contribute more to the estimation of an effect size. Of course, every researcher aims to include both properly executed and relevant studies in their meta-analysis. However, the following example illustrates that this is not always feasible.

In 2010, the European Medicines Agency (EMA) recommended suspension of rosiglitazone-containing medicines because of the possible increased risk of cardiac disorders in patients using this drug. Nevertheless, research results with respect to possible harmful effects were not unambiguous. The arisen doubts about the safety of this drug could be a motivation for a new safety study, in which the results obtained in previous research could be used as historical information. Problematic here is the variety of historical sources. First, there are randomized studies that often focus on the anti-diabetic efficacy of rosiglitazone and are therefore less valuable as a source of information for unintend cardiac end points. On the other hand, there is a wealth of relevant historical information about the (cardiac) side effects of rosiglitazone found in observational (non-randomized) studies, which are often deemed to be of inferior quality in most guidelines on hierarchy of evidence. Where randomized study designs are better suited to rule out alternative explanations for apparent differences between the experimental and control condition, observational studies are not or less able to do so and results from these studies have a higher risk of being biased. Therefore, typically only the results from experimental studies are used and historical observational studies are not considered for quantitative syntheses on the effects of interventions.

However, the observational study is not by definition of poorer quality than the randomized study (Shrier et al., 2007).

The risk of distortion of the final meta-analysis result should be leading in the determination of the weight a randomized or observational study receives in the meta-analysis (Shrier et al., 2007). By assigning weights based on quality and relevance valuable information obtained in, in principle, any historical study could be incorporated in the analysis of the new data. Assessing the risk of bias is a subjective process that seems to clash with the pursuit of scientific objectivity. In intervention research this will often result in the a priori exclusion of observational studies. However, excluding a study for evidence synthesis comes down to assigning a weight to this study that is equal to zero. The choice of this weight is still a subjective one, albeit this choice is made less explicit. This shows that the subjective assessment of studies with respect to quality and relevance is inevitable. It is therefore the task of the researcher to assess these study characteristics in an explicit and informed manner, for example by using validated instruments for quality and relevance assessment and by consulting experts.

1.4 Quantification of quality and relevance

Different scales were designed to assess and quantify the quality and relevance of the historical data as objectively as possible. An example for randomized trials is the Cochrane Collaboration's tool for assessing risk of bias (Higgins et al., 2011). For observational research the Newcastle-Ottawa Scale was developed (Wells et al., 2013). Although these types of scales have been used to evaluate in particular the quality of available studies, it seems that researchers do not use these quality estimates in their quantitative syntheses since they do not know how they should do this (Hopewell et al., 2013; Moja et al., 2005). Currently there is a debate going on in the literature about the proper use of these scales (Groenwold and Rovers, 2010). The main question raised is whether the scores obtained with these scales can be used as weights in a quantitative synthesis or not. A key problem is that the existing scales do not distinguish between the importance of individual items. For example, a study could receive a very poor score on an item that assesses the study duration, because the study was too short to observe the outcome. However, when this study receives high scores on the remaining items these can still add up to a high total score. This way the study can, mistakenly, play an important role in the synthesis and with that in the estimation of the overall effect. In addition, it has been shown that different scales can lead to different assessments of quality for the same study and thus can lead to different weights. Because of this, the result of the quantitative synthesis depends on the type of scale used for quality assessment. The use of these scales might therefore lead to biased results (Jüni et al., 1999). A common solution is to set a lower limit for research quality and relevance; studies that initially did meet the inclusion criteria, can then be excluded in a later stage if they do not meet the quality requirements or when they do not seem relevant enough. A sensitivity analysis can reveal the impact of exclusion of low quality studies on the estimate effect size. This approach is not less subjective than the previously discussed options, but it is more explicit in corroborating the choices made.

1.5 Expert judgments

Researchers can assess the quality and relevance of the available information themselves, but it seems obvious to involve experts from the relevant field in this process. With experts we refer to doctors or researchers with extensive knowledge of or years of experience with a particular subject. Since the accuracy of subjective expert judgments cannot be tested, techniques to elicit judgments of quality and relevance from experts are rarely discussed in the literature. In practice, this means that historical information is either totally ignored or simply fully included. In the first case, one will analyze the new data without taking into account that what is already known about a subject. In the second case, a meta-analysis will be performed in which the study weight is only determined by the sample size. In both cases, the utilized approach appears objective at first sight, since quality and relevance of the included information are not explicitly assessed. However, as mentioned earlier, even when the historical information is fully included or excluded, the decision to adopt either approach is equally subjectively as choosing for an approach in which experts are asked to judge the quality and relevance of the historical information and to determine to what extent the historical information should play a role in the analysis and / or interpretation of new data.

1.6 Power priors and aim of the thesis

One specific approach for the synthesis of evidence obtained in studies of different quality and relevance within the Bayesian framework is proposed by Ibrahim and Chen (2000). Their power prior distribution offers a flexible approach for the inclusion of different types of historical data in the prior distribution that is specified for the analysis of newly collected data. Their procedure uses a weight parameter for the historical studies to deal with differences between the historical and new studies. This approach offers a starting point for the inclusion of historical data in the analysis of new data, and with that the formal synthesis of evidence obtained in different types of studies. Ibrahim and Chen propose two types of power prior distributions: the conditional power prior and joint power prior distribution.

The conditional power prior distribution is obtained by multiplying a low informative initial prior $\pi_0(\theta|c_0)$ with the likelihood of the historical data D_0 raised to a power a_0, as in

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0} \pi_0(\theta|c_0), \tag{1.1}$$

where θ is a vector of parameters of interest and a_0 the weight parameter with $0 \leq a_0 \leq 1$. Naturally, $a_0 = 0$ indicates no inclusion of historical data, and $a_0 = 1$ equals full inclusion of the historical data. This way, the weight parameter enables the researcher to downweight the historical data in case either the quality of the historical data is poor, or in case the historical research design or the research results are somewhat different from the design and results for the new study.

The joint power prior distribution in which θ and a_0 are jointly estimated from the current and historical data, is given by

$$\pi(\theta, a_0|D_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta|c_0)\pi(a_0|\gamma_0), \tag{1.2}$$

Ibrahim and Chen note that by using this joint power prior, for which an additional initial prior $\pi(a_0|\gamma_0)$ for a_0 is to be specified, the tails for the marginal distribution of θ better reflect the uncertainty about a_0 than the marginal posterior following from the conditional power prior distribution that assumes fixed a_0.

Evaluations of the performance of the power prior (see Duan, 2006; Neuenschwander et al., 2009) revealed that for binary and normal data, the joint power prior distribution as given above results in a marginal prior for a_0 with values close to zero, even if the data show evidence of commensurability between the historical and current data. The modified power prior distribution as proposed by Duan (2006) and also Neuenschwander et al. (2009) normalizes the (marginal) prior distribution for a_0, resulting in larger values for a_0. This modified power prior distribution builds on the assumption that both the current and historical data are needed to update the distribution for a_0. As a consequence, higher weights are assigned to the historical data sets when they show more resemblance with the current study result. In addition, Hobbs et al. (2011) present modifications that also allow the commensurability of the historical and current samples to determine how much of the historical information is used.

All above versions of the joint power prior distribution for θ and a_0 result in a marginal posterior distribution for a_0 that is influenced by the degree of agreement between the results obtained in historical study and the current study result. This property is convenient for the researcher who is not willing or able to make a subjective judgement about the commensurability of the research characteristics and prefers to estimate the size of the weight parameter solely on the available data. It is, however, questionable whether the dependence of the weight parameter on the commensurability of the research results is desirable. Some authors, see for example Neelon and O'Malley (2009), suggest that using a conditional power prior distribution, as originally proposed by Ibrahim and Chen, with fixed value for a_0 might sometimes be more appropriate.

In this dissertation the use of power priors in applied (medical) research settings is evaluated. The desirability of the situation in which the weight parameter depends on the agreement in study results rather than study characteristics is assessed. Furthermore, the use of the conditional power prior in which fixed weights are assigned to the historical studies based on relevance and quality of the historical study is examined.

The aim of this thesis is to provide the applied researcher with a practical approach for quantitative evidence synthesis using the conditional power prior that allows for subjective input and thereby provides an alternative to deal with the difficulties associated with the joint power prior distribution.

1.7 Outline of the thesis

In Chapter 2 the use of the power prior distribution is assessed in case a treatment effect is to be estimated in a randomized trial and the historical data come from randomized studies that used slightly different patient populations and slightly different study designs. Sensitivity analyses were performed to evaluate whether simply ranking the historical studies would provide a solid basis for study weight assignment.

Chapter 3 elaborates on the assignment of fixed study weights. In this chapter a Delphi procedure to elicit study weights from a panel of experts is designed and evaluated. Experts were asked to rank and weigh four historical studies with respect to quality and relevance. They were asked to report their motivation for their choices in each round of the Delphi study, and the other experts were able to adjust their ranking and weights if needed, and to respond on the other experts in an anonymous way. This process was monitored to evaluate whether this approach is suitable for the elicitation of study weights in applied research.

As mentioned before, Ibrahim and Chen (2000) do not only propose a power prior approach in which the weight parameter is fixed and user specified, but also describe a procedure in which the size of the weight parameter is estimated from the commensurability of the historical and new studies results. In Chapter 4 the question is asked whether this procedure might lead to biased results, since differences or similarities in study results might be the product of sampling variability. In this case the value of the weight parameter would depend on the sample results, meaning that natural variation in sampling results might lead to varying values for the size of the weight parameter. Through a numerical example and a simulation study the size of this problem is assessed and discussed.

The fifth chapter is dedicated to an application of a power prior procedure for the cross-design synthesis of evidence for drug safety evaluations. A structured approach is proposed for the inclusion of relevant studies per specific research question raised in a meta-analysis and the assessment of study quality is discussed. In this example the prior distribution for the evaluation of drug safety in a meta-analysis of more or less relevant randomized trials is obtained from a set of relatively large observational studies of varying quality. A procedure is evaluated to assess and, where possible, limit the influence of the informative prior distribution on the posterior results.

In Chapter 6 the results are presented from an extensive systematic review on the use of Bayesian statistics in medical and epidemiological literature in general, and the use of and reporting on different types of prior distributions in specific.

The thesis is concluded with a summary discussion on the results obtained in the different chapters, and with some remarks about the consequences of the main results.

2

Incorporation of Historical Data in the Analysis of Randomized Therapeutic Trials

Summary. Historical studies provide a valuable source of information for the motivation and design of later trials. Bayesian techniques offer possibilities for the quantitative inclusion of prior knowledge within the analysis of current trial data. Combining information from previous studies into an informative prior distribution is, however, a delicate case. The power prior distribution is a tool to estimate the effect of an intervention in a current study sample, while accounting for the information provided by previous research. In this study we evaluate the use of the power prior distribution, illustrated with data from a large randomized clinical trial on the effect of ST-wave analysis in intrapartum fetal monitoring. We advocate the use of a power prior distribution with pre-specified fixed study weights based on differences in study characteristics. We propose obtaining a ranking of the historical studies via expert elicitation, based on relevance for the current study, and specify study weights accordingly.

2.1 Introduction

In the process of designing a clinical trial to examine the occurrence relation between determinant and disease-outcome, the design and results from previous studies form a valuable source of information. Quantitative incorporation of results from prior studies into the actual analysis of current trial data would exploit the benefit of the presence of historical studies to an even larger extent. Bayesian techniques offer possibilities for inclusion of prior knowledge within the analysis of the current trial data (for an elaborate introduction on applied Bayesian data analysis we refer the reader to Greenland (2006), Greenland (2007), Spiegelhalter et al. (1999), Goodman (1999a), and Goodman (1999b)). The *a priori* state of belief about the effect of an intervention based on results from previous studies (or other sources) is represented

This chapter has been published as: Rietbergen, C., Klugkist, I., Janssen, K.J.M.,Moons, K.G.M., & Hoijtink, H.J.A.(2011). Incorporation of Historical Data in the Analysis of Randomized Therapeutic Trials. *Contemporary Clinical Trials*, 32(6), 848-855.

by an informative prior distribution. Updating the prior distribution with the current data results in a posterior distribution, that is summarized to obtain an estimate of the intervention effect. Combining information from previous studies into an informative prior is, however, a delicate case. Ibrahim and Chen (2000) present the power prior distribution that enables the historical data to be weighted relative to the current data. This technique allows the researcher to specify either a fixed weight parameter, or a prior distribution for the weight parameter, that determines the amount of historical data to be included in the analysis of the current data.

A first objective of this paper is to give a motivation for the use of the the power prior distribution in the analysis of current trial data. Illustrated with data from a gynaecological randomized clinical trial (RCT) by Westerhuis et al. (2009), that builds on cumulative knowledge obtained from several comparable studies, we construct a conceptual framework for the comparison of the power prior with a meta-analytic and full Bayesian approach.

The technical aspects of power prior specification are a second concern of this paper, and are discussed for the situation in which data from a single historical study is to be included in the analysis of the current study by Westerhuis. As the posterior distribution for the treatment effect might be sensitive to changes in the weight parameter in the power prior, careful selection of the weight parameter is demanded. To circumvent the subjectivity associated with the selection of the study weight, a prior distribution for the weight can be specified in a hierarchical power prior distribution. A short review of the current discussion regarding the hierarchical power prior is given later in this paper. The importance of careful specification of study weights increases as data from more historical studies is included in the prior distribution. However, the power prior for multiple historical studies, and in particular the hierarchical power prior distribution, remains under exposed in current literature. The third object of this study is, therefore, to explore and evaluate the process of power prior specification for the case of multiple historical data sets using fixed study-specific weights. Often multiple historical data sets with varying degrees of informativeness for the current trial are at the researchers disposal. We propose consulting experts within the field of interest to come to a ranking of the historical studies based on relevance each historical study has for the current trial. The power parameters of the power prior distribution are then specified according to this ranking. The amount of influence the choice of weights has on the posterior distribution for the treatment effect is evaluated by means of a sensitivity analysis.

A description of the current and historical data is given in Subsection 2.2.1. In Subsection 2.2.2 the conceptual framework for the use of the power prior is constructed. The full process of power prior specification is described and illustrated in Section 2.3. The fourth section is dedicated to the process of power prior specification and weight elicitation for multiple historical studies. A discussion is given in Section 2.5.

2.2 Conceptual Framework

2.2.1 Short description of the data

The main object of the RCT by Westerhuis et al. (2009) was to assess whether fetal monitoring using cardiotocography (CTG) combined with ECG ST-segment analysis (ST) was associated with a reduced risk of metabolic acidosis and instrumental delivery, compared to CTG alone. The relative risk (RR) and some of the study-characteristics are shown in the first row of Table 2.1. Four previous RCTs with a comparable theoretical and study design were conducted between 1991 and 2006 (Vayssière et al. (2007); Ojala et al. (2006); Amer-Wåhlin et al. (2001); Westgate et al. (1993)). Characteristics and findings of those historical studies are displayed in the last four rows of Table 2.1.

Table 2.1. Relevant study characteristics by first author (Study); last year of inclusion (Year), total sample size (Size), country of commission (Country) and intervention conditions under study (Experimental vs. Control). Results; Maximum Likelihood estimate of the RR of metabolic acidosis and 95% confidence interval (CI).

	Study	Year	Size	Country	Experimental	Control	RR	95% CI
Current	Westerhuis	2007	5667	NL (EU)	CTG + ST	CTG	0.67	[0.38, 1.18]
Historical	Vayssière	2006	799	FR (EU)	CTG + ST	CTG	1.62	[0.53, 4.86]
	Ojala	2004	1436	FI (EU)	ST	CTG	2.45	[0.86, 6.85]
	Amer-Wåhlin	2000	4238	SE (EU)	CTG + ST	CTG	0.46	[0.25, 0.86]
	Westgate	1991	2434	UK (EU)	CTG + ST	CTG	0.38	[0.14, 1.07]

As displayed in the table, the direction and strength of the estimated treatment effect vary from study to study. In only one of the studies a statistically significant effect ($p < .05$) was found. Furthermore, the table reveals some important differences between the study characteristics, for example, the evaluated interventions and year and country of commission. These differences will play a role in the specification of the power prior distribution (see Section 2.4).

2.2.2 The value of historical data.

The differences between the studies in Table 2.1 can be explained from the perspective that each study sample is a random draw from a study-specific subpopulation. This subpopulation, itself, can be seen as drawn from an encompassing population, that is infinite over time. For the two historical studies by Amer-Wåhlin and Vayssière, and for the current study by Westerhuis this concept is illustrated in Figure 2.1. The resulting treatment effect in each of the studies is an estimate of the treatment effect in each study-specific subpopulation. Pooling the treatment effects in all possible subpopulations in the past, present and future would result in a distribution of the treatment effect in the encompassing population. In a meta-analysis, the distribution of the treatment effect in the encompassing population is approximated by pooling the treatment effects estimated in the available study samples. When the focus of interest of a researcher is on the treatment effect in the encompassing population, this would be a natural tool to obtain an effect estimate.

When a researcher is interested in the treatment effect valid for a specific subpopulation (e.g., WH in Figure 2.1), a current data analysis is the common approach. To benefit from the historical data, one could perform a full Bayesian analysis to combine the data from the historical studies (sample A and V in Figure 2.1) into an informative prior distribution. However, updating this prior distribution with the data from the study by Westerhuis rather gives an estimate of the treatment effect for the encompassing population. Conceptually this is equal to performing a meta-analysis. Hence, when a researcher is interested in the effect for the study-specific subpopulation, the historical data would receive too much weight.

The power-prior distribution as described by Ibrahim and Chen (2000) allows the researcher to assign different weights to the historical data relative to the current data. This way the researcher decides on the amount of information obtained in a previous study to be included in the analysis of the current data. In the situation of the trial by Westerhuis we deal with multiple historical studies. Although all studies evaluated similar interventions, it was pointed out in the previous subsection that the studies differ on certain study characteristics. The partial overlap between the ellipses in Figure 2.1 represents the degree of shared characteristics between the study-specific subpopulations A, V and WH. Relevant information for population WH is captured in sample A and, to a lesser extent, sample V. From this it could be argued that the historical trials differ in the degree of informativeness for the current trial. Recent trials and trials from countries with similar obstetric protocols as followed within Dutch hospitals are considered to be more informative. For example, the study by Ojala et al. (2006) uses a different experimental condition than all other trials, which makes this trial possibly less informative.

We propose assigning study-specific weights to the historical data sets included in the power prior, based on this degree of informativeness. This approach provides the researcher with a tool to not only account for the type of information captured in historical studies, but to also control the amount of information included per study. Note that the specification of study weights is not straightforward, we will elaborate on this in the following sections. We first discuss the specification and evaluation of the power prior distribution for a single historical study.

2.3 The Power Prior Distribution for a Single Historical Study

2.3.1 The conditional power prior distribution

In a Bayesian analysis estimates of the treatment effect are obtained through the posterior distribution. According to Bayes' theorem the posterior is proportional to the

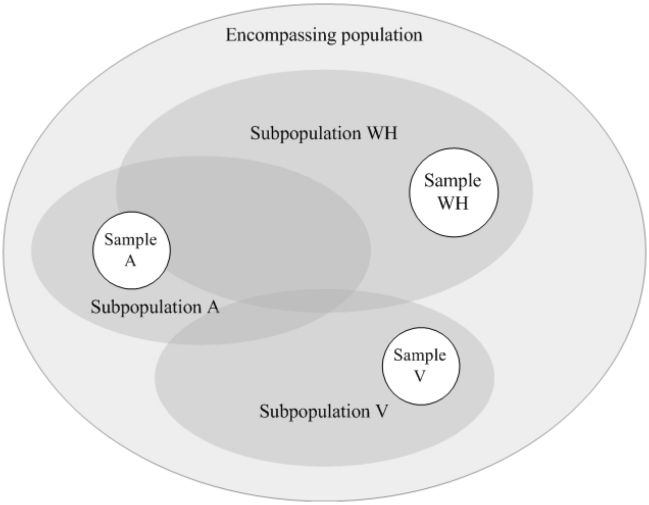

Fig. 2.1. Schematic representation of the historical study samples and subpopulations by Amer-Wåhlin (A) and Vayssière (V), the current sample and subpopulation by Westerhuis (WH), and the encompassing population.

likelihood of the current data multiplied with a prior distribution. In the trial by Westerhuis the primary outcome measure concerns the presence or absence of metabolic acidosis in newborns. The object of the study is to evaluate whether the proportion newborns with metabolic acidosis varies between the intervention conditions. A natural model for dichotomous data like these is the binomial distribution

$$\pi(y_i|\theta_i, n_i) = \binom{n_i}{y_i}{\theta_i}^{y_i}(1-\theta_i)^{n_i-y_i}, \tag{2.1}$$

where θ_i represents the probability of a newborn suffering from metabolic acidosis within intervention condition i, and $i = ctg, ctg + st$. The sample size and number of newborns with metabolic acidosis within each condition are denoted by n_i and y_i respectively.

The next step in a Bayesian analysis is the specification of a prior distribution for the parameters of interest. A conjugate prior for the binomial model is a beta distribution

$$\pi_0(\theta_i) \propto {\theta_i}^{\alpha_i-1}(1-\theta_i)^{\beta_i-1}, \tag{2.2}$$

we take $\alpha_i = \beta_i = 1$ to make it a standard low-informative prior distribution for the binomial model. To update the initial prior distribution with a certain amount of information provided by one of the historical studies, the initial prior is multiplied with the likelihood of the historical data raised to the power a_0 (Ibrahim and Chen, 2000). The value for a_0 determines the amount of data of the historical study to be included. We assume equal weight parameters for both intervention conditions and therefore do not use subscript i for the power parameter. As $0 \leq a_0 \leq 1$, naturally, $a_0 = 0$ indicates no inclusion of the historical data, and $a_0 = 1$ equals full inclusion of the historical data. Assuming a fixed power parameter a_0, the resulting beta-binomial posterior distribution is used as the conditional power prior distribution for θ_i, that is

$$\pi(\theta_i | y_{0i}, n_{0i}, a_0) \propto \theta_i^{y_{0i}a_0 + \alpha_i - 1}(1 - \theta_i)^{(n_{0i} - y_{0i})a_0 + \beta_i - 1}, \tag{2.3}$$

where n_{0i} and y_{0i} denote the size of the historical sample and number of newborns suffering from metabolic acidosis respectively. The resulting posterior is used as a Beta(a_i, b_i) power prior for θ_i, where $a_i = y_{0i}a_0 + \alpha_i$ and $b_i = (n_{0i} - y_{0i})a_0 + \beta_i$.

Updating the power prior distribution with data from the current trial results in a posterior distribution for θ_i of the form

$$\pi(\theta_i | y_i, n_i, y_{0i}, n_{0i}, a_0) \propto \theta_i^{y_i + a_i - 1}(1 - \theta_i)^{n_i - y_i + b_i - 1}, \tag{2.4}$$

which is a Beta$(y_i + a_i, n_i - y_i + b_i)$ distribution. To account for uncertainty regarding the actual value of the power parameter Ibrahim and Chen (2000) propose specifying a prior distribution for a_0. This joint power prior distribution for θ_i and a_0 will be discussed within the context of multiple historical studies in Section 2.4.

2.3.2 Empirical example: The power prior distribution in practice

As an illustration a Bayesian analysis on the data from the study by Westerhuis as presented in Table 2.2 is performed using WinBUGS (Lunn et al., 2000). We use data from the study by Amer-Wåhlin (see Table 2.3) as input for the power prior distribution with weight $a_0 = 0.5$, which comes down to using 50% of the historical data. The posterior distribution for the RR is obtained by taking the following steps:

- From Table 2.2 we obtain $n_{ctg+st} = 2827$ and $y_{ctg+st} = 20$, and $n_{ctg} = 2840$ and $y_{ctg} = 30$ to specify the likelihood function in (2.1).
- Using 50% of the data in Table 2.3 combined with the initial Beta(1,1) we obtain $a_{ctg+st} = 0.5 * 15 + 1 = 8.5$ and $b_{ctg+st} = 0.5 * 2144 + 1$, which gives a Beta(8.5,1073) power prior distribution for the CTG+ST condition, and likewise a Beta(16.5,1025) power prior for the CTG condition (see (2.3)).
- Combining likelihood and power prior as in (2.4) for each of the two intervention conditions results in a Beta(28.5, 3880) and Beta(46.5, 3835) posterior distribution respectively.
- Using WinBUGS the posterior distribution for the RR is approximated by sampling from the posterior distributions of both proportions θ_{ctg+st} and θ_{ctg}, thereby calculating RR_t in each of $T = 10000$ iterations. WinBUGS code to obtain the posterior distribution is given in the Appendix.

After carefully monitoring convergence (see Appendix) we take the mean and median of the resulting posterior distribution for the RR as estimates of the effect of ST-analysis. The mean of the posterior distribution equaled 0.62, whereas the median equaled 0.61, indicating a slightly skewed distribution.

From the posterior distribution a 95% central credibility interval (CCI) can be obtained, that corresponds to the range of values above and below which lies 95% of the posterior probability. This interpretation of the CCI differs from the non-Bayesian 95% confidence interval, that indicates that by repetition in 95% of the samples the true population value lies within that interval. For the current situation 95% of the posterior probability lies within the interval [0.38, 0.96]. As a RR equal to 1 is not included within this interval, we would conclude from this analysis that there is an effect of ST-analysis on metabolic acidosis.

Table 2.2. Westerhuis data

	Type of Intervention	
	CTG+ST	CTG
Metabolic Acidosis	20	30
Total Women Included	2827	2840

Table 2.3. Amer-Wåhlin data

	Type of Intervention	
	CTG+ST	CTG
Metabolic Acidosis	15	31
Total Women Included	2159	2079

2.3.2.1 Sensitivity Analysis for a Single Historical Study

As a_0 is supposed to be quite influential on the posterior distribution and not straightforward to specify, Ibrahim and Chen (2000) recommend always performing a sensitivity analysis to assess the degree to which different weights for the historical study lead to different conclusions.

As an illustration, the posterior distribution for the treatment effect in the trial by Westerhuis was evaluated at different weights for the study by Amer-Wåhlin et al. (2001). The ML-estimates of the treatment effect under both studies were in favor of the CTG plus ST-wave analysis ($RR < 1$). Study weights were set at $a_0 = (0, 0.25, 0.5, 0.75, 1.0)$. The decreasing width of the credible intervals displayed in the first five rows of Table 2.4 show that the amount of evidence for $RR < 1$ increases by incorporating more information from the historical study. The upper bound of the interval drops below 1. Adding more prior knowledge to the analysis altered our conclusions about the treatment effect. From being indecisive about the presence or absences of an effect of ST-wave analysis in addition to CTG after a current data analysis, we would now conclude there is an effect.

A different situation occurs when we take the data from the study by Vayssière et al. (2007) as the only prior knowledge we have. The treatment effect found in this relatively small study was in favor of non-ST fetal monitoring ($RR > 1$). The last five rows of Table 2.4 show the result of the sensitivity analysis when only the study by Vayssière is incorporated in the prior. The smaller the amount of historical data included the closer the RR goes to the ML-estimate in the current trial. As the sample size for the Vayssière-trial is relatively small compared to the current trial, the choice of a_0 did not have much influence on the conclusions about the treatment effect. Whether we included or excluded the historical data, the 95% posterior interval was 'centered' around 1. The historical and current data did not provide evidence of a treatment effect of ST-wave analysis in addition to CTG.

The results in Table 2.4 show that conclusions about the treatment effect in the current subpopulation are sensitive to the power parameter even if only a single historical study is used as prior input. In the presence of multiple historical studies power parameter specification becomes even more consequential. We will elaborate on this in the following section.

Table 2.4. Results sensitivity analysis using a single historical study, given the weight of the study (a_0); posterior mean (PM), median(PMed) and the 95% central credibility interval (CCI) for the RR.

Historical data	a_0	PM	PMed	95% CCI
Amer-Wåhlin	0.00	0.70	0.68	[0.39, 1.16]
	0.25	0.66	0.64	[0.38, 1.05]
	0.50	0.62	0.61	[0.38, 0.96]
	0.75	0.60	0.59	[0.38, 0.90]
	1.00	0.58	0.57	[0.38, 0.86]
Vayssière	0.00	0.70	0.68	[0.39, 1.16]
	0.25	0.74	0.71	[0.42, 1.22]
	0.50	0.77	0.75	[0.44, 1.25]
	0.75	0.80	0.78	[0.47, 1.28]
	1.00	0.83	0.81	[0.49, 1.31]

2.4 The Power Prior Distribution for Multiple Historical Studies

2.4.1 Fixed or random study weights?

To account for the uncertainty regarding the weight parameter a_0, Ibrahim and Chen (2000) propose a hierarchical power prior distribution. This comes down to specifying a joint prior distribution for θ_i and a_0, so that no fixed value for the weight parameter has to be chosen. A desirable feature of this hierarchical prior distribution is that it creates heavier tails for the marginal distribution for θ_i (Ibrahim and Chen, 2000), thereby better reflecting uncertainty with respect to a_0.

Several authors discuss modifications to the hierarchical power prior distribution. The modified power prior distribution as proposed by Duan Duan (2006) builds on the assumption that both the current and historical data are needed to update the distribution for a_0. As a consequence, higher weights are assigned to the historical data when they show more resemblance with the current sample. This is in agreement with the modified power prior distributions for a single historical study as proposed by Neuenschwander et al. (2009) and also by Hobbs et al. (2011).

The modified versions of the joint power prior distribution for θ_i and a_0 result in marginal posterior distributions for a_0 that are influenced by the commensurability of results obtained in the historical study and the current study result. Whether this would also hold for a joint power prior for multiple historical studies, remains uninvestigated.

Assigning higher weights to a study when the estimated treatment effect is similar to the effect found in the current sample, might induce bias towards the latter. Differences (or similarities) in study results, may be an indication of the informativeness of the historical study for the current, but it may just as well be a product of sampling variability (especially for small samples). Therefore, weight specification based on only the sample results and not the characteristic of the subpopulations (see Figure 2.1) could induce bias towards a result observed by chance. This is also communicated by Neelon and O'Malley (2009), who remark that the historical and current sample will always differ to some degree. The authors advocate the use of expert-elicited fixed values for a_0.

Given the lack of insight in the hierarchical power prior for multiple historical studies, and the risk of bias due to overreliance on sampling based results, we too propose the use of a conditional power prior distribution for multiple historical studies with study-specific fixed weights based solely on differences in study characteristics. This power prior distribution with study-specific fixed weights a_{0k} takes the form

$$\pi_i(\theta_i|D_{0ik}, a_{0k}) \propto \left[\prod_{k=1}^{L_0} [L(\theta_i|D_{0ik})]^{a_{0k}}\right] \pi_{0i}(\theta_i), \tag{2.5}$$

for $k = 1, ..., L_0$ historical data sets D_{0k}.

2.4.2 Eliciting study-ranking from experts

The differences between the study-characteristics make some historical trials more informative for the current trial than others. The degree of overlap between the study-characteristics provides a reasonable basis for the specification of study-specific weights. In a paper by Johnson et al. (2010a) several methods for expert elicitation for prior specification are reviewed. They conclude that several biases, such as overconfidence and lack of expertise of the expert, thread the reliability and validity of these elicitation methods (see also Johnson et al. (2010b)).

We propose adopting a form of expert elicitation in which the expert is asked to order the historical studies according to the degree to which he or she thinks these studies are informative for the current one. Using this relatively simple approach we do not have to elicit exact power parameter values. The expert is asked to base the ranking on the degree of overlap in study-characteristics with the current trial. Furthermore, the expert is asked to motivate his or her ranking, by reporting the study-characteristics on which the ranking was based. Note that the expert is considered the qualified person to decide which characteristics are (most) important. In order not to influence the choice of the expert, the interviewer does not provide suggestions on which study-characteristics to use.

To avoid bias toward the treatment effect in the current study, the ranking should preferable be determined without knowledge of the results found in each of the studies. As the expert is expected to be abreast of the published literature, he or she might be familiar with the treatment effects found in the historical studies. It should therefore made explicit to the expert not to include this knowledge in his or her motivation for the ordering.

For the trial by Westerhuis we asked a gynaecologist who is an expert in the field of intrapartum fetal monitoring to order the four historical studies according to the relevance for the current trial. Based on several study-characteristics she preferred the following ranking (in ascending order): Westgate, Vayssière, Ojala, Amer-Wåhlin. The technique used for ST-monitoring used in the study by Westgate was quite different from the technique used in the current trial, placing this study lowest in ranking. The severity of the indication for intrapartum fetal monitoring and the gestational age of the baby were perceived as important inclusion criteria. As the inclusion protocol by Amer-Wåhlin showed the most resemblance with the protocol of the current trial, this study was rated to be most relevant. Of the two remaining studies, the inclusion criteria by Ojala showed more similarities with the current study than the criteria by Vayssière. In constructing the power prior the power parameters for the four studies have to be in agreement with the ordering given above. Furthermore, as none of the studies is considered irrelevant, the smallest weight should be larger than zero.

2.4.2.1 Sensitivity Analysis for Multiple Historical Studies

After the expert has come to an ordering of the historical studies, an essential stage in the ranked power prior specification is the sensitivity analysis on the actual values of the assigned weights. This analysis should give more insight in the sensitivity of the posterior distribution of the treatment effect for the assigned weights. If, given a fixed ranking of the historical studies, the posterior distribution of the treatment effect remains unchanged by different values of the study weights, we conclude determining a study ranking is sufficient. However, sensitivity of the posterior distribution for changes in the power prior reveals the strength of the burden of proof captured in the current study as well as in the ranking of the historical studies. If the actual weights assigned to the studies do result in different posterior conclusions, elicitation of additional information on the weight parameters might be required. We will elaborate on this in the discussion section. In addition, posterior sensitivity to the prior parameters might lead to the conclusion that the evidence is not decisive and further research is required.

A first step in the sensitivity analysis could be to vary the range of the assigned weights, centered around 0.5. Due to the differences in sample size between the four historical studies, the actual amount of data included in the analysis varies over the different runs. Posterior results of these analyses are displayed in the first 3 rows of Table 2.5. The mean and median of the posterior distribution remained about equal, and the width of the 95% CCI decreased slightly as more data was included in the analysis. Different ranges of study weights in the analyses resulted in equal conclusions about the effectiveness of the intervention.

As a second step in the sensitivity analysis, one could vary the average of the assigned weights. The results for these analyses are displayed in the next two rows of Table 2.5. The weights in one setting (fourth line in the table) were all set below 0.5, which led the inclusion of 38% of the historical data. In the next setting each weight is doubled, which is a proportional increase of the included data to 77%. Although the effect of the amount data included is seen in the slight increase (small weights) and decrease (large weights) in the width of the 95% CCI, the posterior means and medians equaled the results in the previous analyses.

In case there are reasons to assume one of the studies is far more or far less relevant for the current study, unequal distances between the study-weights could be considered. The experts should provide a rationale to do so. In the current situation, the expert stated that the study by Westgate is different from the other studies, because this study is relatively old (1991) as compared to the other studies, and the techniques for fetal monitoring used at that time were different than used in the other studies. This could be a reason to assign a much lower weight to the study by Westgate, while keeping the ranking intact. The bottom line of Table 2.5 shows the results from this analysis. Though the posterior mean and median of the estimated treatment effect are somewhat larger than the ones found in the previous analyses, the credibility interval remained centered around one. Assigning a much lower weight to the Westgate study did not lead to a different conclusion with respect to the presence or absence of a treatment effect.

The results of this sensitivity analysis revealed that, given the fixed ordering, the posterior estimates were not particularly sensitive to the actual values of the power prior parameters. All six runs of the sensitivity analysis gave about equal results, and would lead to the conclusion that no evidence was found for an effect of ST-wave analysis over CTG in the current population.

2.5 Discussion

The power prior distribution can be used to estimate the treatment effect in a current study sample, while accounting for the information captured within previous research. Note the distinct difference between this Bayesian technique and the common meta-analytic approach. Simply combining the current with previous studies would result in an estimate of the overall treatment effect in the encompassing population instead of the treatment effect for the current population we are interested in. Just considering the current data by means of a current data analysis, would in turn ignore the extensive load of information provided by the previous studies. The power prior pro-

Table 2.5. Results sensitivity analysis using all four historical studies, given the weights of the study (a_0); posterior mean (PM), median(PMed) and the 95% central credibility interval (CCI) for the RR.

$a_0{}^*$	% Incl.	PM	PMed	95% CCI
{0.35, 0.45, 0.55, 0.65}	53.41	0.71	0.70	[0.47, 1.03]
{0.20, 0.40, 0.60, 0.80}	56.81	0.71	0.69	[0.47, 1.02]
{0.10, 0.30, 0.70, 0.90}	59.55	0.71	0.70	[0.48, 1.01]
{0.20, 0.30, 0.40, 0.50}	38.40	0.71	0.70	[0.45, 1.06]
{0.40, 0.60, 0.80, 1.00}	76.80	0.71	0.70	[0.49, 1.00]
{0.10, 0.70, 0.80, 0.90}	64.73	0.75	0.73	[0.51, 1.06]

* Ordering = Westgate, Ojala, Vayssière, and Amer-Wåhlin

vides the researcher with a tool to not only account for the information captured in historical studies, but to also control the amount of information included.

Several studies examined the use of the hierarchical power prior distribution. Modifications were proposed to ensure the power parameter distribution to reflect the degree of commensurability between the current and historical sample. We advocate a power prior distribution with pre-specified fixed study weights based on differences in study characteristics instead. By eliciting a ranking of the historical studies according to their informativeness for the current trial, we introduced a subjective element into the rather objective process of prior specification based on historical data. In the current study we confronted a single expert with a single open-ended question to come to a ranking of the historical studies. A more formal, and with that, more objective elicitation process could include consulting multiple experts, or for instance a procedure in which one expert determines the relevant study-characteristics and other experts determine a ranking based on these characteristics.

For the data on ST-wave analysis it was shown with a sensitivity analysis, that given the ordering of the studies, the exact values of the weight parameters had no impact on the final conclusions. Under different circumstances the marginal posterior for the treatment effect might not be robust for changes in the weight parameters. This emphasizes the delicacy of prior specification and shows that the actual values of the weight parameter may affect the final conclusions. Such a situation might require an extensive procedure to elicit more specific weights. The interviews could be extended with a stage in which the experts are asked to determine the lower and upper bound of the range, i.e. the weight of the least and most relevant study, as well as the total amount of information included, or to provide information on the relative distances between the study-weights.

The heterogeneity of the effect estimates found in the different studies might ask for a random effects analysis. Future research could focus on the use of a hierarchical model in combination with fixed, ordered weights for the historical studies as discussed within this paper. Furthermore, the power prior distribution evaluated in this paper assumes equal weight parameters for both intervention conditions. The study by Ojala et al. (2006) compared the use of traditional CTG with ST-wave analysis, while the trial by Westerhuis compared CTG with CTG plus ST-analysis. This makes the control condition of the historical study possibly more informative for the current study than the experimental group. It might be desirable to use a larger part of the historical data from the control condition than from the experimental condition. Further research could explore the use of unequal power parameters for different intervention conditions.

2.6 Acknowledgements

The authors thank Ming-Hui Chen for helpful discussions, Michelle Westerhuis for permission to use the dataset on ST-wave analysis in intrapartum fetal monitoring, and Anneke Kwee for her expertise in ordering the historical studies.

2.7 Appendix Chapter 2

2.7.1 Elaborate Example of the Conditional Power Prior Distribution

In the following a step-by-step procedure is described to perform Bayesian inference using R and WinBUGS. In order to perform the analyses, a recent version of WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/) should be installed, as well as the `R2WinBUGS` package (http://cran.r-project.org/). For this example a situation is assumed in which only the study by Amer-Wåhlin et al. (2001) is included in the current data analysis, with $a_0 = 0.5$.

2.7.1.1 Model

First step is to specify the model. The model given below can be written in any text-processor and should be saved as a `.txt` file. The first two lines of the code model the data in both the ST+CTG condition and the CTG-only condition. The unknown parameters of interest are the proportion of newborns suffering from metabolic acidosis in both intervention conditions. The second two lines concern the power prior distributions for these proportions. For each group the prior number of cases and group size follow a beta distribution obtained from the historical data. The last line specifies the contrast of interest, which in this case is the risk ratio (RR). The next step is to define the known elements of this model using R.

```
MODEL {
# likelihood current data
model{

    yTC ~ dbin(pTC, nTC)  #likelihood data experimental condition
    yCC ~ dbin(pCC, nCC)  #likelihood data control condition

# prior for the proportion of cases per intervention condition
    pTC ~ dbeta(alphaT,betaT)
    pCC ~ dbeta(alphaC,betaC)

# contrast (relative risk in current study)
    RRC <- pTC/pCC }
```

2.7.1.2 Data

The data from the study by Westerhuis is summarized in R as follows:

```
y.st  <- 20
n.st  <- 2827
y.ctg <- 30
n.ctg <- 2840
```

For the historical studies the same approach can be used. Data from the study by Amer-Wåhlin could for example be stored like

```
y0.st.a  <- 15
n0.st.a  <- 2144
y0.ctg.a <- 31
n0.ctg.a <- 2048,
```

giving parameter values for beta prior of

```
a.st.a   <- a0*y0.st+1
b.st.a   <- a0*n0.st+1
a.ctg.a  <- a0*y0.ctg+1
b.ctg.a  <- a0*n0.ctg+1.
```

Next the data has to be stored in a list using R. And the parameters we wish to estimate are specified.

```
# storing the data
    data <- list("y.st", "n.st", "y.ctg", "n.ctg", "a.st.a",
                 "b.st.a", "a.ctg.a", "b.ctg.a")

# specifying parameters of interest
    st.parameters <- c("pCC", "pTC", "RRC")
```

2.7.1.3 Calling WinBUGS from R

Below follows the code necessary for calling WinBUGS from R to run the model. The function `bugs` needs input on the data, the parameters to estimate, and the model to use. Furthermore, initial values, the number of chains and iterations per chain have to be specified. Optional is whether to save the history and whether to show the log-file in case an error occurs (`debug=T`).

```
# calling WinBUGS
    st.sim <- bugs(data, inits=NULL, st.parameters, "model.txt",
              n.chains=2, n.iter=10000, save.history=T, debug=T )
```

2.7.1.4 The Posterior Distribution

Before summarizing the posterior distribution, checking wether the sampler has converged is an important step. Several formal approaches to evaluate whether all chains have achieved the stationary distribution are available (e.g. Gelman and Rubin (1992a)). Furthermore, eyeballing trace-plots, running mean plots and density plots are other approaches to get an indication of the convergence of the sampler. After concluding the sampler has converged, the burn-in period should be discarded and possibly more iterations have to be made in order to obtain a large enough sample of the posterior to base inferences on.

3

Expert Elicitation of Study Weights for Bayesian Analysis and Meta-Analysis

Summary. Meta-analysis and Bayesian informative prior distributions are used for updating knowledge about treatment effects with new data. When the available data come from slightly different study populations, or from slightly different trials as compared to the new data, the researcher has to specify study weights to control the influence of the historical data on the current data. This research evaluates whether an internet Delphi technique is useful to elicit valid and reliable study weights from an expert panel. Weights were elicited for four historical studies by four experts. Despite some difficulties regarding the panel, such as respondent burden, the method seemed useful within this context. The authors advise to include sensitivity analyses to assess the required number of Delphi rounds.

3.1 Introduction

Researchers usually have access to previous studies investigating the same treatment effects as examined in their own trials. To estimate an overall treatment effect, the (aggregated) data from the 'historical' studies and the current one can be combined by doing a meta-analysis or by using the older data to construct an informative prior distribution for Bayesian analysis. In case the available studies were done in slightly different study populations, or under a slightly different design, incorporating all of the historical information might not be desirable. Therefore, when there is doubt with respect to the relevance of the historical information, the information is often excluded completely. There are however techniques to weight the historical information. In this article we propose a method for the expert elicitation of study weights for various applications in evidence synthesis.

This chapter has been published as: Rietbergen, C., Groenwold, R.H.H., Hoijtink, H.J.A., Moons, K.G.M., & Klugkist, I.(2014). Expert Elicitation of Study Weights for Bayesian Analysis and Meta-Analysis. *Journal of Mixed Methods Research*, published ahead of print October 19.

The most common way of weighing the data is by doing random-effects meta-analysis, where larger deviations in study results lead to lower study weights. Another option is to control the influence of the historical data on the estimated overall treatment effect by using the power prior distribution as proposed by Ibrahim and Chen (2000). The main principle of this technique is the specification of a weight for each historical study that reflects the differences and similarities between that study and the current one. An example of an application of this technique in medical science is given by Rietbergen et al. (2011). In this study historical data on the effect of ST-wave analysis in intrapartum fetal monitoring were included in the analysis of data obtained in a new trial. A more technical elaboration of an application in the assessment of water quality is given by Duan et al. (2006). In this research area sample sizes of the current studies are usually small and the paper discusses how historical studies and the power prior distribution can be used to deal with this problem.

The study weights can either be specified by researchers themselves or can be automatically deduced from the similarities in current and historical study results, as is done in a specific class of power prior distributions by Ibrahim and Chen (2000). Several papers investigated the properties of the latter (Neuenschwander et al., 2009; Neelon and O'Malley, 2010; Hobbs et al., 2011), and some of them conclude that it might be problematic and advise to use the fixed, user-specified value instead. With this second approach the weights will reflect the degree of commensurability in study population characteristics instead of study results. In other words, the researcher can ensure that studies with more *relevant* study results will receive higher weights, instead of studies with more *similar* results. However, the translation of relevance into valid study weights is not straightforward, and procedures for elicitation remain relatively unexplored.

The fact that little attention is given to the elicitation of study weights might be related to several difficulties that the researcher will encounter, and that will aggravate the elicitation process. First of all the study weight, should reflect the relevance of an historical study for the current one (Neelon and O'Malley, 2010). The relevance can be explained as the degree in overlap between the current and historical study populations and characteristics of the study designs (Rietbergen et al., 2011). In addition, we think that also the quality of the historical studies should be reflected in the weights. However, the Cochrane Collaboration discourages the use of quality measurement scales, since there is no empirical evidence for their validity (Higgins et al., 2011). These factors imply that the specification of study weights depends heavily on study specific substantial aspects, rather than statistical ones, making (clinical) expert knowledge crucial in this process.

To asses the reliability of the elicitation results it is desirable, if not necessary, to elicit the opinion of multiple experts. The degree of agreement among the experts is an intuitive indication of the quality of the information they provide (Dorussen et al., 2005). Reaching consensus among a panel of experts is therefore of major importance, although, of course, not sufficient for conclusions about the validity of the elicitation results, i.e. the elicited study weights.

Another important issue with respect to validity lies in the degree to which the elicited weights are influenced by study outcomes. In judging the relevance of a study the expert panel should therefore be blinded for the outcome of the study, so that weights are chosen irrespective of the study results. Since this is often unfeasible, one must prevent that the specified study weights reflect too strongly the opinion of experts who have a substantial interest in the results of the analysis of the new study. Related to this is the risk of reaching consensus via personal dominance of one of the panel members rather than expertise (O'Hagan et al., 2006).

The Delphi technique might be able to deal with these issues. With this technique a panel of experts can be consulted, and an inventory on their opinion can be made in a setting in which the absence of social pressure is ensured. This is achieved by using a procedure in which each participating expert provides answers to the questions separately from the other participants (Linstone and Turoff, 1975; Delbecq et al., 1975). In subsequent rounds, each panel member receives the anonymized results of previous rounds. Each member is then asked to react on these results, and if necessary to adjust their own answers from the previous round. This iterative process can continue until consensus among the experts is reached, or when the researcher who facilitates the Delphi rounds decides to combine the results mathematically.

The study outlined here evaluates whether an internet based Delphi technique can be used to elicit study weights from an expert panel in an efficient manner, without the loss of reliability of the elicited weights. The goal of this evaluation is two-fold. The first aim of this study is to examine whether this Delphi method can be used to reach consensus among a panel of experts with respect to the elicited study weights and the criteria on which these weights were based. Secondly, since the Delphi technique was not used to elicit study weights within this context before, we aim to provide a detailed description of the elicitation process for those researchers who plan to use this technique for a comparable purpose in the future.

A detailed description of the motivating example is presented in the second section together with the panel formation and a description of the subsequent Delphi rounds. The main elicitation results and reports on the course of each round are presented in the third section. We end with a discussion of the findings in the fourth section.

3.2 Methods

3.2.1 Motivating Example

To illustrate the use of the proposed Delphi procedure for study weight elicitation, we applied the technique to a set of studies that evaluated a drug called rosiglitazone. Rosiglitazone is an insulin sensitizer used for blood glucose lowering in diabetic patients and was subject to controversy due to alleged cardiovascular adverse effects. A recent study by Home et al. (2009) was the first randomized trial that directly assessed the effect of rosiglitazone on cardiovascular outcomes since concerns about the drug arose. We will refer to this trial as the *current* study. Previously, a large body of study reports on this subject was published (see for example Nissen and Wolski (2010) and Loke et al. (2011)). We refer to these prior studies as *historical* or *auxiliary* studies. The results from the current study could be synthesized with the available evidence from previous studies to get an optimal estimate of the safety of rosiglitazone. However, the historical randomized trials were primarily designed to evaluate the efficacy of rosiglitazone for surrogate endpoints and only reported the adverse cardiovascular events as part of the research protocol. Other historical studies that directly addressed the safety of rosiglitazone were non-randomized studies that made use of health care databases. The differences in design but possibly also the differences in the used study populations between the current and historical studies make some historical studies more relevant for answering the research question of the current study than others. Therefore, in the synthesizing the evidence some studies should receive a higher weight than others.

We acknowledge that in an empirical study of rosiglitazone, all available evidence should be taken into account. However, we aimed to illustrate the weight elicitation procedure and to limit the response burden of our expert panel at the same time. We therefore stratified the vast body of available studies according to study design - randomized control trials (RCT) versus observational studies - and selected only two studies of each type. The convenience sample of four historical studies are listed together with the current study in Table 3.1.

Table 3.1. Selected publications on the evaluation of the effect of rosiglitazone on cardiovascular endpoints.

	Study	Journal*	Objective
Current	Home et al. (2009)	Lancet	Efficacy & Safety trial
Historical	Tzoulaki et al. (2009)	BMJ	Safety study
Historical	Lipscombe et al. (2007)	JAMA	Safety study
Historical	Stocker et al. (2007)	AHJ	Efficacy trial
Historical	Hedblad et al. (2007)	JIM	Efficacy trial

* Note: Britisch Medical Journal (BMJ), Journal of the American Medical Association (JAMA), American Hearth Journal (AHJ), Journal of Internal Medicine (JIM).

3.2.2 The Expert Panel

The representativeness of a panel is determined by the qualities of the expert panel rather than by its size (Powell, 2003). In addition, a heterogeneous panel can provide higher quality solutions than more homogeneous groups (Delbecq et al., 1975). Since for this study the experts are asked to make judgements on a clinical as well as on a methodological level, we aimed at forming a panel consisting of experts with a relevant clinical background and experts with a more epidemiological or methodological background. Therefore we identified eligible potential panel members with expertise in either internal medicine, general practice, epidemiology, research methodology, or pharmacy. From the eligible experts we identified five experts who were willing to take part in this study, and covered the five fields listed above. Of this small sample, only four completed the three Delphi rounds. This final panel consisted of four experts all holding a Ph.D in their field of research, including a general practitioner (Expert 1), a vascular internist (Expert 2), a research methodologist (Expert 3), and an epidemiologist (Expert 4). No formal ethical approval was obtained for this study since the study did not concern medical scientific research on humans.

3.2.3 Description of the Delphi Rounds

The study weights were elicited using a three round Delphi process. At the start of the first Delphi round the expert panel was provided with a description of the goal of the study, and an explanation of the Delphi technique. In addition, each panel member received a file containing the references to the publications of the five studies that evaluated the effect of rosiglitazone.

In the first Delphi round the experts were asked to read the five study reports and to judge the relevance of each of the historical studies for the current one. Some of the historical studies provide evidence of the positive association between rosiglitazone use and cardiovascular disease, as myocardial infarction, stroke, and cardiovascular death. Other studies, that focuss on the effect of rosiglitazone on surrogate endpoints, however, found an opposite effect of the drug. However, the experts were asked to ignore these study results when making the judgements of relevance. As a first step, and to guide the weight specification, the experts were asked to rank order the four studies from most to least relevant. Subsequently the experts were asked to assign study weights to the four studies. The experts were instructed that the weights should reflect the relevance of the auxiliary studies expressed as the percentage of the information captured in the auxiliary study he or she was willing to include in the analysis of the current study. For example, a weight of 50 implied that the expert was willing to incorporate 50% of the historical data in the prior distribution for the new data. In a consecutive question the experts were asked to write down their motivation for the rank order and assigned weights.

The subsequent Delphi rounds focussed on reaching consensus among the experts with respect to the rank order of the studies and size of the study weights. The experts were presented with a summary of the results from the previous round. The rank orders, study weights and motivation of the other panel members were provided anonymously. However, the main discrepancies between the panel members' opinions were pointed out and linked to specific lines of written motivation of the concerning experts. The experts were asked whether they wanted to make modifications to their original answers after seeing the answers and motivations of the other participants. If they decided to maintain their rank order and weights, they were asked to give their motivation to do so. As soon as agreement on the ranks was reached no new Delphi round was started.

3.3 Results

3.3.1 Main Results: Relevant Study Characteristics, Ranking and Weights

In the three subsequent Delphi rounds a ranking of the historical studies and associated study weights were collected from the four panel members, in addition the experts were asked to motivate their choices.

Table 3.2 lists the most important study characteristics that were extracted from the written motivation (see Tables A, B and C in the Appendix for full reports). From the discussion among the experts it seems that outcome measure was the most important variable considered. Similarities in the used endpoints were a motivation for higher rankings and weights. Other important issues had to do with the design of the study, for example the type of research design that was used and the type of treatment conditions that were compared. The experts also considered the similarities in the exposed populations, and focussed on variables such as age and comorbidity.

Table 3.2. Relevant Study characteristics of the five studies as perceived by the expert panel

Study	Design	Population	Comparator	Outcome
Home (2009)	RCT	DM2, 40-75 years	Active control ^{1,2}	Cardiovascular
Tzoulaki (2009)	Retrospective cohort	DM2, 35-90 years	^{1,2,3}	Cardiovascular
Lipscombe (2007)	Nested case-control	DM2, $>$ 66 years	^{1,2,3,4}	Cardiovascular
Stocker (2007)	RCT	DM2 21-80 years	Active control ^{2}	Surrogate
Hedblad (2007)	RCT	DM2 + IRS, 35-80 years	Placebo	Surrogate

Notes: Diabetes Mellitus Type 2 (DM2), ^{1} Sulfonylurea, ^{2} Metformin, ^{3} Pioglitazone, ^{4} Other.

Table 3.3 displays the assigned rankings per study assigned by each expert per round, with changes from round to round printed in boldface. The first Delphi round already showed some agreement on the study rankings. The two RCTs are placed in third and fourth position by all experts, and the two observational studies are always placed in the two highest positions in ranking. Three rounds were sufficient for the experts to reach agreement on the ranking of the four studies as follows (from highest to lowest in ranking): 1) Tzoulaki et al. (2009), 2) Lipscombe et al. (2007), 3) Stocker et al. (2007) and 4) Hedblad et al. (2007).

Figure 3.1 shows for each Delphi round a plot of the assigned weights for each study per expert. Overall the plots indicate that the spread in assigned study weights decreases over the different rounds. In other words, it seems that over the three rounds the expert's opinions tend to converge. From the plots we can also see that for this group of experts the variation in assigned study weights is larger for the higher ranked studies than for the lower ranked studies, a phenomenon that is seen in all three Delphi rounds.

Table 3.3. Elicited rankings of the historical studies in three Delphi rounds, with the changes in ranks per expert per round printed in boldface.

	Expert 1	Expert 2	Expert 3	Expert 4
	Round	Round	Round	Round
Study	1 2 3	1 2 3	1 2 3	1 2 3
Tzoulaki (2009)	1 1 1	1 1 1	1 1 1	2 2 **1**
Lipscombe (2007)	2 2 2	2 2 2	2 2 2	1 1 **2**
Stocker (2007)	4 **3** 3	3 3 3	3 3 3	3 3 3
Hedblad (2007)	3 **4** 4	4 4 4	4 4 4	4 4 4

3.3.2 Report of Rounds

3.3.2.1 Report of Round 1

In the third part of the first Delphi round the panel members were asked to present their motivation for the chosen ranking and assigned weights. A report of the translated written motivation of the panel members is provided in Table A in the Appendix.

Both Table 3.3 and Figure 3.1 show a clear distinction between RCTs by Hedblad et al. (2007) and Stocker et al. (2007) versus the observational studies by Lipscombe et al. (2007) and Tzoulaki et al. (2009). As said, the two RCTs are placed in third and fourth position by all experts, and the two observational studies are always placed in the two highest positions in ranking. This dichotomy between the two lower and higher ranked historical studies also shows from the weights that the experts assigned to the studies as shown in Figure 3.1. A reason to place the RCTs lowest in ranking is given by one of the experts as: " *The studies by Hedblad and Stocker were less relevant because of the surrogate endpoint, about which it is unclear whether it is correlated to cardiovascular death/events.*"(Expert 2). This idea found support by two other panel members. Another panel member explains her choice by saying: " *With respect to exposure definition Stocker and Hedblad compared two mono-therapies, Hedblad even compared with placebo, while the current study compared two combi-therapies.*" (Expert 4). This also explains why the study by Hedblad was perceived least relevant by three experts. Another reason for that is given by one expert who said: " *The study by Hedblad included also patients that did not have Diabetes Mellitus yet, but only impaired glucose-tolerance.*" (Expert 2).

A reason to put the observational studies by Lipscombe et al. (2007) and Tzoulaki et al. (2009) highest in ranking is described by one researcher:" *The studies by Tzoulaki and Lipscombe appeared most relevant to me, since the primary endpoint was a direct cardiovascular endpoint.*" (Expert 2). The study by Lipscombe (2007) is perceived less relevant than the study by Tzoulaki (2009) by three out of four experts, which is explained by one of them as:" *...in the study by Lipscombe patients were only included $>$ 66 years of age, because of that a different population was being studied than in the study by Home.*" (Expert 2). Another expert remarks that: "*Lipscombe et al. (2007): Case control and because of that a larger probability of bias.*" (Expert 1). One of the experts however, perceives the study by Lipscombe et al. (2007) as the

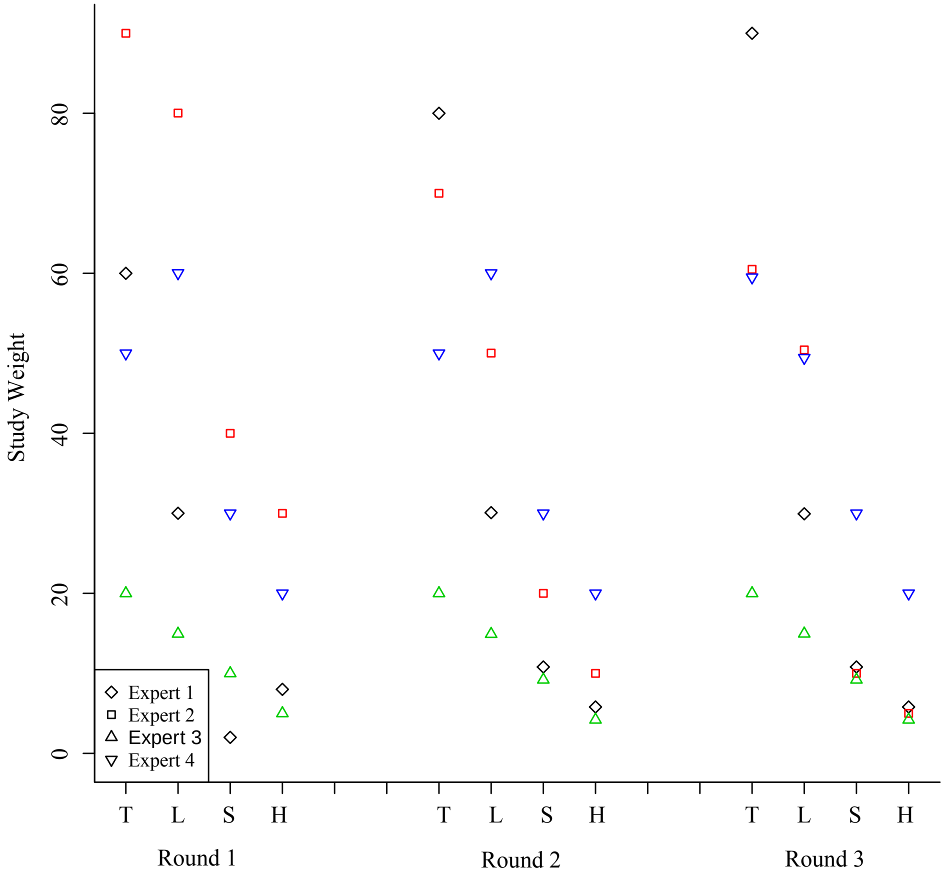

Fig. 3.1. Elicited study weights per Delphi round, where T, L, S, and H denote Tzoulaki et al. (2009), Lipscombe et al.(2007), Stocker et al. (2007) and Hedblad et al. (2007) respectively.

most relevant for the current study because: ” *Tzoulaki included hearth failure in the primary outcome measure, while this is an exclusion criterion for the current study, therefore these are less similar than the study by Lipscombe.*” (Expert 4).

The studies with the lowest ranking receive substantially lower weights than the two highest ranked studies. The weights for the lowest ranked studies are relatively close together, which is also the case for the highest ranked studies. In other words the variation in weights of the lowest ranked studies and highest ranked studies was smaller than the variation in weights between the most and least relevant studies.

In general weights were specified that covered about the entire range of possible weights; with a minimal weight for the least relevant study of 5%, and a maximum weight for the most relevant study of 90%. The weights assigned by Expert 3 are all relatively low. The expert gives the following explanation for this decision: "*... low weights were assigned because the highest studies in ranking were not RCTs but observational studies.*" (Expert 3). Since the most relevant studies were observational studies, this expert decided to assign only small weights to all studies.

3.3.2.2 Report of Round 2

In the second Delphi round the experts were asked to review the answers they provided in Round 1 in the light of the answers of the other experts. They were given the possibility to modify their initial ranking and weights, and again were asked to motivate their choices in this second round (see Table B in the Appendix for full written motivation of the experts).

The results in Table 3.3 show that Expert 1 changed the ranking of the studies by Hedblad et al. (2007) and Stocker et al. (2007) conform the ranking of the other experts: "*I have switched 3 and 4 based on the motivation of Expert 2, Stocker only concerns DM2*"(Expert 1). This resulted in panel agreement on the ranking of these two studies.

With respect to the studies by Lipscombe et al. (2007) and Tzoulaki et al. (2009), none of the experts decided to modify his or here initial ranking. All experts felt that their motivation for their initial ranking was still valid. One of the experts said: "*I was strengthened in my motivation by the fact that 2 out of 3 other experts selected the same ranking.*"(Expert 2). Expert 4, who was the only expert who preferred a second place in ranking for the study by Tzoulaki et al. (2009) and a first position for the study by Lipscombe et al. (2007) was not convinced by the arguments of the other experts to change her initial ranking. The expert gives the following three reasons for her choice: "*This study (Lipscombe et al., 2007) was a nested case control study in a large database. This means that also in this study data regarding the exposure was registered irrespective of the outcome and that the controls are from the same population as the cases. These would be the two main reasons for case control studies to be more biased than cohort studies, but these are both covered by the nested design.*"(Expert 4). She agrees with the other experts that the study population in the study by Lipscombe et al. (2007) might possibly be somewhat more different from the population in the current study than the population in the study by Tzoulaki et al. (2009). However, she remarks that: "*this is only based on the inclusion criteria, we have not seen the actual population characteristics.*" (Expert 4). With respect to the outcome measure she adds that in the study by Tzoulaki et al. (2009) it is not possible to compare the outcome measure with the one used in the current study since heart failure is included in it. For the study by Lipscombe et al. (2007) heart failure is also included as an outcome measure, however this is separately reported and therefore it can be compared to the primary outcome measure used in the current study.

Two of the experts did not change the weights they assigned to the studies in the first round. One reason for this was given by Expert 4 who said that: "*The ratio between the weights is rather similar for all experts, since this remains a difficult matter I decided to keep this equal.*" (Expert 4). Another expert pinpoints that she did not make any changes because she is still convinced that the weights should be low for all studies. She claims that: "*..., I would only have given a high weight in case a study would agree on several factors, that is a similar outcome measure, a similar population, and (perhaps) similar design. Since this was not the case, I choose to assign a low weight even to the study highest in ranking ...*" (Expert 3). For this reason she thinks that the analysis of the current study should be minimally influenced by the results in the previous studies. One expert was convinced by these arguments and changed the size of his initial weights (see Figure 3.1). He said: "*I did somewhat lower the percentages for the studies by Libscombe and Touzlaki, because I think these studies indeed have an inferior study design relative to the study by Home, as some of the other experts rightly remark.*" (Expert 2). In addition, he noticed that: "*after rereading the articles, it is not very well possible to extract a precise comparison between SUD/TZD or Metf/TZD vs SUD/Metf (like in the study by Home) from the studies by Libscombe and Tzoulaki.*" (Expert 2). This idea provided even further motivation for him to change the size of the weights. In addition, because for Expert 1 it was initially not clear that the weights should not necessarily sum op to 100, he made some modifications to the weights in the second round (see Figure 3.1).

3.3.2.3 Report of Round 3

In the third and last round the experts had a final look at the discrepancies in rankings and weights that still remained, and give their final motivation for their ranking and weights of choice (see Table C in the Appendix for full written motivation of the experts).

The results in Table 3.3 show that Expert 4 changed the ranking of the studies by Lipscombe (2007) and Tzoulaki (2009) conform the ranking of the other experts: "*I am willing to go with the other experts with respect to the ordering of the studies. Mainly, because my most important argument to stay with my original ordering was incorrect, namely the argument concerning heart failure as exclusion criterium or outcome. Which was remarked by Expert 2.*"(Expert 4). This resulted in panel agreement on the ranking of all four studies.

Expert 4 also changed the weights of studies accordingly so that they are in agreement with the new ranking. Another expert who made additional weights modifications in this final round was Expert 2. This expert thought that the previously assigned weights for Hedblad et al. (2007) and Stocker et al. (2007) were too high. Furthermore, the expert decreased the weight for the study by Tzoulaki et al. (2009) since: "*... this study does not provide information about adding rosiglitazone to the metformin/SUD condition, like was done in the design of the Home study.*" (Expert 2). The same sort of adjustment in the weight for this study was seen with Expert 1, although no further motivation for this change was provided by the expert.

Discussion

We evaluated whether the Delphi technique can be used as a tool to elicit expert judgement on the relevance of a set of historical studies for a new one. Overall, this study suggests that the Delphi method is useful to create agreement in a group of diverse experts on the ranking of studies and to a great extent also on the quantification of perceived study relevance. The experts stood open for each other's arguments and motivations and eventually were able to convince each other of what would be the best ranking. In our study three Delphi rounds were sufficient to reach agreement between the experts with respect to the ranking of all four studies. This quick convergence of the ranking might be partly due to limited initial variability of the ranks. Already in the first Delphi round the experts agreed on the two most relevant and two least relevant studies for the current trial. It seems reasonable to assume that stronger disagreement between the experts at the start will result in either more Delphi rounds, or even in failure to achieve convergence.

With respect to the study weights the results show that already at an early stage there was more agreement among the experts about the weights for the two lower ranked studies, than for the higher ranked studies. The results clearly showed that over the different rounds the variation in selected study weights tended to decrease as the experts took notice of each other's arguments. This also stresses the importance of multiple expert elicitation, in contrast to the consultation of a single expert. Although overall the selected weights tend to converge, the difference in variability for the two lowest ranked studies and two highest ranked studies remained, which might indicate that the experts found it easier to judge the lower ranked studies than the higher ranked studies.

Despite of the favorable results, some issues might have influenced the elicitation process. We do not know to what extent this had an effect on the validity of the resulting study weights. However, we would like to pinpoint these issues in order for other researchers to consider them when designing such an elicitation study.

In order to make a well informed judgement on the relevance and quality of the historical studies, panel members have to read a considerable amount of literature. Because of this, the response burden on the experts can become quite high. We therefore decided to work with a relatively small, but heterogeneous expert panel. Although a larger panel could increase the reliability of the elicitation results, it is unclear whether a larger panel would improve or reduce the efficiency of the elicitation process. For example, one of the main disadvantages of the internet based Delphi technique is the variation between the panel members in time to submit a response. Although for each of three rounds we planned the deadline for response two weeks after sending the invitation for participation, the time to response ranged from one day to several months. The delay in response time frustrated the efficiency of the process because panel members sometimes had to wait a long time before the new round could be started. One solution for this problem could be found in sticking to a strict deadline for participation in each round. However, we fear that this would have resulted in panel attrition, which becomes directly problematic for small size panels. Another solution might be found in more direct contact with the panel members. In this study the contact between the researchers and the expert panel was established through e-mail. We think however that it might be worthwhile to introduce a more direct way of contact between the researchers and panel members (e.g. by phone or face-to-face) in order to motivate panel members more directly. However, this seems only feasible for smaller panels.

Furthermore, in this study, we asked the panel members to rank order and weight four historical studies. Often however, many more historical studies might be available and interesting to include. It is likely that more studies to rank increases the response burden of the panel. Not only because of the increasing number of papers to read, but with that it will also become more difficult to distinguish between the different studies. In this research a feasible number of studies with deviating study designs were assessed for their eligibility. Probably more rounds will be needed to reach consensus when more studies with smaller differences are included in the synthesis of the new and historical data.

In this Delphi study we were satisfied with full agreement on the study ranks, and partial convergence of the study weights. In case absolute consensus on the actual weights among the experts is required, more rounds will be needed. A sensitivity analysis to evaluate the influence of the disagreement on the estimated treatment effect might help to determine the need for absolute consensus. In addition, a sensitivity analysis can help to determine when to stop organizing more Delphi rounds. Future research on the use of the Delphi method within this context might focus on these aspects.

To summarize, in this study the Delphi method was successful in reaching consensus among a group of experts on the ranking of a set of studies. This was established in few rounds, and without a face-to-face discussion among the panel members. In addition, with the method we approached convergence on the quantification of the relevance of the individual studies in this specific situation. These promising results might inspire other researcher to adopt a similar approach when they have to make judgements on the quality and relevance of historical studies in circumstances where complete inclusion or exclusion of information is not desirable. We encourage researchers with a comparable elicitation question to farther explore the possibilities of this Delphi approach, of course, with careful consideration of the difficulties described here.

3.4 Acknowledgements

The authors thank the participating experts for their time and valuable contributions.

3.5 Appendix Chapter 3

Table A - Translation of written motivation of all experts in Round 1.

Expert 1
”1.Tzoulaki (2009): Cohort study (the data were collected without awareness of the research question; this advocates a form of prospective research), Open population (GP database), Large and diverse population, 15 years follow-up, Real life setting: up until 90 years of age; limited exclusion, Clinical endpoint and directly interpretable. 2. Lipscombe (2007): Case control and because of that a larger probability of bias, Open population, Large and diverse population, 3.8 years follow-up, Primary as well as secondary care, Participants > 66 years, TZD: at that time separately registered (and paid for); bias, Clinical endpoint and directly interpretable. 3. Hedblad (2007): Middle large population, Increased risk of CV events Diabetes and IRS. Bias by indication, Laboratory endpoint / proxy. 4. Stocker (2007): Small population, Secondary/tertiary care. Bias by indication, Laboratory endpoint / proxy,”
Expert 2
”I weighed whether the patient population was similar to the population of the Home study and whether the primary endpoint was relevant for the Home study. The studies by Tzoulaki and Lipscombe appeared most relevant to me, since the primary endpoint was a direct cardiovascular endpoint. The study by Tzoulaki was even a bit more relevant because there was no clear age limit for inclusion: in the study by Lipscombe patients were only included > 66 years of age, because of that a different population was being studied than in the study by Home. The studies by Hedblat and Stocker were less relevant because of the surrogate endpoint, about which it is unclear whether it is correlated to cardiovascular death/events. The study by Stocker appeared a bit more relevant because only DM 2 patients were included, that were suboptimal controlled, as is the case in the study by Home. The study by Hedblat included also patients that did not have DM yet, but only impaired glucose-tolerance.”
Expert 3
”I wanted to base the relevance on design (RCT vs non-RCT), outcome (same outcome vs intermediate outcomes), population (DM2 vs non-DM2, old vs same age), intervention. I based the ranking on the outcome: two studies used the same outcomes as Home et al, these are highest in ranking because I miss the clinical background to judge what the effect of the intervention is, and because I think that the outcome under study is an adverse event, because of which it might be less relevant what the population is (young vs old, DM2 vs non-DM2). So priority in ranking is based on the agreement in outcome measure, and low weights were assigned because the highest studies in ranking were not RCTs but observational studies.”
Expert 4
”I looked at inclusion and exclusion criteria, the population (which country) and the exposure and outcome definition. In doing so, I took the outcome definition as the most important and exposure definition as the second most important. Hedblad and Stocker both have surrogate endpoints and no primary outcome measures, because of which the results are less relevant for the current study. Tzoulaki included hearth failure in the primary outcome measure, while this is an exclusion criterion for the current study, therefore these are less similar than the study by Lipscombe. The study by Lipscombe also included heart failure as an endpoint but separately analyzed from the other outcome measures. With respect to exposure definition Stocker and Hedblad compared two mono-therapies, Hedblad even compared with placebo, while the current study compared two combi-therapies. Because in the studies by Tzoulaki and Lipscombe different mono-therapies and combi-therapies were compared, you can extract relevant information from them for the current study.”

Table B - Translation of written motivation of all experts in Round 2.

Expert 1
”RANKING: I have switched 3 and 4 based on the motivation of Expert 2, Stocker only concerns DM2. WEIGHTS: I thought indeed that the weights should sum to 100, therefore now round numbers. I do not agree with Expert 3 who says that because the high ranked studies are not RCT they should not receive too high weights. The observational studies, like Tzoulaki, are more real life.”
Expert 2
”I have not made changes to the ranking, because I still think that my initial motivation for this ranking is valid. In fact, I was strengthened in my motivation by the fact that 2 out of 3 other experts selected the same ranking. The fourth expert deems the study by Lipscombe more relevant than the study by Tzoulaki because the latter included heart failure as a primary outcome measure,while this an exclusion criterion in the study by Home. The study by Home, however, considers this as an important outcome measure as well (although patients are not allowed to have had heart failure before randomization): this is for me a good reason to place the study by Tzoulaki highest in ranking indeed. I did somewhat lower the percentages for the studies by Lipscombe and Tzoulaki, because I think these studies indeed have an inferior study design relative to the study by Home, as some of the other experts rightly remark. In addition, after rereading the articles, it is not very well possible to extract a precise comparison between SUD/TZD or Metf/TZD vs SUD/Metf (like in the study by Home) from the studies by Lipscombe and Tzoulaki.”
Expert 3
”The explanation provided by the other experts did not motivate me to change the ranking. I assigned low weights, because I would only have given a high weight in case a study would agree on several factors, that is a similar outcome measure, a similar population, and (perhaps) similar design. Since this was not the case, I choose to assign a low weight even to the study highest in ranking, in other words I choose for a minimal influence of the results of the previous studies on the new analyses.”
Expert 4
”I did not make any changes to the ranking because I consider my arguments for this ranking still valid. The other experts switched Lipscombe and Tzoulaki in the ranking. Reasons for doing so were the case control design by Lipscombe, and the fact that the study by Lipscombe only included ¿ 66 year old patients. I do not consider the argument that a case control study would by definition lead to more bias, a valid argument. This study was a nested case control study in a large database. This means that also in this study data regarding the exposure was registered irrespective of the outcome and that the controls are from the same population as the cases. These would be the two main reasons for case control studies to be more biased than cohort studies, but these are both covered by the nested design. There are possibly more differences between the study from Lipscombe and the Home study, than between the study by Home and the one by Tzoulaki. However, this is only based on the inclusion criteria, we have not seen the actual population characteristics. I consider my argument that we can extract heart failure from the outcome measure in the Lipscombe study and not in the study by Tzoulaki still valid, and therefore I stick to my decision to put Lipscombe in first position and Tzoulaki in second place. The ratio between the weights is rather similar for all experts, since this remains a difficult matter I decided to keep this equal.”

Table C - Translation of written motivation of all experts in Round 3.

Expert 1
”Two arguments two place Tzoulaki first and Lipscombe second in ranking that **were** not discussed yet: 1) Tzoulaki 7.1 years follow-up versus 3.8 years follow-up in Lipscombe study, so a higher probability of events in Tzoulaki study, 2) In Lipscombe study a higher probability of bias due to payment for TZD use. In my opinion this aspect remained under-exposed in the discussion.”
Expert 2
”After reading the articles again I do think I assigned too high weights to the studies by Hedblad and Stocker. I also assigned a lower weight to the study by Tzoulaki because this study does not provide information about adding rosiglitazone to the metformin/SUD condition, like was dan in the design of the HOME study.”
Expert 3
”I do not have reasons to change the weights, I do have a suggestion to perform a sort of sensitivity analysis with the weights 10-20-30-40, and 15-30-45-60.”
Expert 4
”I am willing to go with the other experts with respect to the ordering of the studies. Mainly, because my most important argument to stay with my original ordering was incorrect, namely the argument concerning heart failure as exlusion criterium or outcome. Which was remarked by Expert 2. I changed the weights of the studies by Lipscombe and Tzoulaki accordingly.”

4

Power prior distributions and sampling variability

Summary. The power prior distribution in Bayesian statistics allows for the inclusion of data and results from previous studies into the analysis of new data. It enables the researcher to control the influence of the historical data on the new data, by specifying a prior parameter that determines the amount of historical data to be included. This parameter can either be a fixed user-specified value, or can be estimated from the data. For the latter, current literature states that the size of the weight parameter should depend on the commensurability of the historical and new study outcomes. In this research we question whether this is desirable, since differences between study results might be caused by sampling variability.

Illustrated with a numerical example and a real data application we show the joint power prior provides posterior estimates farther from the true value than the estimates provided by a power prior with a fixed, self-chosen weight parameter. This supports our supposition that coincidental differences between new and historical data can affect the posterior distribution for the power parameter, and consequently the parameter of interest, especially for smaller samples.

We advocate the inclusion of additional (expert) knowledge on the commensurability of the new and historical study populations, since only the commensurability of sample results can never fully justify the value of the weight parameter.

4.1 Introduction

The Achilles heel of Bayesianism is depicted in the deriding description of the Bayesian as someone who, vaguely expecting a horse, and catching the glimpse of donkey, strongly believes he has seen a mule (Senn, 2007). It refers to the delicate process of prior specification to prevent drawing fruitless posterior conclusions after combining prior and data. Often an uninformative prior distribution is specified in an attempt to

This chapter is based on: Rietbergen, C., Chen, M.-H., & Klugkist, I. Power prior distributions and sampling variability. (Under revision).

deal with this problem. Using information from previous studies for prior specification is another way to try to decrease the probability of inadequate prior specification and, with that, invalid posterior conclusions.

Differences between the new and previous studies make the inclusion of historical data for prior specification a challenging matter. To control the influence of the historical data on the new study Ibrahim and Chen (2000), therefore, propose the use of a power prior distribution. The weight parameter in this special class of prior distributions enables the historical data to be weighed relative to the new data. This parameter could either be a carefully selected fixed parameter, or could be considered a random parameter for which a hyper prior is to be specified. In case of the latter, current literature states that the (marginal) posterior distribution for the weight parameter should depend on the commensurability of the results of the historical and new data (Hobbs et al., 2011; Duan, 2006; Neuenschwander et al., 2009).

In this paper we question whether it is desirable to let the data, i.e. the commensurability of new and historical study results, determine the weight assigned to the historical data. Differences in sample statistics between studies can be systematic or can be naturally occurring sampling variability, i.e. the variation due to unsystematic sampling error. Sampling variability is often underestimated and various papers, in for example psychological research, are dedicated to this issue. For example Maxwell (2004) and Cumming et al. (2004) explain how sampling error can lead to erroneous conclusions about results obtained in social research if we focus on (small sample) single studies. In current power prior literature the existence of sampling variability and its implications for the estimation of the distribution for the weight parameter remains underexposed.

Illustrated with a numerical example and a real data application we discuss how sampling variability might induce bias in the estimation of the marginal posterior for the weight parameter. We show how this can lead to distorted posterior estimates for the measure of interest. We make a case for a specification procedure of the weight parameter that is, at least partially, based on the commensurability of the study populations themselves, instead of a procedure that fully depends on the commensurability of study results.

In Section 4.2 we will first present some frequentist properties and technical aspects of the specification of the power prior with fixed weight. We continue with the specification of the power prior distribution with random weight. The possible influence of sampling variability on the estimation of the marginal posterior for the weight parameter will be discussed in Section 4.3. The process is illustrated with a numerical example and real data application. Implication of the findings will be discussed in Section 4.4.

4.2 Types of Power prior distributions

4.2.1 Power prior distributions with fixed weights

The power prior distribution for some parameter vector θ given the historical data D_0 and a fixed weight parameter a_0 as proposed by Ibrahim and Chen (2000) is obtained by multiplying an initial prior $\pi_0(\theta)$ with the likelihood of the historical data raised to a power a_0, as in

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta), \tag{4.1}$$

where and $0 \leq a_0 \leq 1$, and $a_0 = 0$ indicates no inclusion of historical data, and $a_0 = 1$ equals full inclusion of the historical data. As this prior distribution conditions on a_0 it is often referred to as the conditional power prior distribution. Often, no information in addition to D_0 with respect to θ will be included, that is, a low informative initial prior $\pi_0(\theta)$ is specified.

4.2.2 Frequentist properties of the posterior estimates for normal data

In this subsection, we examine frequentist properties of the conditional power prior for normal data with known σ^2 and σ_0^2.

Let $D = (y, n)$ denote the current data with $y = (y_1, y_2, \ldots, y_n)'$, where $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$, and also let and $D_0 = (y_0, n_0)$ denote the historical data with $y_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})'$, where $y_{i0} \overset{iid}{\sim} N(\mu, \sigma_0^2)$, $i = 1, 2, \ldots, n_0$. We further assume that D and D_0 are independent and σ^2 and σ_0^2 are known. We take an improper uniform initial prior for μ, i.e., $\pi_0(\mu) \propto 1$. Then, the power prior of μ given D_0 with a fixed a_0 in (4.1) is given by

$$\pi(\mu|D_0, a_0) \propto \exp\Big\{ -\frac{a_0 n_0}{\sigma_0^2}(\mu - \bar{y}_0)^2\Big\},$$

where $\bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_{0i}$, and the corresponding posterior distribution of μ is given by

$$\pi(\mu|D, D_0, a_0) \propto \exp\Big\{ -\frac{1}{2}\Big(\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}\Big)\Big(\mu - \frac{\frac{n}{\sigma^2}\bar{y} + \frac{a_0 n_0}{\sigma_0^2}\bar{y}_0}{\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}}\Big)^2\Big\}, \tag{4.2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Using (4.2), we obtain the posterior mean and the posterior variance of μ as

$$\bar{\mu} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{a_0 n_0}{\sigma_0^2}\bar{y}_0}{\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}} \text{ and } \mathrm{Var}(\mu|D, D_0, a_0) = \Big(\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}\Big)^{-1}. \tag{4.3}$$

Consequently, we obtain the frequentist variance of $\bar{\mu}$ as

$$\mathrm{Var}(\bar{\mu}) = \frac{\frac{n}{\sigma^2} + a_0^2\frac{n_0}{\sigma_0^2}}{\big(\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}\big)^2}. \tag{4.4}$$

Although the variance in the historical data is different than the one in the current data, the posterior mean $\bar{\mu}$ is an unbiased estimator of μ as long as an improper unform initial prior is specified for μ.

Result 1: *Using (4.3) and (4.4), we have*

$$\mathrm{Var}(\bar{\mu}) \le \mathrm{Var}(\mu|D, D_0, a_0) \tag{4.5}$$

for $0 \le a_0 \le 1$. *In addition, the equality in (4.5) holds if and only if* $a_0 = 0$ *or* $a_0 = 1$ *and the maximum difference between* $\mathrm{Var}(\mu|D, D_0, a_0)$ *and* $\mathrm{Var}(\bar{\mu})$ *is reached at* $a_0 = 0.5$.

It is easy to see that

$$\mathrm{Var}(\mu|D, D_0, a_0) - \mathrm{Var}(\bar{\mu}) = \frac{a_0(1-a_0)\frac{n_0}{\sigma_0^2}}{\left(\frac{n}{\sigma^2} + \frac{a_0 n_0}{\sigma_0^2}\right)^2} \ge 0.$$

Thus, the proof of Result 1 is straightforward. Result 1 implies that the frequentist coverage probability of the $100(1-\alpha)\%$ highest posterior interval will exceed a prespecified credible level $1-\alpha$, where $0 < \alpha < 1$ for $0 < a_0 < 1$. To see this, we let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$ percentile of the standard normal $N(0,1)$ distribution and using (4.2) and (4.3), we obtain the $100(1-\alpha)\%$ highest posterior interval as

$$\begin{aligned} &I(1-\alpha, D, D_0, a_0) = \\ &\{\mu :\ \bar{\mu} - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\mu|D, D_0, a_0)} < \mu < \bar{\mu} - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\mu|D, D_0, a_0)}\}. \end{aligned} \tag{4.6}$$

Then, we are led to the following result.

Result 2: *The frequentist coverage probability is given by*

$$P(\mu \in I(1-\alpha, D, D_0, a_0)) = 2\Phi\Big(\frac{\sqrt{\mathrm{Var}(\mu|D, D_0, a_0)}}{\sqrt{\mathrm{Var}(\bar{\mu})}} z_{1-\alpha/2}\Big) - 1, \tag{4.7}$$

where $\Phi(\cdot)$ *denotes the* $N(0,1)$ *cumulative distribution function. Consequently, we have*

$$P(\mu \in I(1-\alpha, D, D_0, a_0)) \begin{cases} = 1-\alpha & \text{if } a_0 = 0 \text{ or } a_0 = 1, \\ > 1-\alpha & \text{if } 0 < a_0 < 1, \end{cases} \tag{4.8}$$

and the maximum coverage probability is given by

$$P(\mu \in I(1-\alpha, D, D_0, a_0 = 0.5)) = 2\Phi\Bigg(\sqrt{\frac{\frac{n}{\sigma^2} + 0.5\frac{n_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + 0.25\frac{n_0}{\sigma_0^2}}} z_{1-\alpha/2}\Bigg) - 1. \tag{4.9}$$

Remark 1: Results 1 and 2 imply that when the historical data and the current data are comparable in terms of the parameter μ, the efficiency will be lost when the historical data are not fully borrowed. Also note that the results derived here are slightly more general than the ones obtained in Ibrahim et al. (2014) since we assume different variances in the historical and current data. In addition, we derive an explicit formula of the frequentist coverage probability of the HPD interval.

4.2.3 Power prior distributions with random weights

Despite the useful properties of the conditional power prior distribution as shown above, specification of this prior can be difficult, since a fixed weight has to be determined by the user. To overcome this issue Ibrahim and Chen propose to specify a prior distribution for a_0. The resulting joint power prior distribution for (θ, a_0) is given by

$$\pi(\theta, a_0|D_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta)\pi(a_0|\gamma_0). \tag{4.10}$$

Here, $\pi(a_0|\gamma_0)$ is the (hyper) prior distribution for a_0 with hyperparameter vector γ_0. Ibrahim and Chen note that by using this joint power prior the tails for the marginal distribution of θ, better reflect the uncertainty about a_0 than the marginal posterior following from the conditional power prior distribution that assumes fixed a_0.

Evaluations of the performance of the power prior (see for example Duan (2006) and Neuenschwander et al. (2009)) revealed that, for binary and normal data, the joint power prior distribution as given in (4.10), with an uninformative prior for a_0, results in a marginal prior for a_0 with values close to zero. To solve this problem, Duan (2006) and also Neuenschwander et al. (2009) propose a power prior distribution that includes the normalizing constant. This modified power prior distribution normalizes the (marginal) prior distribution for a_0, resulting in larger values for the weight parameter. This modified power prior distribution takes the form

$$\pi(\theta, a_0|D_0) = \frac{L(\theta|D_0)^{a_0}\pi_0(\theta)\pi(a_0|\gamma_0)}{\int L(\theta|D_0)^{a_0}\pi_0(\theta)d(\theta)}. \tag{4.11}$$

The joint power prior distribution as presented in (4.10) and (4.11) builds on the assumption that both the new and historical data are needed to update the distribution for a_0. For the case of a single historical study this implies that a higher weight is assigned to the historical study when its results show more resemblance with the new sample results.

4.3 Sampling variability and the estimation of a_0

The weight parameter a_0 can be seen as the probability that the new and historical samples come from the same population (Duan, 2006). Values of a_0 closer to zero indicate it is highly *unlikely* the samples come from the same underlying population and values closer to one indicate it is highly *likely* the samples come from the same population. This probability is based solely on the sample results, and therefore can only be correct in two situations. First, in case of similar sample results the similarities should not be the result of overlapping sampling distributions of the new and historical situations. Second, in case of different sample results the differences should be systematic and not the manifestation of sampling variability.

Figure 4.1 shows four plots with sampling distributions to illustrate and explain the situations described above. In each case two samples are drawn to estimate some population parameter θ: a new sample N denoted by a dot, and an historical sample H denoted by a square. The new and historical samples can both be random draws from the same population with population mean θ as in the top row of the figure. In this situation the estimated value for a_0 will be correct in case the sample statistics are similar, but incorrect in case the sample statistics are different. In the bottom row of the figure, the historical and new sample are random draws from two different populations with mean θ_H for the historical population and θ_N for the new population. In this case the estimated a_0 will only be correct in case the sample statistics are different, and will be incorrect in case the sample statistics are similar. The problem, however, is that we can never know from the data whether the observed differences or similarities between the sample estimates are random or systematic. And in case we are wrong the resulting marginal posterior for a_0 is the product of chance, thereby overestimating or underestimating the true value of the weight parameter, and shifting the posterior estimate for θ towards a result observed by chance.

To illustrate this further, we describe for a number of situations the sensitivity of the joint power prior results for sampling variability and make a comparison with the conditional power prior. Figure 4.2 presents a selection of situations that might occur if we want to estimate a population parameter θ with new sample N (displayed in circles) and historical sample H (displayed in squares).

Let's assume that both samples N and H are both random draws from the same underlying population with population mean θ. In this scenario the observed differences between each new and historical sample estimate for θ are a manifestation of sampling variability.

Since we do not have any additional knowledge about samples N and H we fix $a_0 = 1$ for the conditional power prior distribution. For the situations presented in the figure we can now describe how sampling variability can influence the posterior distributions for a_0 and θ under both prior distributions.

When the new and the historical sample estimates are close together, as in situation 1 in Figure 4.2, the estimated posterior mean for a_0 for the joint power prior will be close to one. The conditional and joint power prior will result in similar posterior estimates of θ, with estimates closer to θ in situation 2.1.a in 4.2 and farther from θ in situation 2.1.b in 4.2.

Larger differences between the samples (see situation 2 in Figure 4.2), as may occur with smaller samples, lead to an increased disregard of the historical data under the joint power prior distribution (smaller posterior values for a_0). Since in situation 2.a in Figure 4.2 the historical estimate is closer to the true value than the new estimate, including more historical data results in estimates of θ closer to the true value. The joint power prior will disregard the historical data, pulling the posterior estimate for θ toward the new sample result and away from the true population mean. The opposite is true for situation 2.b in Figure 4.2, where including less of the historical data will provide results closer to the truth. In situation 2.c in Figure 4.2, with extreme sample results, equal inclusion of new and historical data, as is done with the conditional prior with $a_0 = 1$ will give results closest to the true value.

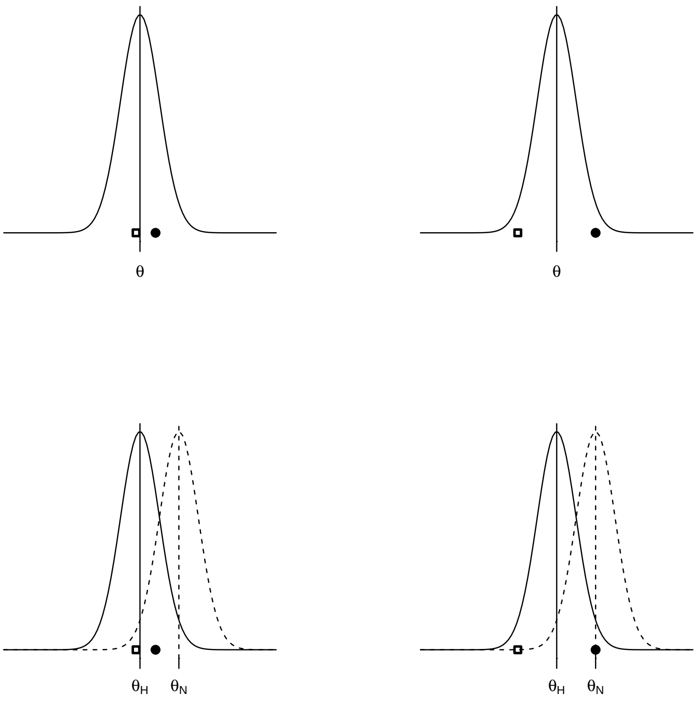

Fig. 4.1. Sampling distributions with the new and historical sample statistics indicated by dots and squares respectively.

To further illustrate the possible bias due to sampling variability when using the joint power prior we present a numerical example and real data application in which we compare the two types of power prior distributions.

4.3.1 An example with normal data

In Table 4.1 we compare the posterior results from the analysis of four pairs of historical and new data sets using the conditional and joint power prior distributions. Both the new and historical data sets, related to IQ measurements, are random draws of $n = 20$ from a common normal population with mean $\mu = 100$, and known population standard deviation $\sigma = 15$. We fix $a_0 = 1$ under the conditional power prior, and estimate a_0 in the joint power prior (3) using a non-informative Beta(1,1) initial prior distribution. The first two columns present the sample means $\bar{x}$ and $\bar{x}_0$, for each pair of new and historical data sets respectively. The third and fourth column show

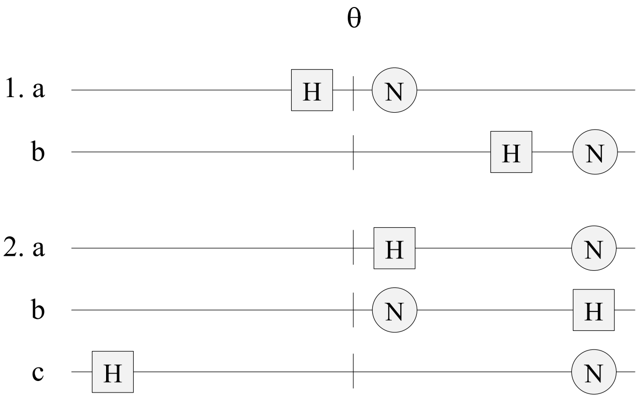

Fig. 4.2. Possible positions of new samples N and historical samples H, relative to the true population parameter θ.

the posterior means for μ under the conditional and joint power prior. For the latter, the marginal posterior mean for a_0 is given within brackets.

Table 4.1. New and historical sample estimates $\hat{x}$ and $\hat{x}_0$; Posterior estimates for μ (posterior means), under the conditional power prior (CPP) with $a_0 = 1$, and joint power prior (JPP) with within brackets the marginal posterior mean for a_0; for $N = 20$

		$\hat{\mu}$	
$\bar{x}$	$\bar{x}_0$	CPP	JPP (a_0)
110.65	103.11	106.85	108.32 (0.49)
102.18	108.08	105.10	104.06 (0.52)
108.25	88.54	98.36	106.26 (0.12)
88.54	108.25	98.36	90.54 (0.12)

The first row, with $\bar{x} >> \mu$ and $\bar{x}_0 > \mu$, resembles situation 2.a in Figure 4.2. As expected, we find $\hat{\mu}_{conditional} = 106.85$ to be a bit closer to μ than $\hat{\mu}_{joint} = 108.32$. For $\bar{x}_0 >> \mu$ and $\bar{x} > \mu$, as in situation 2.b in the figure, the second row of the table shows slightly better results for the joint power prior, with $\hat{\mu}_{joint} = 104.06$ versus $\hat{\mu}_{conditional} = 105.10$.

The extreme lack of commensurability between the two sets in the third row (see situation 2.c in Figure 4.2), has caused a large decrease in the posterior mean for a_0. Therefore, much of the information captured in the historical data is ignored under the joint prior, leading to an overestimate of μ with $\hat{\mu}_{joint} = 106.26$. Under the conditional power prior, where the new and the historical sample receive equal weights, the estimated mean IQ $\hat{\mu}_{conditional} = 98.36$ is very close to the true population mean of 100 points. In addition, exchanging the historical and new set, as presented in the fourth row, $\hat{\mu}_{conditional}$ remains at 98.36 points. Under the joint distribution $\hat{\mu}_{joint} = 90.54$, which implies a big shift towards $\bar{x}$, that is towards an extreme result observed by chance.

Taking a closer look at the marginal posterior estimates for a_0 in Table 4.1, it immediately shows how the weight parameter is increasingly underestimated when the historical and new data are farther apart. Especially, the last two rows in the table show how the joint power prior might provide poor estimates for a_0, and subsequently the parameter of interest.

4.3.2 Real data example: the safety of rosiglitazone

In the previous section we evaluated a situation for which we knew the new and historical samples were random draws from the same underlying population. This way the value for the weight parameter could be fixed to one. When presented with real life data a researcher can never be sure about whether the new and historical sample share a common underlying population. To illustrate this, the following example discusses such a situation historical data is available for inclusion in a power prior distribution although the degree of relevance is uncertain.

In a randomized control trial (RCT) by Gerstein et al. (2010) on the efficacy of rosiglitazone in preventing cardiovascular disease in 333 type 2 diabetes patients using this antidiabetic drug between 2005 - 2008 the researchers observed eight events of myocardial infarction (MI). In a previously conducted RCT by Dargie et al. (2007) on the same therapeutical intervention, researcher observed seven events of MI in 108 type 2 diabetes patients in the period 2001 - 2003. In a previously conducted observational study by Lipscombe et al. (2007) on a comparable therapeutical intervention, researchers found 53 events of MI in 229 patients between 2002 - 2005. In Table 4.2 some characteristics of the three studies are presented, the data of the three studies are displayed in Table 4.3. To estimate the risk of MI from the new RCT data, we specify a conditional and joint power prior distribution using the data from either the previously conducted RCT or the observational study as historical prior input.

For specification of the conditional power prior a fixed value for the weight parameter a_0 had to be chosen based on the degree of commensurability between the study and patient characteristics of the most recent RCT and the historical studies. To do so we would need clinical expertise to determine the commensurability between the studies on the most important of these characteristics. Since elaborate expert elicitation falls beyond the scope of this article, we interviewed a single expert in the field to provide a selection of six most relevant study characteristics. For both historical studies we subsequently counted the proportion of items on which the historical study was similar to the new RCT. The resulting proportions were interpreted as values for a_0. Based on the agreement in characteristics between the studies Table 4.2 shows

Table 4.2. Selection of relevant study characteristics of new and historical study.

	Gerstein (2010)	Dargie (2007)	Lipscombe (2007)
Data collection	2005 - 2008	2001 - 2003 (-)	2002 - 2005 (-)
Age (in years)	61 (9)	64.3 (8.8) (+)	73.9 (5.7) (-)
Insuline	Included	Included (+)	Excluded (-)
Cardiovascular history	Included	Included (+)	Included (+)
Rosiglitazone alone	Yes	Yes (+)	All TZDs (-)
Country	Canada	UK (-)	Canada (+)
Score		4 / 6 = .67	2 / 6 = .33

Table 4.3. Data of new and historical studies; number of events, and total sample size per study (N).

	Gerstein (2010)	Lipscombe (2007)	Dargie (2007)
Events	8	53	7
Total N	333	229	108

$a_0 = .67$ for the historical RCT, and $a_0 = .33$ for the historical observational study. For both models an uninformative Beta(1,1) initial prior was specified for the parameter of interest θ. In addition, for the joint power prior an uninformative Beta(1,1) hyperprior for the weight parameter a_0 was specified.

Table 4.4 presents the results of the analyses of the new data using either the RCT or observational study as power prior input in both models. The estimates for θ and accompanying 95% CIs are presented, as well as the fixed and estimated values for a_0. The results show how the estimates for a_0 under the joint power prior are in the same direction of the pre-specified fixed values for a_0; under both power priors the historical observational study will receive a much lower weight than the historical RCT. However, the location of these values do differ, i.e. the estimated values for a_0 are closer to zero than the pre-specified values. This means that under the conditional power prior a larger part of the historical information was included, and therefore the posterior results for θ are influenced in a higher extent by the historical data than the results under the joint power prior. In the table this effect shows from the resulting estimate for θ under the conditional power prior when using the observational study by Lipscombe data as prior input. Since the raw estimate in this study is much higher than the one in the RCT by Gerstein, and more of this historical information is included, the estimate for θ is much higher than in the other situations.

The agreement in the relative size of the weight parameters under both power prior distributions suggest that both procedures might be heading in the right direction. However, the actual size of the weight parameters remains uncertain. Does the pre-specified weight parameter overestimate the relevance of the historical studies? Or are the estimated values under the joint prior the result of coincidental differences between the new and historical sample estimates and therefore too small? Unfortunately, since, in this situation, one cannot know the true value of a_0, the questions remain unanswered.

Table 4.4. Results of analysis under both power prior distributions.

Type of power prior	$\hat{\theta}$	95% CI	a_0
Dargie (2007)			
Conditional	.033	[.018;.053]	.67
Joint	.032	[.016;.052]	.48
Lipscombe (2007)			
Conditional	.065	[.043;.090]	.33
Joint	.029	[.014;.050]	.20

4.4 Discussion

Current literature on the power prior distribution focussed mainly on the higher degree of objectivity associated with the joint power prior in comparison to the conditional power prior distribution. Since the degree of commensurability between the new and historical data determines the size of the weight parameter, no subjective judgements with respect to the value of this parameter have to be made. Our main concern with this approach was that with disregarding the subjective input inherent to the specification process, we make way for an increasing influence of chance. That is, differences or similarities between new and historical studies might be the manifestation of sampling variability, and therefore should not be used as a motivation for weight specification.

In Section 4.3 we evaluated the extent to which sampling variability distorts the marginal posterior estimates of the weight parameter, and with that the posterior estimates for the parameter of interest. The numerical example presented in Section 4.3 supports our idea that in some cases the use of the joint power prior does lead to posterior estimates that are farther from the true value than the estimates provided by the conditional power prior. Coincidental differences between the new and historical data can affect the posterior distribution for the power parameter a_0, and the parameter of interest consequently. This implies that when working with smaller (new) samples, the effect of sampling variability can be worrisome. This, while at the same time the inclusion of results from prior studies can be particularly interesting when the new sample size is relatively small.

In the current paper we did not illustrate the situation where both samples come from (somewhat) different populations. Therefore, we did not evaluate how coincidental similarities might lead to the overestimation of the commensurability of the new and historical data. Incautious use of the joint prior distribution under these conditions might lead to the crossing of apples and oranges. That is, one might combine study outcomes that are unrelated due to similarities observed by chance, and thereupon draw fruitless posterior conclusions. Of course, it is good practice in meta-analysis-like settings, to make well argued choices about what studies to include, based on study and population characteristics of the studies in question. Given this careful process of study inclusion, very extreme situations will (hopefully) never occur. However, when a researcher judged a set of studies to be eligible to be included in the analysis, would it be reasonable to assign study weights close to zero, as happened in the real data example in Section 4. Here, the estimated value for the weight parameter for the study by Lipscombe was only .10. Consequently only a very small part of the historical data was included in the prior for the analysis of the new data. If a study is down weighed to such a large extent, then why was it considered eligible in the first place?

The above shows how only the commensurability of *sample results* can never be fully sufficient for the justification of the value of a_0, and we advocate the inclusion of additional knowledge with respect to the commensurability of the new and historical study *populations*. Qualitative knowledge on the commensurability of the new and historical data is provided by the careful inclusion process of the prior studies. This implicit knowledge can be translated into fixed study weights for the conditional power prior distribution as was done in Rietbergen et al. (2011) and Rietbergen et al. (2014). Another option is to incorporate this information in a hyper prior distribution for a_0 in the joint power prior distribution. In the presented examples we used an uninformative hyper prior for a_0, this way only the commensurability of the new and historical data influenced the posterior distribution for the weight parameter. However, one might want to specify an informative hyper prior to guide the estimation of the weight parameter. This way we combine both procedures, and acknowledge the value of the subjective judgement of relevance and quality next to the objective process of joint power prior specification.

Especially when only a limited amount of new data is available, the inclusion of information obtained in previous studies could be attractive. When doing so, a sensitivity analysis should always be performed to evaluate the effect of the chosen weights or hyper prior parameters on the posterior estimate of the parameter of interest. Quantification of the available knowledge to make it suitable for prior specification remains a big challenge. Therefore, future research on power prior specification should also focus on the elicitation of the information to specify the weight parameter itself, or the parameters of the hyper prior for the weight parameter.

5

Evidence synthesis in drug safety assessment: the example of rosiglitazone

Summary. The current system of benefit-risk assessment of medicines has been criticized for relying on intuitive expert judgment. There is a call for more quantitative approaches and transparency in decision making. Illustrated with the case of cardiovascular safety concerns for rosiglitazone we aimed to explore a structured procedure for the collection, quality assessment and statistical modelling of safety data from observational and randomized studies.

We distinguished five stages in the synthesis process. In Stage I the general research question, population and outcome and general inclusion and exclusion criteria are defined and a systematic search is performed. Stage II focusses on the identification of sub-questions examined in the included studies and the classification of the studies into the different categories of sub-questions. In Stage III the quality of the identified studies is assessed. Coding and data extraction are performed in Stage IV. Finally, meta-analyses on the study results per sub-question are performed in Stage V.

A Pubmed search identified 30 randomized and 14 observational studies meeting our search criteria. From these studies, we identified 4 higher level sub-questions and 4 lower level sub-questions. We were able to categorize 29 individual treatment comparisons into one or more of the sub-question categories, and selected study duration as an important covariate. We extracted covariate, outcome and sample size information at the treatment arm level of the studies. We extracted absolute numbers of myocardial infarctions from the randomized study, and adjusted risk estimates with 95% confidence intervals from the observational studies. Overall, few events were observed in the randomized studies which were frequently of relatively short duration. The large observational studies provided more information since these were often of longer duration. A Bayesian random effects meta-analysis on these data showed no significant increase in risk of rosiglitazone for any of the sub-questions.

The proposed procedure can be of additional value for drug safety assessment because it provides a stepwise approach that guides the decision making in increasing process trans-

This chapter is based on: Rietbergen, C., Stefansdottir, G., Leufkens, H.G.M., Knol, M., Klugkist, I., Grobbee, D.E., & De Bruin, M.L. Evidence synthesis in drug safety assessment: the example of rosiglitazone. (Under revision).

parency. The procedure allows for the inclusion of results from both randomized an observational studies, which is especially relevant for this type of research.

5.1 Background

The current system of benefit-risk assessment of medicines has been criticized as it primarily relies on intuitive expert judgement (Coplan et al., 2011) and there is a call for more quantitative approaches and transparency (Guo et al., 2010). With respect to the risk-arm of the benefit-risk balance, safety information from different sources accumulates throughout the life cycle of the products (Weaver et al., 2009; O'Neill, 1998). At market approval, information on adverse drug reactions (ADRs) of drugs comes from pre-clinical studies and randomized controlled trials (RCTs) whereas post-marketing data mostly include spontaneous ADR reports and epidemiologic studies. Regulators base their pharmacovigilance decisions on both pre-marketing and post-marketing data, which can be conflicting and of deviating relevance and quality, and hence difficult to integrate into a single opinion.

A typical example of a product where information on (cardiovascular) safety accumulated throughout the products life cycle causing an ongoing debate is rosiglitazone (Diamond et al., 2007; Kaul et al., 2010; FDA and Administration, 2011). Rosiglitazone is an insulin sensitizer used to treat diabetes type II, which was approved by the United States Food and Drug Administration (FDA) in 1999 and by the European Medicines Agency (EMA) in 2000. Subsequently, rosiglitazone was suspended from the European market by the EMA in 2010 due to cardiovascular risk, while it still remains marketed in the United States under severe restrictions (FDA and Administration, 2011; EMA, 2010). The decision to withdraw rosiglitazone from the EU market was based on data that accumulated during the post-marketing phase through use in the general population, which tends to differ from the trial population. The different labels of rosiglitazone in Europe and the US and subsequent market withdrawal in Europe shows how the evaluation of evidence in different regulatory systems can lead to different decisions. Discrepancies like these occur often and the regulatory systems could benefit from a structured approach to come to a more consistent conclusion.

For integrating information from different sources, post-marketing safety evaluation could benefit from an evidence synthesis strategy for data from both randomized controlled trials (RCTs) and observational studies. Especially in combination with judgements of quality and relevance to enable overt combining of data from different sources. Previously, some efforts have been made to combine information from RCTs and observational studies (Kuoppala et al., 2008; Bertrand et al., 2012). Bayesian statistics can be a useful tool, since data from RCTs and observational studies can either be jointly modelled to estimate an effect, or the observational data can serve as input for the specification of prior distribution for the analysis of the RCT data, or the other way around.

The aim of this paper is to implement such a Bayesian approach in which the prior distribution is derived from observational data, within a structured procedure for data gathering and quality assessment to combine safety data from RCTs and observational studies. We used the cardiovascular safety of rosiglitazone as an example. With this we aim to add to the operationalization of the framework provided by Coplan et al. (2011) and to provide the regulators with a tool to structure the decision making when data from many sources are available.

5.2 Methods

Figure 5.1 presents our integrated approach comprising five stages for searching and combining relevant study results from different sources. In the following we elaborate on each stage of the process and at the same time apply this approach to the rosiglitazone example.

5.2.1 Stage I

Step 1: Defining the research question, population and outcome The assessor has to clearly specify the main research question that one is interested in, and with that, the outcome and population of interest. A preliminary literature search at this point can aid the decision-making with respect to these elements.

Based on information from the literature and the different conclusions about the safety of rosiglitazone we posed the following overall research question, outcome definition and population description: Does rosiglitazone increase cardiovascular risk in otherwise healthy adult patients with type II diabetes. A quick scan of several available studies on this topic revealed a great variety in the interpretation of the specified outcome. Some studies reported the total number of myocardial infarctions, strokes, and cardiovascular deaths. Some reported only on one of these major adverse cardiovascular events (MACE). We decided to focus on myocardial infarction (MI) only since the majority of the studies so far reported information on this specific event. With respect to the treatment conditions all possible comparators were considered and listed i.e. placebo, no treatment, and other diabetic agents. However, some studies did not report the use of any control group. These studies, in which rosiglitazone was not compared with any other treatment or placebo, were excluded at this stage since these could not contribute to answering the research question.

Step 2: Data sources, searches, and general inclusion criteria The inclusion and exclusion criteria have to be listed, and a proper query to search for relevant publications has to be specified.

We performed a Pubmed search, searching for any randomized controlled trials (RCTs) and observational studies on rosiglitazone among adult patients, published before December 31st 2010. All studies that mentioned rosiglitazone in the title or abstract were searched. Furthermore, overall inclusion criteria included studies with a duration of at least 24 weeks that included a comparable non-exposed group. We allowed for studies with specific populations e.g. Mexicans and Taiwanese. Only original research articles in English were considered for inclusion. Furthermore, it was required that the number of myocardial infarction (MI) events during the study period was mentioned in the result section, or that a safety section was included that discussed all major adverse events during the study period, whether MI was mentioned or not. If MI was not mentioned in this section it was considered to have zero events. Since our domain was patients with type II diabetes that were otherwise healthy, only studies that included patients with diabetes type II were considered for inclusion. The search identified 683 abstracts, after excluding non eligible studies either with no original data, case-reports, in-vitro studies, animal studies and/or studies without rosiglitazone, 91 publications were retrieved (see Figure 5.1). From these publications we excluded 47 studies in which either the study population did not include patients with type II diabetes (23 studies), the studies were not of sufficient duration (14 studies), there was no comparable exposure group (6 studies) and/or adverse events during the study period were not listed (4 studies). Finally, 14 observational studies and 30 RCTs were considered for inclusion for one or more of our research questions (see Figure 5.1 and separate reference list).

5.2.2 Stage II: Inclusion of studies per research question

Although, the information on the safety outcomes of interest is reported in each of the studies selected in Stage I, important differences between the studies might exist with respect to the exact research questions addressed. Some studies are designed to examine the efficacy of the drug under study compared to placebo or other therapies, while others are designed to directly assess the safety of the drug. As a consequence, simple pooling of all available data ignores the underlying safety questions that can actually be answered with the different studies. Therefore, in Stage II we propose an approach in which a close inspection of the used study designs is made in order to extract the actual research (sub) questions that are addressed. Subsequently, one should extract from each study only those treatment arms that are relevant for one or more of the specified sub questions. In doing so, the originally intended comparisons should be retained and one should never extract single study arms from any of the studies. Furthermore, study arm selection may not be influenced by study results, i.e. the selection process should take place without consideration of the study results.

For the example of rosiglitazone we extracted different types of research questions. Figure 5.2 shows the four higher level questions (1-4) and four lower level sub questions (a-d). For each specific research question we assessed the relevance of the treatment arms and whether the study included a comparable non-exposed group. All studies considered for inclusion were reviewed by one of two researchers, and in case of uncertainties reviewed by both. Studies were included only if consensus was reached.

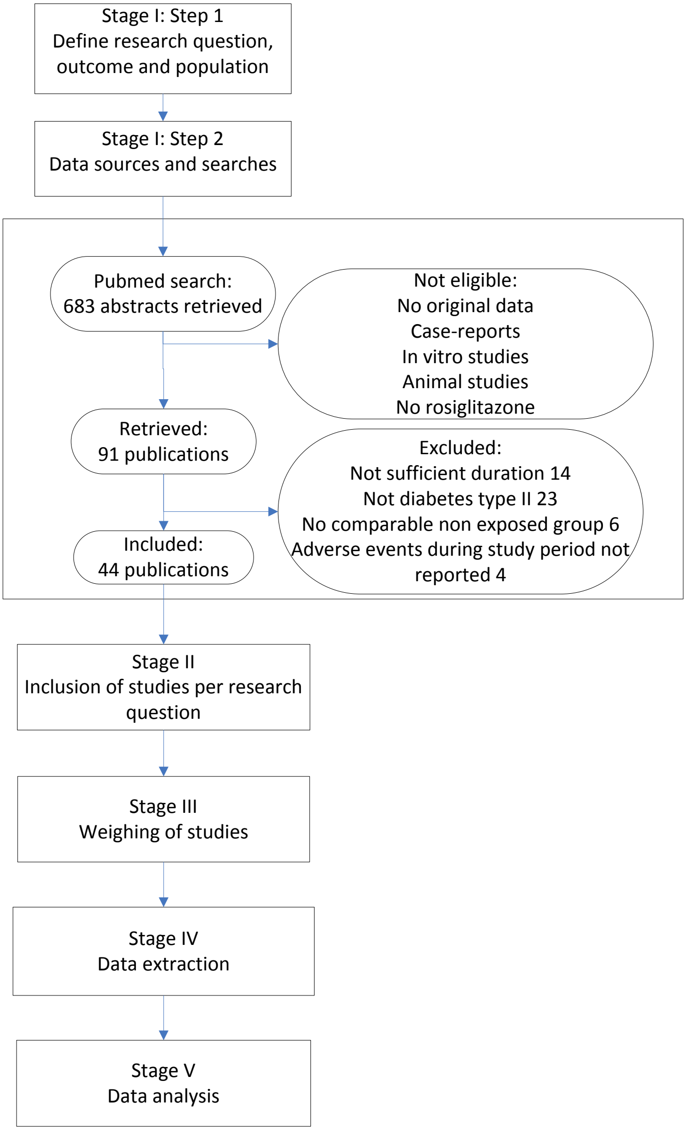
Stage I: Step 1
Define research question, outcome and population
Stage I: Step 2
Data sources and searches
Pubmed search: 683 abstracts retrieved
Not eligible:
No original data
Case-reports
In vitro studies
Animal studies
No rosiglitazone
Retrieved: 91 publications
Excluded:
Not sufficient duration 14
Not diabetes type II 23
No comparable non exposed group 6
Adverse events during study period not reported 4
Included: 44 publications
Stage II
Inclusion of studies per research question
Stage III
Weighing of studies
Stage IV
Data extraction
Stage V
Data analysis

Fig. 5.1. Safety of Rosiglitazone research questions.

The upper part of Figure 5.2 addresses research questions 1 and 2. For research question 1, concerning the risk of rosiglitazone compared to no treatment (or placebo), we included a) studies that had rosiglitazone only arms compared to either placebo or untreated controls (which included observational studies properly adjusted for other glucose control treatments) and b) studies that evaluated rosiglitazone plus another glucose control agent (rosiglitazone as add-on therapy) vs. the same glucose control agent as monotherapy. For research question 2, that concerned the risk of rosiglitazone compared to other treatments, we included c) studies that had arms comparing rosiglitazone monotherapy with another monotherapy and d) studies that compared dual therapy with rosiglitazone and another agent (rosiglitazone as add-on therapy) versus treatment with that same agent plus another glucose control agent (as an add-on).

The lower part of Figure 5.2 addresses research question 3, about the risk for MI associated with rosiglitazone monotherapy, which is a combination of a and c and research question 4, on the risk of rosiglitazone add-on therapy, which is a combination of b and d.

Table A in the Appendix presents a list of all included studies and the selected study arms for each comparison (a, b, c, and d) and the relevant study characteristics. To explain the selection procedure we take the example of the observational study by McAffee et al. (2007). This study included several treatment arms where rosiglitazone was prescribed both as a monotherapy and add-on therapy. Since rosiglitazone was compared to metformin as well as sulphonylurea monotherapy the study is listed twice in Table A under sub question c. In addition, rosiglitazone was used as add-on to metformin, sulphonylurea and insulin and compared to treatment arms where these treatments were used with add-on of sulphonylurea, metformin and other diabetes agents, respectively. Therefore, these treatment arms were included for sub question d.

Another example is the RCT by Home et al. (2009) which randomized patients on metformin to either rosiglitazone or sulphonylurea and patients on sulphonylurea to either rosiglitazone or metformin. Based on this randomization we would have included all four study arms in sub question d. However, in the analysis the researchers combined both rosiglitazone arms and compared it with both non-rosiglitazone arms. This introduced a problem of whether there was a comparable non-exposed group. What was compared in the end is a group of patients on rosiglitazone and either metformin or sulphonylurea with patients using both metformin and sulphonylurea. Therefore, we concluded that this comparison should be included in sub question b. Since the results were not adjusted for background medication use (metformin and sulphonylurea), this study was considered less optimal than the observational studies with the same comparisons that do adjust for co-medication.

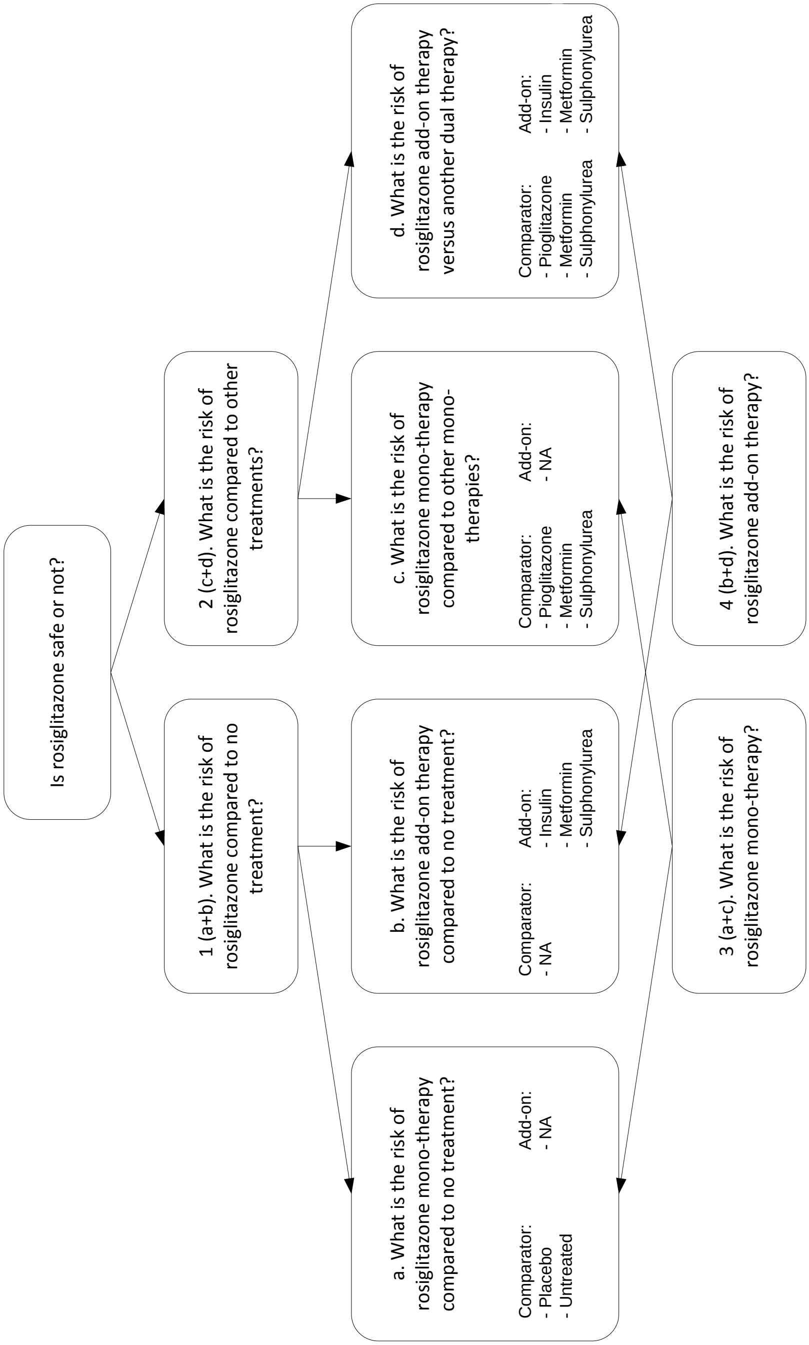
Is rosiglitazone safe or not?
1 (a+b). What is the risk of rosiglitazone compared to no treatment?
2 (c+d). What is the risk of rosiglitazone compared to other treatments?
a. What is the risk of rosiglitazone mono-therapy compared to no treatment?
Comparator:
- Placebo
- Untreated
Add-on:
- NA
b. What is the risk of rosiglitazone add-on therapy compared to no treatment?
Comparator:
- NA
Add-on:
- Insulin
- Metformin
- Sulphonylurea
c. What is the risk of rosiglitazone mono-therapy compared to other mono-therapies?
Comparator:
- Pioglitazone
- Metformin
- Sulphonylurea
Add-on:
- NA
d. What is the risk of rosiglitazone add-on therapy versus another dual therapy?
Comparator:
- Pioglitazone
- Metformin
- Sulphonylurea
Add-on:
- Insulin
- Metformin
- Sulphonylurea
3 (a+c). What is the risk of rosiglitazone mono-therapy?
4 (b+d). What is the risk of rosiglitazone add-on therapy?

Fig. 5.2. Research questions on the safety of rosiglitazone .

5.2.3 Stage III: Quality assessment

An important stage in this procedure is the assessment of quality and relevance of the selected studies. Different scales are available to assess the quality. The Cochrane risk of bias tool for RCTs (Higgins et al., 2011) to assess the weight of randomized studies takes into account method of study treatment allocation and concealment, blinding, completeness of outcome data and reporting and other sources of bias. The Newcastle-Ottawa quality assessment scale (Wells et al., 2013) that was designed to assess the risk of bias in case-control studies and cohort studies, consists of three sections that take into account selection, comparability of groups and exposure in case-control or outcome in cohort studies. The quality scores can be transformed into study weights such that studies with lower quality receive less weight in the meta-analysis and studies of higher quality receive higher weights. In the ongoing debate about the use of quality scores as weights in meta-analysis (Bown and Sutton, 2010; Jüni et al., 1999; Herbison et al., 2006, see)many experts argue that using such weights might induce bias in the estimation of the treatment effect of interest. Since we do attach importance to the process of quality assessment, we propose to use the quality judgment to set a criterion for study (arm) inclusion. For the example of rosiglitazone we set a cut off value (lower limit) of 0.7 for studies to be included in our meta-analyses, but different choices in this respect can be made.

We used the above mentioned tools to assess the quality of the included randomized and observational studies in the rosiglitazone example. Since the Cochrane risk of bias tool allows for self-specified potential threats of bias we also included representability of the study population, duration (> 24 weeks) and size (>1000 patients). For each topic on the scale a study could score 1 point making up a total of 10 points. The final weight is represented as a percentage of the maximum 10 points. The Newcastle-Ottawa scale consists of 3 sections which take into account selection, comparability of groups and exposure in case-control or outcome in cohort studies. Each study could get a maximum of 9 points. The final score was represented as a percentage of the total 9 points and is presented for each study in Table B in the Appendix.

Due to the nature of these scales studies that are substantially different may receive the same weight. For example, the Newcastle-Ottawa scale allows the user to specify the most important factors that determine the comparability of cases and controls. Each study can earn one point if the included cases and controls match on these factors. A second point can be earned if the study matches cases and controls on an additional important factors. We selected age and gender as primary matching factors and diabetic co-medication and previous cardiovascular events as important additional factors. The studies by Dormuth et al. (2009) and Dore et al. (2009) both received two points for comparability. Unlike Dormuth et al. and many other studies, Dore et al. additionally adjusted for previous diagnosis of obesity and smoking which are important risk factors for cardiovascular disease. The used quality scale, however, does not allow to account for these additional factors.

5.2.4 Stage IV: Data extraction

In the fourth stage the focus is on data extraction of general study characteristics, information about the experimental conditions and study outcome. From the randomized studies the number of adverse events per study arm and accompanying size of the arms have to be extracted. In addition, adjusted risk estimates and standard errors from all included study arms of the observational studies have to be extracted. The resulting data can be found in Table B. In this stage variables that might have influenced the study outcome and therefore have to be included as covariates in the final analyses should be considered. For different studies different variables might be of importance. For some adverse events the latency time (time to event) is much longer than the time needed to measure the efficacy of a drug, hence, the duration of the studies is an important covariate. In other cases the year of publication might be especially relevant for example when there are substantial changes to a drugs label which will affect the population that is being exposed to the drug.

From the randomized trials on rosiglitazone we extracted information on the absolute number of myocardial infarction in all relevant treatment arms and the number of patients in each arm. From the observational study publications we extracted all adjusted odds ratios (OR) or hazard ratios (HR) and associated 95% confidence intervals for MI (see Table B). Furthermore, from all included publications we extracted information on publication year, baseline medication (untreated, wash-out period, continued treatment with metformin, sulphonylureas, insulin or other glucose lowering treatment), comparison treatment (metformin, sulphonylureas, insulin or other glucose lowering treatment), mean age (in years at baseline), male rate, and duration in weeks. Information was collected at study arm level. Consequently, study information may vary from comparison to comparison.

5.2.5 Stage V: Data analysis

In the final stage the data extracted in Stage IV should be arranged per research question, in order to include them in the analyses. Decisions on the statistical models to use have to be made at this stage as well.

For the rosiglitazone example we used and compared three models: Model A) a crude analysis of all studies, Model B) a crude analysis of studies with weight ≥ 0.7, and Model C) an analysis of studies with weight ≥ 0.7 adjusted for study duration. The rarity of the outcome event of interest in both the treated and untreated (i.e. unexposed to rosiglitazone) patient groups allowed for pooling of odds ratios, risk ratios and hazard ratios. Therefore, we will from now on refer to all three ratios as odds ratios. A Bayesian random effects meta-analysis was performed, in which the adjusted odds ratios and associated standard errors reported in the observational studies were pooled and used to derive the prior distribution for the meta-analysis of the randomized studies.

We performed a random effects meta-analyses to pool the observational data per research (sub) question, which consisted of adjusted risk ratios and their 95% confidence intervals. The pooled effect estimates, as presented in Table 5.1, were used to derive an informative prior distribution per research (sub) question. Since, large populations were used in the included observational studies and since the risk of bias is much larger for these studies than for randomized trials, we adopted a power prior approach Ibrahim and Chen (2000) to limit the influence of the observational data on the estimated effect. In this approach the likelihood of the (pooled) observational data is raised to the power α. If this parameter is set to zero the observational data is fully discounted, while a value equal to one would allow full inclusions of the observational evidence (for a simple introduction on the application of the power prior distribution we refer to Rietbergen et al. (2011)).

To monitor the size of the influence of the posterior we ran the analyses for each research question with three different values for α. First, we used $\alpha = 0$ to fully ignore the observational data and $\alpha = 1$ to fully include observational evidence. In addition we determined the size for α based on the variance of the estimated treatment effect in the RCTs. That is, we shrunk the size of the weight parameter such that the variance of the pooled effect found in the observational studies was as large as the variance of the pooled effect in the RCTs. Because the full data was available for the randomized studies, the following model was used for the Bayesian random effect meta-analysis for this part of the data (see also Welton et al. (2012)):

$$r_i{}^C \sim Bin(n_i{}^C, \pi_i{}^C)$$

$$r_i{}^T \sim Bin(n_i{}^T, \pi_i{}^T)$$

$$\mu_i = logit \pi_i{}^C$$

$$logit \pi_i{}^T = \mu_i + \delta_i$$

$$\delta_i \sim N(\delta, \tau^2)$$

where $r_i{}^C$ and $r_i{}^T$ are the estimated risk in study i in the control group and treatment group respectively. Furthermore, $\delta_i = logit(\pi_i{}^T - \pi_i{}^C)$ is the log-odds ratio in study i, which follows a normal distribution with mean δ and between-study variance τ^2. Calculating odds ratios for all RCTs required a continuity correction for those studies with empty cells. To decrease the problem of possible swamping of the real effect, 0.1 was added to all cells instead of the usual 0.5, with one exception: the randomized data for research question d were so sparse that 0.5 was added to the cells to enable estimation at all.

Although in each analysis different combinations of studies and study arms were included, the same model was used. The data for all studies and study arms included per analysis are presented in Table B. In addition to Model B, in Model C a study level covariate to adjust for the duration of the study was added to the model. These analyses were only conducted for those research questions for which multiple observational studies as well as multiple intervention studies could be included. All analyses were performed using OpenBUGS 3.2.1. and R (code available upon request), we used non-informative prior distribution for all parameters other than the estimated treatment effect.

5.3 Results

Overall, we found 58 treatment arm comparisons from 30 RCTs and 14 observational studies. From these studies we included 7 study arm comparisons for research question a (1 observational and 6 RCTs), 16 study arm comparisons for research question b (1 observational and 15 RCTs), 21 study arm comparisons for research question c (13 observational and 8 RCTs) and 14 study arm comparisons for research question d (8 observational studies and 6 RCTs). The majority of the patients included in the trials were men above 50 years of age. Nearly half of included study arm comparisons had duration between 24 and 52 weeks (28 comparisons, 48.3%); consequently, the overall duration of exposure to rosiglitazone was relatively short considering that diabetes is a chronic condition requiring long term treatment. In the first years after marketing of rosiglitazone only randomized studies were found as expected, from 2007 onwards we found publications of observational studies as well. The characteristics of included study arm comparisons per research question can be seen in Table A in the Appendix.

The number of MIs in the randomized studies and the adjusted risk estimates (hazard ratios and odds ratios) along with the risk of bias weights can be found in Table B. Overall, few events were observed in the randomized studies, many did not report any events of MI. RCTs of longer duration such as the one by Home et al. (2009) and Kahn et al. (2006) reported MI events in both patients exposed to rosiglitazone and the comparison group.

Table 5.1 presents the results per model for each research question. For each model we present the results for analysis with prior weights $\alpha = 0$, $\alpha = 1$ and with α chosen such that the precision in the observational studies is as large as in the RCTs. By means of this sensitivity analysis we could evaluate the influence of the prior on the posterior estimates. For research question a we could not present results for the analysis in which we only included high quality studies, since only one prior study was available, and this study was of poor quality.

The estimates for the models in research questions a, b, d, 1 and 4 gave very unstable results when no or little prior information was taken into account. Although the estimated mean effect sizes were sometimes large, the associated credibility intervals were so large, that we cannot interpret the point estimates. This problem is caused by the fact that the included studies for these research questions reported very few cases of MI. Interestingly, for research question a, all MI events in the rosiglitazone arms of the randomized studies came from the same study. This study is characterised by its long duration (52 weeks) and an exclusive inclusion of patients with a history of cardiovascular disease. Notably, the older studies (published in 2005 or earlier) are shorter than the ones published later. However, the results for Model C as presented in the last column of Table 5.1 indicates that adjusting for study duration did not noticeably change the results. This result was supported by a simple plot the effect sizes against study duration, which showed no relationship between the two.

Table 5.1. Results of the Bayesian-meta analysis per sub questions; prior weight (α), mean ES (mean), median ES (med), and lower and upper bounds of the 95% Central Credibility Interval (95%LB and 95%UB respectively).

		Model A				**Model B**				**Model C**			
Question	Prior weight (α)	mean	med	95% LB	95% UB	mean ES	med	95% LB	95% UB	mean	med	95% LB	95% UB
a	$\alpha = 0$	12.54	0.61	0.01	34.77	14.16	0.59	0	41.33	55.84	0.37	0	123.7
	$\alpha = 1$	1.79	1.71	0.95	3.06	xx	xx	xx	xx	xx	xx	xx	xx
	$\alpha = 0.0013$	6.59	0.67	0.01	32.79	xx	xx	xx	xx	xx	xx	xx	xx
b	$\alpha = 0$	3.48	1.55	0.48	15.94	5.43	1.66	0.44	25.25	7.3	2.54	0.42	35.58
	$\alpha = 1$	1.04	1.04	0.94	1.14	1.04	1.04	0.94	1.14	1.04	1.03	0.94	1.14
	$\alpha = 0.0002, 0.00006, 0.00003$	2.81	1.53	0v48	12.48	4.73	1.64	0.64	23.5	6.52	2.45	0.46	33.74
c	$\alpha = 0$	1.65	1.34	0.47	3.72	1.69	1.35	0.44	3.79	1.38	1.04	0.32	3.48
	$\alpha = 1$	1.05	1.05	0.99	1.12	1.04	1.04	0.95	1.13	1.03	1.03	0.95	1.12
	$\alpha = 0.0014, 0.0023, 0.0029$	1.36	1.3	0.61	2.49	1.37	1.3	0.59	2.56	1.11	1.02	0.43	2.33
d	$\alpha = 0$	3.56	1.73	0.21	16.44	35.72	2.64	0.07	140.9	132.2	2.91	0.02	339.6
	$\alpha = 1$	0.97	0.96	0.83	1.12	0.97	0.96	0.83	1.12	0.97	0.96	0.82	1.12
	$\alpha = 0.00041, <0.00001, <0.00001$	2.97	1.69	0.22	12.47	41.59	2.53	0.06	137.3	287.9	2.98	0.02	287.5
1 (a+b)	$\alpha = 0$	2.25	1.51	0.4	8.4	2.71	1.55	0.35	11.32	3.12	1.88	0.34	12.69
	$\alpha = 1$	1.08	1.07	0.98	1.17	1.04	1.04	0.94	1.14	1.03	1.03	0.94	0.14
	$\alpha = 0.00051, 0.0003, 0.0002$	1.89	1.42	0.39	6.12	2.26	1.5	0.37	8.67	3.05	1.87	0.34	12.52
2 (c+d)	$\alpha = 0$	1.65	1.43	0.67	3.81	1.75	1.44	0.63	4.38	1.61	1.22	0.46	4.25
	$\alpha = 1$	1.05	1.04	0.98	1.11	1.03	1.03	0.96	1.1	1.02	1.02	0.95	1.1
	$\alpha = 0.0015, 0.0014, 0.0014$	1.41	1.35	0.73	2.44	1.47	1.39	0.67	2.84	1.33	1.19	0.53	3.01
3 (a+c)	$\alpha = 0$	1.48	1.32	0.34	3.52	1.49	1.32	0.29	3.78	1.27	1.02	0.21	3.69
	$\alpha = 1$	1.07	1.07	1.01	1.14	1.04	1.04	0.95	1.13	1.03	1.03	0.95	1.12
	$\alpha = 0.0014, 0,0024, 0,0024$	1.35	1.29	0.51	2.6	1.35	1.29	0.48	2.66	1.15	1.04	0.36	2.61
4 (b+d)	$\alpha = 0$	3.01	1.74	0.65	12.33	5.94	1.98	0.62	26.53	7.48	3.26	0.69	36.9
	$\alpha = 1$	1.03	1.02	0.95	1.1	1.03	1.03	0.95	1.11	1.02	1.02	0.95	1.1
	$\alpha = 0.00016, 0.00003, 0.00003$	2.54	1.68	0.64	9.37	4.58	1.95	0.63	22.72	6.9	3.22	0.65	32.67

Research question c (rosiglitazone mono-therapy versus any other mono-therapy) was the only question for which we could include a number of studies in which a substantive amount of MI cases were reported in both the control and treatment arms of the trials. Therefore, only for research question c and for the two questions including c, that is 2 and 3, we found stable results for the meta-analysis for the RCTs. No significant effects were found for research question c and 2 and the majority of models for question 3, meaning that we did not find support in these data for the expectation that patients on rosiglitazone mono-therapy are at higher risk for MI than patients on any of the other mono-therapies. Nevertheless, these results show nicely how including more prior information pulls the effect size from the estimated effect in the RCTs towards the overall effect size found in the observational studies. At the same time, including more information reduces the size of the 95% credibility intervals, indicating that including prior information provides more confidence in the estimated effect. Take for example research question c model C, here the estimated odds ratio found in the meta-analysis of the RCTs alone was found to be equal to 1.38 with 95% credibility interval between 0.32 and 3.48. Adding some information obtained in the observational data pulls the estimated mean in the direction of the OR equal to 1.03 as found in the observational studies, resulting in a posterior mean equal to 1.11. At the same time the 95% interval was reduced to an interval closer around 1 with a lower bound equal to 0.43 and upper bound equal to 2.33.

Only for research question 3 with model A we found a borderline significant effect of 1.07 with 1.01-1.14 95% CI in case the observational data were fully included. This research question asked whether patients receiving rosiglitazone mono-therapy were at higher risk for MI compared to patients receiving no treatment (a) or patients receiving any other mono-therapy. According to this estimate we could conclude there is some evidence that patients receiving rosiglitazone only are at somewhat higher risk. However, by using only a small amount of observational information ($\alpha = 0.0014$), or by omitting the low quality observational studies and RCTs from the analysis we could not reproduce this close to significant effect, and found the lower bounds of the credibility intervals shifting back to < 1.

5.4 Discussion

In this paper we propose a procedure for drug risk assessment which may be used by regulators or others that work in the field of pharmacovigilance. The objective of the procedure was to guide and formalize the process of post-marketing safety evaluation and decision making. We proposed a five stage approach with which results from carefully selected Phase III and Phase IV studies can be combined. Using rosiglitazone as an example in a meta-analysis, offered a complete case study to clarify the proposed procedure and to illustrate the difficulties that can be encountered when evaluating safety.

One of the first difficulties we encountered in the first stage had to do with the choice of outcome measure. Initially we were interested in all major adverse cardiovascular events (MACE) which includes MI, stroke and cardiovascular death. However, we learned that the included studies reported these adverse events inconsistently. Some studies reported only MIs, others also reported stroke and cardiovascular death. Also, it may not be clear which patients had either MI or stroke and later cardiovascular death or whether the same patient had both MI and stroke, hence, there is a risk of counting these patients twice. Furthermore, the definition of MACE was not homogenous between the studies. Therefore, we decided to exclusively focus on MI for the rosiglitazone example.

An important and distinguishing part of the procedure is discussed in the second stage, where we zoom in on the research questions that underlie the study arms in each randomized and observational study. Selecting the proper study arms and considering the underlying research questions can be quite challenging at times. Many studies use rather complex study designs, and sometimes even propose a different design in the method section than what is actually reported in the result section. An example of this is the study by Home et al. on which we elaborated in the method section and that has been previously criticised for the adjudication of the outcome (Psaty and Furberg, 2007; Psaty and Prentice, 2010).

Another difficulty that was encountered in Stage II has to do with the fact that study arms were selected for inclusion in the analysis for lower level sub questions, which were re-used for the analysis at the higher level question as well. Although, we are aware of the fact that with this approach some rosiglitazone patients will be counted twice we consider this the best approach available at this point because of the importance of identifying the right non exposed group for each sub question. Another approach may for example be to down weigh all arms from the same study so that together the weight of these comparisons equals that of one study.

We did not find a significant association between rosiglitazone and myocardial infarction for any of our comparisons, which is possibly due to the overall lack of events. From the observational studies we learned that, when reported, the mean time to event was relatively long, over 1 year. However, most randomized studies were of much shorter duration; for many the follow-up time was between 24 - 28 weeks, and for the majority the follow-up time was between 24 and < 52 weeks (16 out of 24 RCTs, 66.7%). Based on this, we hypothesize that the randomized studies may have been too short to investigate the association between rosiglitazone use and myocardial infarction, even though we did not observe any difference in risk in the model adjusted for duration. This should be kept in mind while designing any future RCTs that test the safety or efficacy of drugs intended for long term treatment.

When events of interest are rare, which is often the case with adverse drug reactions, classical meta-analysis methods may not perform well (Cai et al., 2010; Sutton et al., 2002). Therefore, Bayesian methods have been discussed as an appropriate alternative (Sutton and Abrams, 2001). Previously, others have successfully combined information from randomized controlled trials and observational studies with Bayesian methods and various methods to achieve this have been described (Sutton and Abrams, 2001). It was estimated in one meta-analytic comparison that observational studies do not overestimate the effect size of treatment compared to randomized controlled trials (Concato et al., 2000). Hence, if we have evidence from properly carried out studies we should not be hesitant to explore innovative methods to combine these data as it will increase the underlying body of evidence and hence, the power of the analysis.

5.5 Conclusion

The procedure discussed here can be of additional value for drug safety assessment because it provides a stepwise approach that guides the decision making in order to increase transparency. With this approach results from randomized and observational studies that include treatment arms which are relevant for the research question are pooled with a Bayesian meta-analysis. Bayesian meta-analysis can be a useful tool to study drug safety because it provides a flexible way of modelling and is considered appropriate for studying rare outcomes which adverse events often are.

5.6 Appendix Chapter 5

Table A - Included studies per research question and study characteristics, with study duration in weeks (Duration) and exposure duration in weeks (Exposure).

a) in patients on rosiglitazone monotherapy compared to no treatment (includes placebo)

Year	1^{st}Author	Type of study	Comparator	Add-on	Duration	Mean age	% Male	Exposure
2001	Lebovitz	Randomized	placebo		26	60.0	65.7	26
2001	Phillips	Randomized	placebo		26	57.5	63.4	26
2005	Wang	Randomized	placebo		26	61.2	82.9	26
2007	Lipscombe	Observational	Adjusted		198	73.9	39.3	197.6
2007	Dargie	Randomized	placebo		52	64.1	81.2	52
2009	Finn	Randomized	placebo		35	62.6	76.9	33.6
2010	Bertrand	Randomized	placebo		52	64.6	91.2	52

b) in patients on rosiglitazone add-on therapy compared to the same add-on therapy alone

Year	1^{st}Author	Type of study	Comparator	Add-on	Duration	Mean age	% Male	Exposure
2000	Wolffenbuttel	Randomized	placebo	placebo	26	61.2	58.5	26
2000	Fonseca	Randomized	metformin	metformin	26	58.2	68.1	26
2001	Raskin	Randomized	insulin	insulin	26	56.8	55.6	26
2002	Gomz-Perez	Randomized	metformin	metformin	26	53.1	25.7	26
2003	Barnett	Randomized	sulphonylurea	sulphonylurea	26	54.2	77.5	26
2003	Zhu	Randomized	sulphonylurea	sulphonylurea	24	58.9	44.8	24
2004	Raskin	Randomized	other	other	24	57.6	57.6	24
2004	Kerenyi	Randomized	sulphonylurea	sulphonylurea	26	59.9	58.5	26
2005	Wong	Randomized	no treatment	insulin	24	62.3	NA	24
2005	Sarafidis	Randomized	sulphonylurea	sulphonylurea	26	62.8	45.0	26
2005	Bailey	Randomized	metformin	metformin	24	57.9	57.6	24
2007	Yilmaz	Randomized	Insulin	Insulin	26	59.8	44.1	24
2007	Davidson	Randomized	sulphonylurea	sulphonylurea	24	52.5	38.5	24
2008	Chou	Randomized	sulphonylurea	sulphonylurea	28	54.1	58.7	28
2008	Koro	Observational	Pioglitazone	Pioglitazone	110	55.5	55.1	109.2
2009	Home	Randomized	Any treatment	Any treatment	260-364	58.4	51.6	286

Table A - *(Continued)* Included studies per research question and study characteristics, with study duration in weeks (Duration) and exposure duration in weeks (Exposure).

c) in patients on rosiglitazone monotherapy vs. another monotherapy

Year	1^{st}Author	Type of study	Comparator	Add-on	Duration	Mean age	% Male	Exposure
2004	Raskin	Randomized	other		24	57.6	57.6	24
2004	Baksi	Randomized	Sulphonylurea		26	61.5	60.1	26
2005	Hanefeld	Randomized	Sulphonylurea		52	60.4	65.6	52
2006	Kahn	Randomized	Metformin		208	56.9	57.7	208
2006	Kahn	Randomized	Sulphonylurea		208	56.9	57.7	208
2007	Gerrits	Observational	Pioglitazone		65	58	58	67.6
2007	McAfee	Observational	Metformin		260	52	55	57.2
2007	McAfee	Observational	Sulphonylurea		260	52	55	57.2
2008	Chou	Randomized	Sulphonylurea		28	53.3	58.8	28
2008	Winkelmeyer	Observational	Pioglitazone		31	76.3	26.2	30.7
2009	Dore	Observational	Other		52	66	32.12	54.3
2009	Juurlink	Observational	Pioglitazone		32	73	52.7	47.1
2009	Tzoulaki	Observational	Metformin		364	66.2	50.6	369.2
2009	Rosenstock	Randomized	Other		104	54.3	56.5	104
2009	Hsiao	Observational	Sulphonylurea		48	60.7	54.1	48
2009	Ziyadeh	Observational	Pioglitazone		38	53.3	57.5	40.3
2009	Stockl	Observational	Pioglitazone		234	73	53.9	103.4
2009	Hsiao	Observational	Metformin		48	58.9	45.3	48
2010	Graham	Observational	Pioglitazone		156	74.4	59.8	15
2010	Bilik	Observational	Pioglitazone		76	58.5	45.6	76
2010	Gerstein	Randomized	Other		78	61	67.9	78.1

d) in patients on rosiglitazone add-on therapy vs. another glucose lowering agent as an add-on

Year	1^{st}Author	Type of study	Comparator	Add-on	Duration	Mean age	% Male	Exposure
2004	Derosa	Randomized	pioglitazone	other	52	53.5	49.4	50.4
2005	Derosa	Randomized	Other	Metformin	52	53,0	50.5	50.4
2005	Weissman	Randomized	Metformin titration	Metformin	24	55.6	NA	24
2007	McAfee	Observational	Sulphonylurea	Metformin	260	52,0	59	62.4
2007	McAfee	Observational	Metformin	Sulphonylurea	260	52,0	59	62.4
2007	McAfee	Observational	Other	insulin	260	52,0	59	62.4
2007	Yilmaz	Randomized	Metformin	insulin	24	57.7	43.8	24
2007	Yilmaz	Randomized	Other	insulin	24	60.1	50	24
2009	Dormuth	Observational	Sulphonylurea	Metformin	520	66.0	73.9	66.0
2009	Dormuth	Observational	Pioglitazone	Metformin	520	68.3	66.4	68.3
2009	Hsiao	Observational	pioglitazone	Metformin	48	58.9	45.3	48
2009	Hsiao	Observational	pioglitazone	Sulphonylurea	48	60.7	54.1	48
2009	Hsiao	Observational	pioglitazone	other	48	55.5	52.7	48
2009	Raskin	Randomized	pioglitazone	other	26	55.2	54.3	26

Table B - Reported adjusted effectsizes (ES) and 95% Confidence Interval (95%CI); or number of MI events and groups size (n); and assigned study weights for observational and randomized studies per research question.

a) in patients on rosiglitazone monotherapy compared to no treatment (or placebo)

Study type	1st Author	ES	95% CI	Weight
Observational	Lipscombe	1.76	[1.27; 2.44]	0.60

		Rosiglitazone		Comparator		
		Events	n	Events	n	Weight
Randomized	Lebovitz	0	335	0	158	0.80
	Phillips	0	735	0	173	0.80
	Finn	0	32	2	33	0.80
	Dargie	5	108	0	110	0.80
	Bertrand	0	98	1	95	0.80
	Wang	0	35	0	35	0.40

b) in patients on rosiglitazone add-on therapy compared to the add-on therapy

Study type	1st Author	ES	95% CI	Weight
Observational	Koro	1.03	[0.93;1.12]	0.90

		Rosiglitazone		Comparator		
		Events	n	Events	n	Weight
Randomized	Raskin	0	62	0	63	0.60
	Wolfenbuttel	0	382	0	192	0.70
	Fonseca	0	239	0	116	0.90
	Raskin	0	209	0	104	0.90
	Zhu	1	425	0	105	0.80
	Gomz-Perez	0	71	0	34	0.80
	Barnett	1	84	0	87	0.80
	Kerenyi	0	165	0	170	0.80
	Chou	0	442	0	222	0.80
	Wong	0	26	0	26	0.30
	Yilmaz	0	15	0	19	0.40
	Bailey	1	288	0	280	0.90
	Davidson	1	116	0	117	0.80
	Home	64	2220	56	2227	0.70
	Sarafidis	0	20	0	20	0.70

Table B - *(Continued)* Reported adjusted effectsizes (ES) and 95% Confidence Interval (95%CI); or number of MI events and groups size (n); and assigned study weights for observational and randomized studies per research question.

(c) in patients on rosiglitazone monotherapy vs. another monotherapy

Study type	1st Author	ES	95% CI	Weight
Observational	Hsiao	1.49	[0.99; 2.24]	0.80
	Hsiao	2.09	[1.36; 3.24]	0.80
	Bilik	0.75	[0.33; 1.67]	0.80
	Dore	1.04	[0.72; 1.51]	0.70
	Ziyadeh	1.35	[1.12; 1.62]	0.70
	Graham	1.06	[0.96; 1.18]	0.60
	Juurlink	0.95	[0.81; 1.11]	0.80
	Tzoulaki	0.79	[0.41; 1.53]	0.80
	Gerrits	0.78	[0.63; 0.96]	0.80
	Winkelmeyer	1.08	[0.93; 1.25]	0.60
	McAfee	1.19	[0.84; 1.68]	0.70
	McAfee	0.79	[0.58; 1.07]	0.70
	Stockl	0.93	[0.72; 1.21]	0.80

		Rosiglitazone		Comparator		
		Events	n	Events	n	Weight
Randomized	Raskin	0	62	0	63	0.60
	Hanefeld	0	384	0	203	0.80
	Baksi	0	225	0	241	0.70
	Kahn	25	1456	21	1454	1.00
	Kahn	25	1456	15	1441	1.00
	Chou	0	230	0	222	0.80
	Rosenstock	0	202	0	396	0.80
	Gerstein	8	333	7	339	0.70

d) in patients on rosiglitazone add-on therapy vs. another glucose lowering agent as an add-on

Study type	1st Author	ES	95% CI	Weight
Observational	Hsiao	0.69	[0.3; 1.55]	0.80
	Hsiao	6.34	[1.8; 22.31]	0.80
	Hsiao	1.04	[0.73; 1.47]	0.80
	McAfee	0.41	[0.16; 1.04]	0.70
	McAfee	1.45	[0.76; 2.75]	0.70
	McAfee	0.79	[0.46; 1.36]	0.70
	Dormuth	0.90	[0.69; 1.17]	0.80
	Dormuth	1.00	[0.67; 1.49]	0.80

		Rosiglitazone		Comparator		
		Events	n	Events	n	Weight
Randomized	Derosa	0	42	0	47	0.90
	Derosa	0	48	0	47	0.90
	Weissman	2	358	0	351	0.90
	Yilmaz	0	15	0	17	0.40
	Yilmaz	0	15	0	15	0.40
	Raskin	0	187	0	187	0.50

6

Reporting of Bayesian methods in epidemiological and medical research: a systematic review

Summary. Despite the increasing acknowledgement of Bayesian data analysis several reviews suggest the underuse of these techniques in epidemiological and medical research. The objective of this systematic review is to investigate the time trend and current status of Bayesian statistics within these research areas. We focus on the types of models and computational methods used for analyses and asses the quality of reporting.

Complete volumes of 6 major epidemiological journals and 6 major medical journals in the period 2005-2013 were searched via Pubmed. In addition we performed an extensive within-manuscript search using a specialized Java-application. Details of reporting on Bayesian statistics were examined in original research papers with primary Bayesian data-analyses.

An upward trend in the number of publications referring to Bayesian statistics is revealed for the medical journals. For the epidemiological journals the number of studies in which Bayesian analyses were used remain constant over the years (except for a dip in 2011). Though many authors presented thorough descriptions of the analyses they performed and the results they obtained, several reports presented incomplete method sections, and even some incomplete results sections. Especially, information on the process of prior elicitation, specification and evaluation was often lacking.

Though available guidance papers concerned with reporting of Bayesian analyses emphasize the importance of transparent prior specification, the results obtained in this systematic review show that these guidance papers are often not used. Additional efforts should be made to increase the awareness of the existence and importance of these checklists in order to overcome the controversy with respect to the use of Bayesian techniques. The reporting quality in epidemiological and medical literature could be improved by updating existing guidelines on the reporting of frequentist analyses to address issues that are important for Bayesian data analyses.

This chapter is based on: Rietbergen, C., Debray, T.P.A., Klugkist, I., Janssen, K.J.M., & Moons, K.G.M. Reporting of Bayesian methods in epidemiological and medical research: a systematic review. (Under revision for *Journal of Clinical Epidemiology*).

6.1 Background

Over the past few decades an extensive body of literature has been published describing the rationale and (potential) advantages of Bayesian data analysis techniques within epidemiological and medical research (see for example Berry and Stangl (1996); Spiegelhalter et al. (1999); Goodman (1999a,b); Spiegelhalter et al. (2004); Greenland (2006, 2007)). These articles discuss the advantages and flexibility of Bayesian approaches in the process of, for example, prediction model development, interim analysis, and sample size calculation.

Despite the attention Bayesian techniques receive in methodological literature, at the beginning of the millennium the use of Bayesian methods in applied research seemed limited. This conclusion followed from an non-systematic search in the medical literature by Altman (2000) and from a systematic review by Spiegelhalter et al. (2000) which focussed on statistical methods in health technology assessment. The persistency of the underuse of Bayesian methods in current research was reported by Pibouleau and Chevret (2011) in a review on the evaluation of the effectiveness of implantable medical devices.

Given the increasing acknowledgement of Bayesian statistics we question whether the conclusion of the underuse of these techniques is still justified. Therefore, with the current study we aim to update the series of reviews of Bayesian techniques that were done in specific research areas, with an extensive systematic review on the use of Bayesian techniques in epidemiological and medical research in general in the period 2005-2013. The objective of this review is to examine the trend over time and current status of the use of Bayesian data analyses techniques in epidemiological and medical research. We aim to map the degree of Bayesian statistics used in primary data analysis, the type of statistical models and computational methods used and to identify the quality and transparency of reporting corresponding study results.

6.2 Methods

6.2.1 Search strategy

The search for studies reporting Bayesian data analysis focussed on original research papers published in the top-6 epidemiological and the top-6 medical journals (ISI Web of Knowledge, 2010) as displayed in Table 6.1. Of the epidemiological journals we systematically searched issues that appeared in 2005, 2007, 2009, 2011 and 2013. Due to the frequent appearance of the medical journals we searched only those issues that appeared in 2005, 2009 and 2013.

To select only original research reports and exclude publications such as editorials, letters, and commentaries, we made use of the PubMed Publication Characteristics (Publication Types). Eligible publication types are clinical trials (phase I-IV), journal articles, multicenter studies, randomized controlled trials, comparative studies, technical reports, controlled clinical trials, twin studies, evaluation studies, and validation studies.

Table 6.1. Epidemiological and Medical Journal Rankings on 5-year Impact Factor According to ISI Web of Knowledge

Top 6 Epidemiological	Abbr.	Top 6 Medical	Abbr.
Epidemiologic Reviews	ER	New England J. of Medicine	NEJM
American J. of Epidemiology	AJE	J. of the American Medical Association	JAMA
International J. of Epidemiology	IJE	The Lancet	LANCET
Epidemiology	EPI	Annals of Internal Medicine	ANNALS
J. of Epidemiology & Community Health	JECH	Public Library of Science Medicine	PLOS MED
J. of Clinical Epidemiology	JCE	British Medical J.	BMJ

Identification of eligible papers published within the selected journals and journal types, followed the two search paths as displayed in Figure 6.1. The left-hand side of the flow diagram shows the identification of epidemiological or medical studies using PubMed with the search terms [Bayes* OR MCMC OR "credible interval"] in combination with the name of each of the epidemiological and medical journals separately and the period of interest (e.g., [(Bayes* OR MCMC OR credible interval) AND "Lancet"[Journal] AND "2005/01/01"[Entrez Date] : "2005/12/31"[Entrez Date]]).

To catch also those research papers not reporting the use of Bayesian methods in titles, abstracts and keywords, a full-text within-manuscript search was performed. The course of this part of the search is displayed on the righthand side of Figure 6.1. A specialized Java-application that enabled automated downloading and indexing of manuscripts was written for this purpose Debray (2014). This application employs Entrez Programming Utilities to perform search queries and retrieve article information (such as title, abstract, authors, keywords) from the Pubmed library. The article information is identical to the information provided by the Pubmed website, and can be used for retrieving its full-text (original PDF files), provided that the user is subscribed to the corresponding journal. The application was extended with a module to allow searching within the full-text of the retrieved articles. This module is based on Apache Lucene, a text search engine library written in Java.

We used the specialized application to retrieve all research papers from the journals and time periods under consideration (e.g., ["Lancet"[Journal] AND "2005/01/01"[Entrez Date] : "2005/12/31"[Entrez Date]]). Afterwards, we employed the text search module for searching full-text publications of all research papers on the occurrence of "Bayes*" in the title, abstract, keywords and full-text of the retrieved articles. This strategy ensures that at least all results from the Pubmed website search are reproduced, as the article information provided by Pubmed as well as the corresponding content are analyzed.

To compare our results with the total number of research papers published by each journal we estimated this number via Pubmed. We searched the database using the search strategy as presented above, while omitting the search terms [Bayes* OR MCMC OR "credible interval"] from the query.

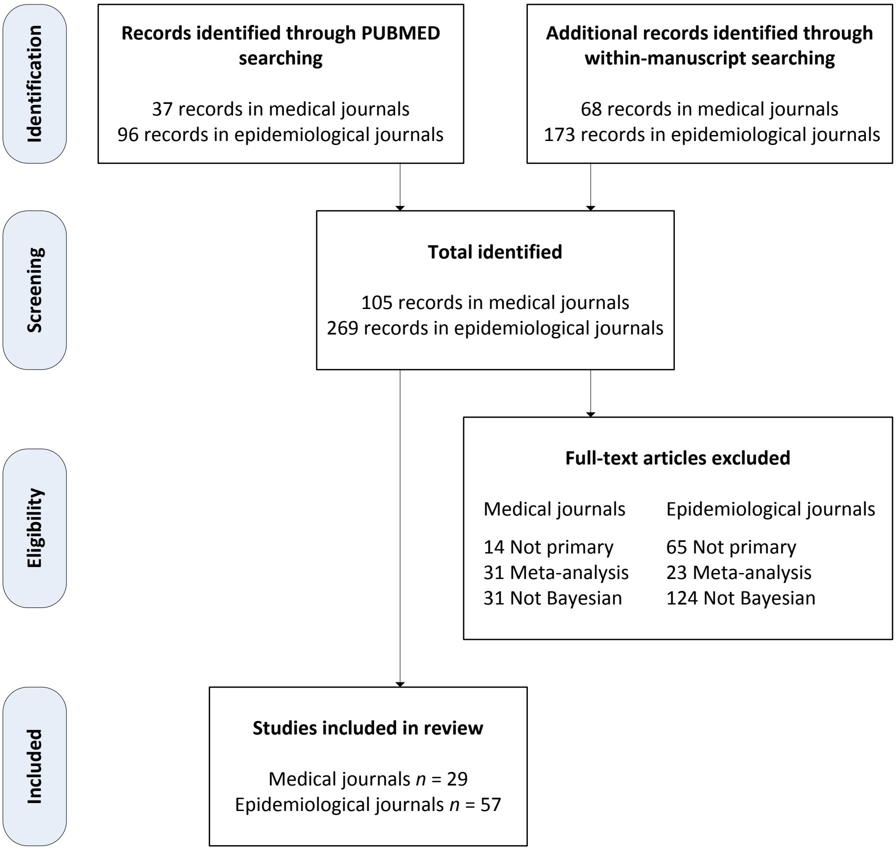
Identification
Screening
Eligibility
Included
Records identified through PUBMED searching
37 records in medical journals
96 records in epidemiological journals
Additional records identified through within-manuscript searching
68 records in medical journals
173 records in epidemiological journals
Total identified
105 records in medical journals
269 records in epidemiological journals
Full-text articles excluded
Medical journals
14 Not primary
31 Meta-analysis
31 Not Bayesian
Epidemiological journals
65 Not primary
23 Meta-analysis
124 Not Bayesian
Studies included in review
Medical journals n = 29
Epidemiological journals n = 57

Fig. 6.1. Flow diagram of systematic review results

6.2.1.1 Inclusion and Exclusion criteria

From the reports that were identified with either search strategy, papers that only mentioned Bayesian statistics as an alternative to the classical approach, or that only reported the Bayesian Information Criterion were excluded and classified as not Bayesian. Furthermore, Bayesian meta-analyses fell outside the scope of this review. Finally, a category of studies that did not report on primary Bayesian analyses were excluded. These studies did not report the use of Bayesian statistics for the main part of the primary data analysis, e.g. studies in which Bayesian statistics were used to develop a scale to measure one of the predictor variables, but in which Bayesian statistics were not used to assess the association between the determinants and the main outcome. In addition, papers reporting empirical Bayesian analyses were excluded and classified as being not a primary Bayesian analysis. The numbers of excluded reports are listed in Figure 6.1.

6.2.2 Methodological assessment

From the primary Bayesian analyses additional information on the design and analysis of the study was obtained using a standardized item-list. This item-list was based on the guidelines of the BayesWatch as introduced by Spiegelhalter et al. (2000), the BaSiS guidelines for reporting Bayesian Analysis (The BaSiS Group, 2001), and the ROBUST criteria as specified by Sung et al. (2005). The list was designed to collect information on the objective of the study, the design of data collection, the statistical model used, how prior distributions were specified, whether sensitivity analyses were part of the prior specification procedure, the statistical package used for analysis, and how the results were reported.

6.3 Results

6.3.1 Bayesian statistics in medical and epidemiological journals

Medical journals. The flow diagram in Figure 6.1 shows result of our search and screening for both medical and epidemiological journal papers. We identified a total of 37 medical research articles reporting the term [(Bayes* OR MCMC OR credible interval)] in titles, abstracts or keywords when using a direct search in PubMed. Via the within-manuscript search an additional 68 studies were found. After screening and eligibility assessment of all 105 hits, we excluded 76 studies that were either not primary Bayesian analyses, or were Bayesian meta-analyses, or that were not Bayesian at all, for example when Bayes was only mentioned in reference lists. We were able to include 29 medical journal articles as reporting on primary Bayesian analysis. In Table 6.2 we present our search results separately per journal. In the first column of the table we report the estimate of the total number of research papers published by these journals. Although we emphasize that these numbers are only a rough estimate, since we could not further screen these results, they do provide some insight in the amount of Bayesian primary analyses as a fraction of the total number of analyses reported in the top epidemiological and medical journals.

Figure 6.2 visualizes the development of the use of Bayesian statistics within the 6 medical top journals printed in dark shades over the period 2005-2013 in the order as presented in the left column of Table 6.1. The bars for the medical journals show a clear upward trend for the use of Bayesian statistics in clinical studies. Annals of internal medicine was excluded from this figure, since no primary Bayesian analyses were identified within this journal.

Epidemiological journals. Using PubMed we found a total of 96 epidemiological research articles reporting one of our search terms. An additional 173 hits were found through the within-manuscript search. After screening the 269 articles we excluded

212 papers for the reasons mentioned in Figure 6.1. We included 57 epidemiological studies that used Bayesian statistics for primary data analysis. Search results per journal separately can be found in Table 6.2. The development of the use of Bayesian statistics within the epidemiological journals over the period 2005-2013 is presented in Figure 6.2. For the epidemiological journals the number of Bayesian analyses per year remained constant, except for a dip in 2011. The results for Epidemiologic Reviews were excluded from the figure since no primary Bayesian analyses could be found within this journal.

Table 6.2. Total number of 5-Year Epidemiological and 3-Year Medical Research Papers; Number of Hits found with Pubmed (Hits); Number of Primary Bayesian Analyses Identified with Pubmed (Primary); Number of Hits Found with Within-Manuscript Search (Hits) and Additional Primary Analyses (Additional) found with Within-Manuscript search; Total Number of Hits.

Epidemiological	5-Year	PubMed search		Within-Manuscript		
		Hits	Primary	Hits	Additional	Total
ER	81	0	0	3	0	0
AJE	1684	35	14	66	12	26
IJE	1040	12	2	48	2	4
EPI	617	23	9	53	9	18
JECH	992	3	2	18	6	8
JCE	923	23	1	81	0	1
Medical	3-Year					
NEJM	2132	3	2	15	2	4
JAMA	1651	4	1	15	1	2
LANCET	2540	10	10	30	6	16
ANNALS	1030	5	0	14	0	0
PLOS MED	643	7	4	17	1	5
BMJ	2359	8	1	14	1	2

6.3.1.1 Description of the included studies: objective, design of data collection and statistical model

Results obtained using the itemlist on the reporting of Bayesian analyses are displayed in Table A in the Appendix. The first three items provide a general description on the type of studies published using Bayesian data analysis. The majority of the epidemiological studies had an etiologic research objective (35/57), against only 4 out of 29 medical studies. Most epidemiological studies made use of observational data obtained from registers, census and surveillance data (19/59) or from prospective, retrospective or cross-sectional cohorts (16/57), where medical studies used mainly

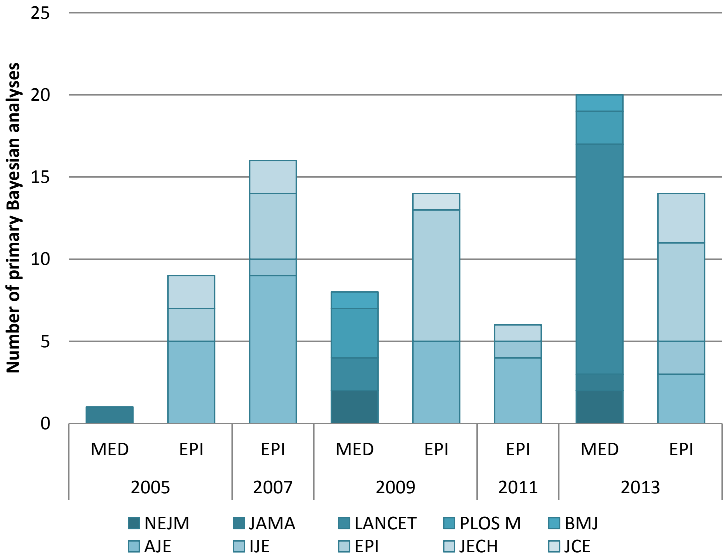

Fig. 6.2. Number of primary Bayesian analysis for medical and epidemiological in publications between 2005-2013.

existing registries (17/29). In total only 10 studies were randomized clinical trials (RCT).

Various statistical models were reported in the research papers of which a larger part (34/57 and 10/29 respectively) can be classified as a (hierarchical) regression model. This category comprises various types of (hierarchical) regression models with different distributions assumed for the outcome variables (Poisson, normal, multinomial, etc.), and different link functions used (logit, log etc.).

6.3.1.2 Prior and sensitivity analysis

A total of 28 out of 57 epidemiological studies and 15 out of 29 medical studies did not report any information about the specification of the priors used in the analyses. About twice as many uninformative than informative priors were used by both the epidemiological and medical researchers.

Of the 29 epidemiological studies that reported information on the type of priors used, only 4 reported the analysis of the sensitivity of the posterior distribution for changes in the prior distribution. The same holds for 3 of the 14 medical studies that reported on the type of prior used.

6.3.1.3 Computations and software

MCMC sampling was the common estimation type used for Bayesian analysis in both epidemiological and medical studies. In two epidemiological studies nested Laplace approximation was used. Only in one (medical) study, parameter estimates were obtained analytically. In 9/57 epidemiological and 10/29 medical articles the computation approach was not mentioned.

Four types of software used for Bayesian analysis were distinguished. The first category concerns packages especially designed for MCMC simulation like OpenBUGS (or WinBUGS) and JAGS. These packages were used in 24 of the 57 epidemiological studies and in 3 of the medical studies. The second category concerns specialized packages (e.g., PHASE, BEAST, MrBayes, Genotyping Console) used for analyses in haplotyping, genotyping and phylogenetical studies. These packages were used in 8 of epidemiological and 5 of the medical studies. The third category concerns more general (statistical) languages and packages that can be used for Bayesian analysis (e.g., R, SAS and Fortran) and that were sometimes used in combination with software from the first category. These were used in 10 of epidemiological studies and 7 of the medical studies. The fourth category is a rest category for (stand-alone) packages that did not fit any of the other categories and in which only a total of 3 studies were classified. In addition, about a fourth of the epidemiological studies and about half of the medical studies did not provide any information on the software package used for analyses.

6.3.1.4 Reporting of results

Usually the epidemiological studies reported a combination of posterior estimates with a type of 95% posterior interval (mostly a Central Credibility Interval). In only about a third of these cases the authors clarified whether a posterior mean, mode or median was reported. The same trends in reporting were found in the medical literature where only in about a fourth of the reports the type of posterior estimate was mentioned. In addition to the posterior estimates and intervals 10 of the epidemiological studies reported the use of the DIC for model selection, while this criterion was not reported in any of the medical papers.

Bayes Factors were only used in 6 of the included studies. The posterior predictive distribution was only used in one medical study that reported a posterior predictive estimate. And 3 studies that reported Bayesian p-values. Strangely, one epidemiological and one medical paper reported no Bayesian results in the end.

6.4 Discussion

Over the past few decades, the use of Bayesian statistics has increasingly been suggested as a promising alternative for frequentist data analysis. Detailed information on their actual implementation in the epidemiological and medical literature is, however, lacking. We therefore investigated the current position of Bayesian statistics in epidemiological and medical research by means of a systematic review. In this review, we inventoried the type of statistical models and computational methods used in primary Bayesian data analysis and evaluated the transparency of their reporting.

We identified a modest upward trend in the number of publications adopting a Bayesian analysis in the medical literature. For the epidemiological papers the number of primary Bayesian analysis remained constant over the past ten years, despite a dip in 2011. Despite of these trends, the actual number of studies using Bayesian techniques for primary data-analysis remains rather low. This is consistent with the findings reported in the reviews by Altman (2000), Spiegelhalter et al. (2000) and Pibouleau and Chevret (2011).

In addition to earlier studies, we found that the quality of reporting on Bayesian analyses was rather low. Though many authors presented thorough descriptions of the analyses they performed, we encountered several reports that presented incomplete method sections, and even some incomplete results sections. In particular, about half of the epidemiological and medical studies failed to report any information on the type of prior distributions used for the analyses. This information is crucial to evaluate their influence on the final results, and to interpret differences that may exist when analyzing the data using frequentist approaches. In addition to this, the majority of the studies did not inform the reader whether a sensitivity analysis with respect to the prior distribution was performed. It remains unclear whether sensitivity analyses were not performed, or performed but unreported. This information is not only valuable to the primary investigator, but also informs the reader on the stability and reliability of the results. It comes to no surprise that the importance of sensitivity analyses is often emphasized in existing checklists on Bayesian data analyses (see Spiegelhalter et al. (2000); The BaSiS Group (2001); Sung et al. (2005)).

Transparent reporting has become increasingly important over the past few years, and several guidance papers have recently been published. Our review indicates that adherence to these guidelines seems particularly urgent for Bayesian data analyses, as controversy around their implementation often relates to the choice of prior distributions. Transparent reporting on prior specification could help to partly overcome this issue. As mentioned, several checklists for reporting on Bayesian analyses are widely available, and they all include items on the reporting of prior specification and sensitivity analysis. Additional efforts should be made to increase the awareness of the existence and importance of these checklists among researchers. Unfortunately, there are no uniform recommendations about the implementation of available guidance papers. The reporting quality of papers in the epidemiological and medical literature could therefore be improved more pragmatically by updating existing guidelines (see for example Schulz et al. (2010); von Elm et al. (2007)) to specifically address issues that are relevant for frequentist and Bayesian data analyses.

6.5 Conclusion

Despite the increasing acknowledgement of Bayesian statistics, this review confirmed the underuse of these methods in contemporary clinical and epidemiological research. However, the use of Bayesian methods is not as limited as presumed by some authors, and an increasing trend in the number of papers referring to Bayesian statistics was found. Unfortunately, a substantive part of the studies that used Bayesian methods, provide little information on the process of data analysis, such as prior specification and sensitivity analyses. The low quality of reporting is likely to feed to controversy and skepticism around the use of Bayesian techniques, as overt prior specification is essential for proper interpretation of research results. Though many checklists for reporting on Bayesian analysis are available, this review revealed that researchers seldom adhere to these guidance papers. We think the use of checklists should be encouraged, and may ultimately improve the reporting on Bayesian methods and the reproducibility of research results.

6.6 Acknowledgement

We would like to thank Fayette Klaassen for her assistance in searching the literature and coding the included research papers.

6.7 Appendix Chapter 6

Table A - Reporting Bayesian Analyses in Epidemiological (E) and Medical (M) Research Papers.

Item	Category	Epi	Med
Objective	Etiologic	35	4
	Prognostic	11	1
	Intervention	0	6
	Diagnostic	3	1
	Prevalence	1	7
	Disease mapping / transmission	6	9
	Economic evaluation	1	1
Design of data collection*	Existing registry [a]	19	17
	(Longitudinal / pro/-retrospective) cohort	16	3
	Cross-sectional	7	1
	(Multicenter) RCT	3	7
	Quasi-experiment	1	0
	(Nested) case-control	6	0
	Survey	5	1
	Clinical sample	0	1
Statistical model*	(Hierarchical) regression model [b]	34	10
	Survival model	6	3
	Geostatistical / spatial model	1	3
	Time-series [c]	3	0
	Haplotyping / genotyping / phylogenetics	8	6
	Mathematical transition models	2	2
	Proportions / percentages	0	4
	Bayesian model averaging	0	1
	Latent class analysses	1	0
	Not reported	2	1
Type of prior*	Uninformative	22	12
	Informative	10	6
	Data driven	4	1
	Not reported	28	15
Sensitivity analysis on prior	Yes	4	3
	Not reported	53	26
Computations	Markov Chain Monte Carlo method	46	18
	Nested Laplace approximation	2	0
	Analytic	0	1
	Not reported	9	10
Software *	WinBUGS / JAGS	24	3
	PHASE, BEAST, MrBayes, Genotyping console	8	5
	R / SAS / Python / Fortran / S-plus	10	7
	bic.glm / MLWin / TLNise	2	1
	Not reported	14	15
Reporting*	Central credibility interval (CI)	43	20
	Highest posterior density (HPD)	0	2
	Posterior estimate (unspecified)	37	19
	Posterior mean, mode, median (specified)	16	6
	Posterior probabilities	3	5
	Posterior predictive estimates	0	1
	Deviance Information Criterion	10	0
	Bayes Factors	4	2
	Posterior SD/SE, MCerror, Inter-quartile-range	2	2
	Bayesian p-values	1	2
	Not reported	1	1

* More categories might be applicable; column totals do not add up to total number of articles. [a] Census, register, surveillance. [b] Normal, logistic, Poisson, multinomial, etc. [c] Distributed lag, discrete time etc.

7

General Discussion

7.1 The power prior distribution with random study weights

Quantitative synthesis of evidence is in line with our natural tendency to consider new information in the light of that what we already know. The introduction of this thesis illustrates how researchers tend to interpret research results while accounting for the results obtained in previous studies in an informal and qualitative way. The power prior distribution as proposed by Ibrahim and Chen (2000) offers a quantitative and flexible approach to formally incorporate historical research results from different sources in the analysis of new data. Two procedures for the specification of this informative prior distribution were discussed and evaluated in this thesis. One of them is the joint power prior distribution in which the weight parameter and the parameter of interest are jointly estimated from the historical and new data. With this approach the size of the weight parameter, that is the weight the historical evidence receives in the analysis, depends on the degree to which the historical and new data resemble each other. This means that when the historical data is in agreement with the new data, the historical study would receive a higher weight than when the historical and current data differ a lot. This thesis discussed whether this property of the joint power prior distribution is desirable since differences and similarities between the datasets can be the result of sampling variability. A numerical example demonstrated how two samples coming from the same population can result in very different estimates of a parameter, leading to an underestimation of the size of the weight parameter. It was concluded that the size of the weight parameter should depend on the similarities and differences of the study characteristics rather than on the study results.

These conclusions imply that not the data should determine the weight of the historical studies, but the researchers themselves should judge the similarities between the current and historical studies in order to specify a power prior distribution. This idea is shared with other researchers that evaluated the power prior distribution. For example, Neuenschwander et al. (2009) state that they would not recommend taking the weight parameter as an unknown parameter. In addition, Neelon and O'Malley (2009) describe how the power prior tends to heavily discount the historical data even in case of minor discrepancies between the new and historical data sets. They discuss an alternative strategy of assigning fixed weights, which provides the user with more control over the impact of the historical data. They mention the need for external expert knowledge as the downside of this approach and stress the importance of sensitivity analyses. Available literature, however, does not elaborate on the approaches for expert elicitation of these fixed weights, possibly since including expert knowledge introduces additional and often unwanted subjectivity into the analysis process.

7.2 Elicitation of fixed study weights

Using external expert knowledge means that the expert has to weigh the quality and relevance of the available knowledge. Implicitly, this is something that researchers are already used to do. In the formal quantitative synthesis of scientific research results one is faced with the same considerations, but now it is demanded to be explicit in the choices made and the arguments used to support these decisions. Each time one has to corroborate what historical data to include or exclude, in which way and to what extent. The subjectivity associated with these considerations is inevitable. That is, the common approach of focusing only on the current data in one's analyses and ignoring previous results, in other words using a weight equal to zero for the historical studies, is a subjective choice too.

A sensitivity analysis on the data from gynaecological clinical trials used to illustrate the second chapter showed that simply ordering the historical studies based on relevance and quality, and assigning weights accordingly, already resulted in stable posterior estimates for the effect of intrapartum fetal monitoring using ECG, irrespective of the actual size of the weights that were used. The third chapter elaborated on this idea. To deal with the subjectivity associated with the specification procedure, this time a panel of experts was asked to rank and assign weights to a number of historical studies. Their motivation was monitored and anonymously shared among the other panel members using a Delphi technique. This way the experts were encouraged to convince each other and together come to a final judgment on the set of historical studies. This research showed that in this particular study only few rounds were needed for the experts to reach agreement on the ranking of the studies and convergence with respect to the actual study weights. These results together with the positive results obtained in the analysis in the second chapter make this approach an attractive procedure for study weight elicitation.

Despite the fact that the case studies as presented in Chapter 2 and 3 are rather specific, some general recommendations and conclusions can be derived from the results. The weights obtained with the Delphi procedure can be used directly to specify the weight parameters for the historical studies in a conditional power prior approach. The between-rater variation in weights can serve as guide values for sensitivity analyses. For example, it can be assessed whether the study weights as established by one expert lead to different posterior results than would have been obtained when using the weights from another expert. In case the researcher is reluctant to use the conditional power prior, due to the more subjective nature of this procedure, another option is to use the information obtained from the expert panel as input for the specification of a hyperprior for the weight parameter in the joint power prior distribution. Hobbs et al. (2011) elaborate on the specification of this hyperprior but let the specification of this prior fully depend on the commensurability of the historical and current data. Using external expert knowledge about the commensurability of the historical and current study characteristics to guide the specification of this prior is not evaluated in the power prior current literature yet. However, further examination of this topic might be worthwhile since this way the researcher can benefit from the use of the joint power prior distribution that does not ask for the selection of fixed study weights, while at the same time valuable expert knowledge can be used to elicit probable values for these weights to make it easier to select a sensible hyperprior.

An important issue addressed in the elicitation study concerns the burden on the expert panel in case large numbers of historical studies are available for the specification of the power prior distribution. A Delphi approach as proposed in the third chapter would possibly not be feasible due to increased panel workload. Furthermore, with more studies it will be more difficult to reach convergence among the experts opinions. Future research could focus on the development of a more efficient procedure to elicit larger numbers of weights (or rankings) to enable the specification of the conditional power prior distribution in this situation. For example, from the second chapter we learned that for some cases the ranking of the historical studies is more important than the actual weights assigned. In addition, in the third chapter it was found that reaching agreement among the expert with respect to the ranking of the historical studies is much easier than reaching agreement with respect to the actual weights. Combining these conclusions suggests that for some cases eliciting only a ranking from the experts might be sufficient. The efficiency of the elicitation process could be improved by implementing an intermediate evaluative phase after obtaining the ranking of the historical studies. Assessing the posterior sensitivity for the actual weights given this ranking, might lead to the conclusion that the posterior is robust for changes in the actual weights given the ranking. Continuing the elicitation process to establish these weights would in that case needlessly increase the expert burden. Implementing this intermediate evaluation is likely to increase the efficiency of the elicitation process, especially in the case of multiple historical studies. Efforts should be made to decide on proper stopping rules and for cases where stopping is not allowed more efficient procedures have to be designed to elicit weights for larger numbers of historical studies at the same time.

7.3 Power priors for multiple historical studies

Not only is the specification of the conditional power prior problematic in the presence of multiple historical studies, also for the joint power prior distribution major difficulties remain existent. That is, although Ibrahim and Chen (2000) propose a joint power prior distribution with individual study weights for multiple historical studies, the practical application of this procedure is limited to the case of two historical data sets. With more than two historical studies present the prior is not suitable since the weights are unidentified. More research on this topic is needed in order to find a solution for this problem. One way to deal with this problem is to pool all available prior studies into one prior distribution, and to assign a weight to the likelihood of the total prior evidence as discussed in Welton et al. (2012). In the fifth chapter this approach was adopted for the conditional power prior in a situation in which the goal was to synthesize evidences obtained in studies with different study designs, that is observational and randomized studies. In this chapter the observational evidence was used as input for prior specification for the analysis of the data obtained in the randomized trials. Per research question the weights for the power priors were selected as such that the prior precision was similar to the precision estimated in the current data. In addition, so called reference analyses were conducted with weights set to 0 and 1 as proposed by Neelon and O'Malley (2009) to assess the sensitivity of the posterior distribution for the estimated parameters with different prior specifications. These sensitivity analyses showed that for this case the influence of the prior can be substantial, especially when there is limited current evidence. With that the necessity of sensitivity analyses to evaluate the influence of the prior distribution on the posterior estimates is once again emphasized.

7.4 Sensitivity analyses and guidelines on reporting Bayesian data analyses

One important conclusion drawn from the systematic review on the use of and reporting on Bayesian analyses is that sensitivity analyses regarding the posterior robustness for the choice of prior are only rarely reported. This finding is remarkable since their use is promoted in every guideline on reporting of Bayesian data analysis published in the literature. In addition, the review revealed that the use of these guidelines in common research practice is in general very limited, and the quality of reporting rather low. Especially, when it comes to prior specification research reports are often unclear about the priors used in the analyses, while this information is crucial in order to evaluate the influence of the prior on the posterior results. In addition, the researchers that mentioned the use of informative priors provided very little insight in what information was used for prior specification and how they decided on the inclusion and exclusion of information for this purpose.

As discussed in the systematic review on Bayesian reporting recently the transparency of reporting in research has received increasing attention and several guidance papers on this topic have been published (see for example von Elm et al. (2007); Schulz et al. (2010). Adherence to these guidelines seems especially urgent for Bayesian data analyses since the controversy around these techniques often refers to the prior specification. Evidence synthesis using Bayesian techniques could be improved by inclusion of the available guidelines on Bayesian data-analysis and reporting within the existing guidelines. This way the development of methods for including information from studies with a different design and the consultation of experts to assess the quality and relevance of the available information provide opportunities for more complete quantitative evidence syntheses in the future.

7.5 Concluding remarks

The majority of the illustrative examples used throughout this thesis are examples of evidence synthesis in medical research. Within this field of research the need for synthesis of evidence has increased enormously during the last decades due to the requirements for evidence based practice; the integration of the best available research with clinical expertise. For social sciences similar developments are currently taking place, since the growing amount of research results concerning specific hypotheses and theories demands for sophisticated approaches of evidence synthesis and meta-analyses. The enormous heterogeneity between studies with respect to outcome definition and measurement, interventions, study design and study populations make sensible synthesis of evidence in social sciences extra challenging. These difficulties on the one hand complicate the procedures for power prior specification. For example, as discussed in the third chapter, it is expected that elicitation procedures will be more time consuming in case there is more panel disagreement on the relevance of the studies at the start. At the same time, these difficulties emphasize the urgency for the use of quantitative evidence synthesis techniques. That is, in the light of the available but complex historical research results, focusing only on that what is found in a single study becomes less valuable. Moreover, pooling these results irrespective of the comparability of their study characteristics becomes less meaningful. Just like in medical sciences the subjective assessment of relevance and quality is inevitable, making the results obtained in this thesis applicable to social sciences as well. The synthesis of evidence and specifically the application of power priors can benefit from the efforts already made in medical research.

References

Altman, D. G. (2000). Statistics in medical journals: some recent trends. *Statistics in Medicine*, 19:3275–3289.

Amer-Wåhlin, I., Hellsten, C., Norn, H., Hagberg, H., Herbst, A., Kjellmer, I., Lilja, H., Lindoff, C., Månsson, M., Mårtensson, L., Olofsson, P., Sundström, A.-K., and Marsál, K. (2001). Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. *The Lancet*, 358:534–538.

Berry, D. A. and Stangl, D. K. (1996). *Bayesian Biostatistics.* Marcel Dekker, New York.

Bertrand, O. F., Belisle, P., Joyal, D., Costerousse, O., Rao, S. V., Jolly, S. S., Meerkin, D., and Joseph, L. (2012). Comparison of transradial and femoral approaches for percutaneous coronary interventions: a systematic review and hierarchical bayesian meta-analysis. *American Heart Journal*, 163(4):632–648.

Bown, M. J. and Sutton, A. J. (2010). Quality control in systematic reviews and meta-analyses. *European journal of vascular and endovascular surgery : the official journal of the European Society for Vascular Surgery*, 40(5):669–677.

Cai, T., Parast, L., and Ryan, L. (2010). Meta-analysis for rare events. *Statistics in medicine*, 29(20):2078–2089.

Concato, J., Shah, N., and Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England journal of medicine*, 342(25):1887–1892.

Coplan, P. M., Noel, R. A., Levitan, B. S., Ferguson, J., and Mussen, F. (2011). Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clinical pharmacology and therapeutics*, 89(2):312–315.

Cumming, G., Williams, J., and Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, 3:299–311.

Dargie, H. J., Hildebrandt, P. R., Riegger, G. A., McMurray, J. J., McMorn, S. O., Roberts, J. N., Zambanini, A., and Wilding, J. P. (2007). A randomized, placebo-controlled trial assessing the effects of rosiglitazone on echocardiographic function and cardiac status in type 2 diabetic patients with new york heart association functional class i or ii heart failure. *Journal of the American College of Cardiology*, 49(16):1696–1704.

Debray, T. (2014). Reference vortex. Available from: `http://www.netstorm.be/downloads/`.

Delbecq, A. L., Van de Ven, A. H., and Gustafson, D. H. (1975). *Group Techniques for Program Planning: A Guide to Nominal and Delphi Processes*. Scott, Foresman and Co., Glenview, IL.

Diamond, G. A., Bax, L., and Kaul, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Annals of Internal Medicine*, 147(8):578–581.

Dorussen, H., Lenz, H., and Blavoukos, S. (2005). Assessing the reliability and validity of expert interviews. *European Union Politics*, 6(3):315–337.

Duan, Y. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17:95–106.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106.

EMA (2010). Press release: European medicines agency recommends suspension of avandia, avandamet and avaglim. antidiabetes medication to be taken off the market. Available from: `http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2010/09/news_detail_001119.jsp&mid=WC0b01ac058004d5c1`.

FDA and Administration, U. (2011). Fda drug safety communication: updated risk evaluation and mitigation strategy (rems) to restrict access torosiglitazone-containing medicines including avandia, avandamet and avanda. Available from: `http://www.fda.gov/Drugs/DrugSafety/ucm255005.htm`.

Gerstein, H. C., Ratner, R. E., Cannon, C. P., Serruys, P. W., Garcia-Garcia, H. M., van Es, G.-A., Kolatkar, N. S., Kravitz, B. G., Miller, D. M., Huang, C., Fitzgerald, P. J., Nesto, R. W., and the APPROACH Study Group (2010). Effect of rosiglitazone on progression of coronary atherosclerosis in patients with type 2 diabetes mellitus and coronary artery disease: The assessment on the prevention of progression by rosiglitazone on atherosclerosis in diabetes patients with cardiovascular history trial. *Circulation*, 121(10):1176–1187.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine*, 130(12):995.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The bayes factor. *Annals of Internal Medicine*, 130(12):1005–1013.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, 35(3):765–775.

Greenland, S. (2007). Bayesian perspectives for epidemiological research: II. Regression analysis. *International Journal of Epidemiology*, 36(1):195–202.

Groenwold, R. and Rovers, M. (2010). The catch-22 of appraisals on the quality of observational studies. *Journal of Clinical Epidemiology*, 63(10):1059 – 1060.

Guo, J. J., Pandey, S., Doyle, J., Bian, B., Lis, Y., and Raisch, D. W. (2010). A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ispor risk-benefit management working group. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 13(5):657–666.

Hedblad, B., Zambanini, A., Nilsson, P., Janzon, L., and Berglund, G. (2007). Rosiglitazone and carotid IMT progression rate in a mixed cohort of patients with type 2 diabetes and the insulin resistance syndrome: main results from the Rosiglitazone Atherosclerosis Study. *Journal of Internal Medicine*, 261(3):293–305.

Herbison, P., Hay-Smith, J., and Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of clinical epidemiology*, 59(12):1249–1256.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., and Sterne, J. A. C. (2011). The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*.

Home, P. D., Pocock, S. J., Beck-Nielsen, H., Curtis, P. S., Gomis, R., Hanefeld, M., Jones, N. P., Komajda, M., and McMurray, J. J. (2009). Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. *The Lancet*, 373(9681):2125 – 2135.

Hopewell, S., Boutron, I., Altman, D. G., and Ravaud, P. (2013). Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. *BMJ Open*, 3(8).

Ibrahim, J., Chen, M.-H., Gwon, Y., and Chen, F. (2014). The power prior: Theory and applications. Technical Report 21, Department of Statistics, University of Connecticut.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., and Feldman, B. M. (2010a). Methods to elicit beliefs for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology*, 63(4):355 – 369.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A., and Feldman, B. M. (2010b). A valid and reliable belief elicitation method for Bayesian priors. *Journal of Clinical Epidemiology*, 63(4):370 – 383.

Jüni, P., Witschi, A., Bloch, R., and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, 282(11):1054–1060.

Kaul, S., Bolger, A. F., Herrington, D., Giugliano, R. P., Eckel, R. H., Association, A. H., and Foundation, A. C. O. C. (2010). Thiazolidinedione drugs and cardiovascular risks: a science advisory from the american heart association and american college of cardiology foundation. *Journal of the American College of Cardiology*, 55(17):1885–1894.

Kuoppala, J., Lamminpaa, A., and Pukkala, E. (2008). Statins and cancer: A systematic review and meta-analysis. *European journal of cancer (Oxford, England : 1990)*, 44(15):2122–2132.

Linstone, H. and Turoff, M. (1975). *The Delphi Method: Tecniques and Applications.* Addison-Wesley, Reading, MA.

Lipscombe, L. L., Gomes, T., Lévesque, L. E., Hux, J. E., Juurlink, D. N., and Alter, D. A. (2007). Thiazolidinediones and cardiovascular outcomes in older patients with diabetes. *JAMA: The Journal of the American Medical Association*, 298(22):2634–2643.

Loke, Y. K., Kwok, C. S., and Singh, S. (2011). Comparative cardiovascular effects of thiazolidinediones: systematic review and meta-analysis of observational studies. *BMJ*, 342.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2):147 – 163.

Moja, L. P., Telaro, E., D'Amico, R., Moschetti, I., Coe, L., and Liberati, A. (2005). Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. *BMJ*, 330(7499):1053.

Neelon, B. and O'Malley, James, A. (2009). The use of power prior distributions for incorporating historical data into a Bayesian analysis. technical report. Department of Health Care Policy, Harvard Medical School.

Neelon, B. and O'Malley, A. J. (2010). Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics & Biostatistics*, 0(0):1–103.

Neuenschwander, B., Branson, M., and Spiegelhalter, David, J. (2009). A note on the power prior. *Statistics in Medicine*, 28:3562–3566.

Nissen, S. and Wolski, K. (2010). Rosiglitazone revisited: An updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. *Archives of Internal Medicine*, 170(14):1191–1201.

O'Hagan, A., Buck, C. E., Daneshkah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities.* Chichester: John Wiley and Sons Ltd.

Ojala, K., Vääräsmäki, M., Mäkikallio, K., Valkama, M., and Tekay, A. (2006). A comparison of intrapartum automated fetal electrocardiography and conventional cardiotocography - a randomised controlled study. *BJOG: An international journal of obstetrics and gynaecology*, 113:419–423.

O'Neill, R. T. (1998). Biostatistical considerations in pharmacovigilance and pharmacoepidemiology: linking quantitative risk assessment in pre-market licensure application safety data, post-market alert reports and formal epidemiological studies. *Statistics in medicine*, 17(15-16):1851–8; discussion 1859–62.

Pibouleau, L. and Chevret, S. (2011). Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices. *Journal of Clinical Epidemiology*, 64(3):270 – 279.

Powell, C. (2003). The Delphi technique: myths and realities. *Journal of Advanced Nursing*, 41(4):376–382.

Psaty, B. M. and Furberg, C. D. (2007). The record on rosiglitazone and the risk of myocardial infarction. *The New England journal of medicine*, 357(1):67–69.

Psaty, B. M. and Prentice, R. L. (2010). Minimizing bias in randomized trials: the importance of blinding. *JAMA : the journal of the American Medical Association*, 304(7):793–794.

Rietbergen, C., Groenwold, R. H. H., Hoijtink, H. J. A., Moons, K. G. M., and Klugkist, I. (2014). Expert elicitation of study weights for bayesian analysis and meta-analysis. *Journal of Mixed Methods Research*.

Rietbergen, C., Klugkist, I., Janssen, K., Moons, K., and Hoijtink, H. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary clinical trials*, 32(6):848–855.

Schulz, K. F., Altman, D. G., Moher, D., and (2010). Consort 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, 152(11):726–732.

Senn, S. (2007). *Statistical Issues in Drug Development.* John Wiley & Sons, Ltd, West Sussex, England, 2nd edition.

Shrier, I., Boivin, J.-F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., Brophy, J., and Rossignol, M. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? a critical examination of underlying principles. *American Journal of Epidemiology*, 166(10):1203–1209.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* John Wiley and Sons, UK: Chichester.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., and Abrams, K. (1999). An introduction to Bayesian methods in health technology assessment. *BMJ*, 319(7208):508–512.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., and Abrams, K. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessment*, 4(38).

Stocker, D. J., Taylor, A. J., Langley, R. W., Jezior, M. R., and Vigersky, R. A. (2007). A randomized trial of the effects of rosiglitazone and metformin on inflammation and subclinical atherosclerosis in patients with type 2 diabetes. *American Heart Journal*, 153(3):445.e1–445.e6.

Sung, L., Hayden, J., Greenberg, M. L., Koren, G., Feldman, B. M., and Tomlinson, G. A. (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology*, 58(3):261 – 268.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical methods in medical research*, 10(4):277–303.

Sutton, A. J., Cooper, N. J., Lambert, P. C., Jones, D. R., Abrams, K. R., and Sweeting, M. J. (2002). Meta-analysis of rare and adverse event data. *Expert review of pharmacoeconomics & outcomes research*, 2(4):367–379.

The BaSiS Group (2001). Bayesian standards in science (BaSiS). Draft available at: http://lib.stat.cmu.edu/bayeswork shop/2001/BaSis.html.

Tzoulaki, I., Molokhia, M., Curcin, V., Little, M. P., Millett, C. J., Ng, A., Hughes, R. I., Khunti, K., Wilkins, M. R., Majeed, A., and Elliott, P. (2009). Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs: retrospective cohort study using UK general practice research database. *BMJ*, 339.

van der Graaf, Y. (2013). De dokter als bayesiaan. *Nederlands Tijdschrift voor Geneeskunde*, 157:B965.

Vayssière, C., David, E., Meyer, N., Haberstich, R., Sebahoun, V., Roth, E., Favre, R., Nisand, I., and Langer, B. (2007). A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. *American Journal of Obstetrics and Gyneacology*, 197:299.e1–299.e6.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gtzsche, P. C., Vandenbroucke, J. P., and (2007). The strengthening the reporting of observational studies in epidemiology (strobe) statement: Guidelines for reporting observational studies. *Annals of Internal Medicine*, 147(8):573–577.

Weaver, J., Grenade, L. L., Kwon, H., and Avigan, M. (2009). Finding, evaluating, and managing drug-related risks: approaches taken by the us food and drug administration (fda). *Dermatologic therapy*, 22(3):204–215.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., and Tugwell, P. (2013). The newcastle-ottawa scale (nos) for assessing the quality of nonrandomised studies in meta-analyses.

Welton, N., Sutton, A., Cooper, N., Abrams, R., and Ades, A. (2012). *Evidence Synthesis for Decision Making in Healthcare.* John Wiley & Sons, Ltd.

Westerhuis, M., Moons, K., van Beek, E., Bijvoet, S., Drogtrop, A., van Geijn, H., van Lith, J., Mol, B., Nijhuis, J., Oei, S., Porath, M., Rijnders, R., Schuitemaker, N., van der Tweel, I., Visser, G., Willekes, C., and Kwee, A. (2009). A randomised clinical trial on cardiotocography plus fetal blood sampling versus cardiotocography plus ST-analysis of the fetal electrocardiogram (STAN) for intrapartum monitoring. *BMC Pregnancy and Childbirth*, 7(13).

Westgate, J., Harris, M., Curnow, J., and Greene, K. (1993). Plymouth randomized trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. *American Journal of Obstetrics and Gyneacology*, 169:1151–1160.

Studies included in the meta-analysis in Chapter 5

Bailey, C. J., Bagdonas, A., Rubes, J., McMorn, S. O., Donaldson, J., Biswas, N., and Stewart, M. W. (2005). Rosiglitazone/metformin fixed-dose combination compared with uptitrated metformin alone in type 2 diabetes mellitus: a 24-week, multicenter,randomized, double-blind, parallel-group study. *Clinical therapeutics*, 27(10):1548-1561

Baksi, A., James, R. E., Zhou, B., and Nolan, J. J. (2004). Comparison of uptitration of gliclazide with the addition of rosiglitazone to gliclazide in patients with type 2 diabetes inadequately controlled on half-maximal doses of a sulphonylurea. *Acta Diabetologica*, 41(2):63-69

Barnett, A. H., Grant, P. J., Hitman, G. A., Mather, H., Pawa, M., Robertson, L., Trelfa, A., and Investigators, I.-A. T. (2003). Rosiglitazone in type 2 diabetes mellitus: an evaluation in british indo-asian patients. *Diabetic medicine : a journal of the British Diabetic Association*, 20(5):387-393

Bertrand, O. F., Poirier, P., Rodes-Cabau, J., Rinfret, S., Title, L. M., Dzavik, V., Natarajan, M., Angel, J., Batalla, N., Almeras, N., Costerousse, O., Larochelliere, R. D., Roy, L., Despres, J. P., and Investigators, V. T. (2010). Cardiometabolic effects of rosiglitazone in patients with type 2 diabetes and coronary artery bypass grafts: A randomized placebo-controlled clinical trial. *Atherosclerosis*, 211(2):565-573

Bilik, D., McEwen, L. N., Brown, M. B., Selby, J. V., Karter, A. J., Marrero, D. G., Hsiao, V. C., Tseng, C. W., Mangione, C. M., Lasser, N. L., Crosson, J. C., and Herman, W. H. (2010). Thiazolidinediones, cardiovascular disease and cardiovascular mortality: translating research into action for diabetes (triad). *Pharmacoepidemiology and drug safety*, 19(7):715-721

Chou, H. S., Palmer, J. P., Jones, A. R., Waterhouse, B., Ferreira-Cornwell, C., Krebs, J., and Goldstein, B. J. (2008). Initial treatment with fixed-dose combination rosiglitazone/glimepiride in patients with previously untreated type 2 diabetes. *Diabetes, obesity and metabolism*, 10(8):626-637

Dargie, H. J., Hildebrandt, P. R., Riegger, G. A., McMurray, J. J., McMorn, S. O., Roberts, J. N., Zambanini, A., and Wilding, J. P. (2007). A randomized, placebocontrolled trial assessing the effects of rosiglitazone on echocardiographic function and cardiac status in type 2 diabetic patients with new york heart association functional class i or ii heart failure. *Journal of the American College of Cardiology*, 49(16):1696-1704

Davidson, J. A., McMorn, S. O., Waterhouse, B. R., and Cobitz, A. R. (2007). A 24-week, multicenter, randomized, double-blind, placebo-controlled, parallel-group study of the effcacy and tolerability of combination therapy with rosiglitazone and sulfonylurea in african american and hispanic american patients with type 2 diabetes inadequately controlled with sulfonylurea monotherapy. *Clinical therapeutics*, 29(9):1900-1914

Derosa, G., Cicero, A. F., Gaddi, A., Ragonesi, P. D., Fogari, E., Bertone, G., Ciccarelli, L., and Piccinni, M. N. (2004). Metabolic effects of pioglitazone and rosiglitazone in patients with diabetes and metabolic syndrome treated with glimepiride: a twelve-month, multicenter, double-blind, randomized, controlled, parallel-group trial. *Clinical therapeutics*, 26(5):744-754

Derosa, G., Gaddi, A. V., Piccinni, M. N., Ciccarelli, L., Salvadeo, S., Peros, E., Ghelfi, M., Ferrari, I., and Cicero, A. F. (2005). Antithrombotic effects of rosiglitazone-metformin versus glimepiride-metformin combination therapy in patients with type 2 diabetes mellitus and metabolic syndrome. *Pharmacotherapy*, 25(5):637-645

Dore, D. D., Trivedi, A. N., Mor, V., and Lapane, K. L. (2009). Association between extent of thiazolidinedione exposure and risk of acute myocardial infarction. *Pharmacotherapy*, 29(7):775-783

Dormuth, C. R., Maclure, M., Carney, G., Schneeweiss, S., Bassett, K., and Wright, J. M. (2009). Rosiglitazone and myocardial infarction in patients previously prescribed metformin. *PloS one*, 4(6):e6080

Finn, A. V., Oh, J. S., Hendricks, M., Daher, M., Cagliero, E., Byrne, R. M., Nadelson, J., Crimins, J., Kastrati, A., Schomig, A., Bruskina, O., Palacios, I., John, M. C., and Gold, H. K. (2009). Predictive factors for in-stent late loss and coronary lesion progression in patients with type 2 diabetes mellitus randomized to rosiglitazone or placebo. *American Heart Journal*, 157(2):383.e1-383.e8

Fonseca, V., Rosenstock, J., Patwardhan, R., and Salzman, A. (2000). Effect of metformin and rosiglitazone combination therapy in patients with type 2 diabetes mellitus: a randomized controlled trial. *JAMA : the journal of the American Medical Association*, 283(13):1695-1702

Gerrits, C. M., Bhattacharya, M., Manthena, S., Baran, R., Perez, A., and Kupfer, S. (2007). A comparison of pioglitazone and rosiglitazone for hospitalization for acute myocardial infarction in type 2 diabetes. *Pharmacoepidemiology and drug safety*, 16(10):1065-1071

Gerstein, H. C., Ratner, R. E., Cannon, C. P., Serruys, P. W., Garcia-Garcia, H. M., van Es, G. A., Kolatkar, N. S., Kravitz, B. G., Miller, D. M., Huang, C., Fitzgerald, P. J., Nesto, R. W., and Group, A. S. (2010). Effect of rosiglitazone on progression of coronary atherosclerosis in patients with type 2 diabetes mellitus and coronary artery disease: the assessment on the prevention of progression by rosiglitazone on atherosclerosis in diabetes patients with cardiovascular history trial. *Circulation*, 121(10):1176-1187

Gomez-Perez, F. J., Fanghanel-Salmon, G., Barbosa, J. A., Montes-Villarreal, J., Berry, R. A., Warsi, G., and Gould, E. M. (2002). Effcacy and safety of rosiglitazone plus metformin in mexicans with type 2 diabetes. *Diabetes/metabolism research and reviews*, 18(2):127-134

Graham, D. J., Ouellet-Hellstrom, R., MaCurdy, T. E., Ali, F., Sholley, C., Worrall, C., and Kelman, J. A. (2010). Risk of acute myocardial infarction, stroke, heart failure, and death in elderly medicare patients treated with rosiglitazone or pioglitazone. *JAMA : the journal of the American Medical Association*, 304(4):411-418

Hanefeld, M., Patwardhan, R., Jones, N. P., and Group, R. C. T. S. (2007). A one-year study comparing the effcacy and safety of rosiglitazone and glibenclamide in the treatment of type 2 diabetes. *Nutrition, metabolism, and cardiovascular diseases : NMCD*, 17(1):13-23

Home, P. D., Pocock, S. J., Beck-Nielsen, H., Curtis, P. S., Gomis, R., Hanefeld, M., Jones, N. P., Komajda, M., McMurray, J. J., and Team, R. S. (2009). Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (record): a multicentre, randomised, open-label trial. *Lancet*, 373(9681):2125-2135

Hsiao, F. Y., Huang, W. F., Wen, Y. W., Chen, P. F., Kuo, K. N., and Tsai, Y. W. (2009). Thiazolidinediones and cardiovascular events in patients with type 2 diabetes mellitus: a retrospective cohort study of over 473,000 patients using the national health insurance database in taiwan. *Drug safety : an international journal of medical toxicology and drug experience*, 32(8):675-690

Juurlink, D. N., Gomes, T., Lipscombe, L. L., Austin, P. C., Hux, J. E., and Mamdani, M. M. (2009). Adverse cardiovascular events during treatment with pioglitazone and rosiglitazone: population based cohort study. *BMJ (Clinical research ed.)*, 339:b2942

Kahn, S. E., Haffner, S. M., Heise, M. A., Herman, W. H., Holman, R. R., Jones, N. P., Kravitz, B. G., Lachin, J. M., O'Neill, M. C., Zinman, B., Viberti, G., and Group, A. S. (2006). Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *The New England journal of medicine*, 355(23):2427-2443

Kerenyi, Z., Samer, H., James, R., Yan, Y., and Stewart, M. (2004). Combination therapy with rosiglitazone and glibenclamide compared with upward titration of glibenclamide alone in patients with type 2 diabetes mellitus. *Diabetes research and clinical practice*, 63(3):213-223

Koro, C. E., Fu, Q., and Stender, M. (2008). An assessment of the effect of thiazolidinedione exposure on the risk of myocardial infarction in type 2 diabetic patients. *Pharmacoepidemiology and drug safety*, 17(10):989-996

Lebovitz, H. E., Dole, J. F., Patwardhan, R., Rappaport, E. B., Freed, M. I., and Group, R. C. T. S. (2001). Rosiglitazone monotherapy is effective in patients with type 2 diabetes. *The Journal of clinical endocrinology and metabolism*, 86(1):280-288

Lipscombe, L. L., Gomes, T., Levesque, L. E., Hux, J. E., Juurlink, D. N., and Alter, D. A. (2007). Thiazolidinediones and cardiovascular outcomes in older patients with diabetes. *JAMA : the journal of the American Medical Association*, 298(22):2634-2643

McAfee, A. T., Koro, C., Landon, J., Ziyadeh, N., and Walker, A. M. (2007). Coronary heart disease outcomes in patients receiving antidiabetic agents. *Pharmacoepidemiology and drug safety*, 16(7):711-725

Phillips, L. S., Grunberger, G., Miller, E., Patwardhan, R., Rappaport, E. B., Salzman, A., and Group, R. C. T. S. (2001). Once- and twice-daily dosing with rosiglitazone improves glycemic control in patients with type 2 diabetes. *Diabetes care*, 24(2):308-315

Raskin, P., Lewin, A., Reinhardt, R., Lyness, W., and Group, R. F.-D. C. S. (2009). Twice-daily dosing of a repaglinide/metformin fixed-dose combination tablet provides glycaemic control comparable to rosiglitazone/metformin tablet. *Diabetes, obesity and metabolism*, 11(9):865-873

Raskin, P., McGill, J., Saad, M. F., Cappleman, J. M., Kaye, W., Khutoryansky, N., Hale, P. M., and Group, R. S. (2004). Combination therapy for type 2 diabetes: repaglinide plus rosiglitazone. *Diabetic medicine : a journal of the British Diabetic Association*, 21(4):329-335

Raskin, P., Rendell, M., Riddle, M. C., Dole, J. F., Freed, M. I., Rosenstock, J., and Group, R. C. T. S. (2001). A randomized trial of rosiglitazone therapy in patients with inadequately controlled insulin-treated type 2 diabetes. *Diabetes care*, 24(7):1226-1232

Rosenstock, J., Niggli, M., and Maldonado-Lutomirsky, M. (2009). Long-term 2-year safety and effcacy of vildagliptin compared with rosiglitazone in drug-naive patients with type 2 diabetes mellitus. *Diabetes, obesity and metabolism*, 11(6):571-578

Sara

dis, P. A., Lasaridis, A. N., Nilsson, P. M., Mouslech, T. F., Hitoglou- Makedou, A. D., Stafylas, P. C., Kazakos, K. A., Yovos, J. G., and Tourkantonis, A. A. (2005). The effect of rosiglitazone on novel atherosclerotic risk factors in patients with type 2 diabetes mellitus and hypertension. an open-label observational study. *Metabolism: clinical and experimental*, 54(9):1236-1242

Stockl, K. M., Le, L., Zhang, S., and Harada, A. S. (2009). Risk of acute myocardial infarction in patients treated with thiazolidinediones or other antidiabetic medications. *Pharmacoepidemiology and drug safety*, 18(2):166-174

Tzoulaki, I., Molokhia, M., Curcin, V., Little, M. P., Millett, C. J., Ng, A., Hughes, R. I., Khunti, K., Wilkins, M. R., Majeed, A., and Elliott, P. (2009). Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs: retrospective cohort study using uk general practice research database. *BMJ (Clinical research ed.)*, 339:b4731

Wang, G., Wei, J., Guan, Y., Jin, N., Mao, J., and Wang, X. (2005). Peroxisome proliferator-activated receptor-gamma agonist rosiglitazone reduces clinical inflammatory responses in type 2 diabetes with coronary artery disease after coronary angioplasty. *Metabolism: clinical and experimental*, 54(5):590-597

Weissman, P., Goldstein, B. J., Rosenstock, J., Waterhouse, B., Cobitz, A. R., Wooddell, M. J., and Strow, L. J. (2005). Effects of rosiglitazone added to submaximal doses of metformin compared with dose escalation of metformin in type 2 diabetes: the empire study. *Current medical research and opinion*, 21(12):2029-2035

Winkelmayer, W. C., Setoguchi, S., Levin, R., and Solomon, D. H. (2008). Comparison of cardiovascular outcomes in elderly patients with diabetes who initiated rosiglitazone vs pioglitazone therapy. *Archives of Internal Medicine*, 168(21):2368-2375

Wolffenbuttel, B. H., Gomis, R., Squatrito, S., Jones, N. P., and Patwardhan, R. N. (2000). Addition of low-dose rosiglitazone to sulphonylurea therapy improves glycaemic control in type 2 diabetic patients. *Diabetic medicine : a journal of the British Diabetic Association*, 17(1):40-47

Wong, T. Y., Szeto, C. C., Chow, K. M., Leung, C. B., Lam, C. W., and Li, P. K. (2005). Rosiglitazone reduces insulin requirement and c-reactive protein levels in type 2 diabetic patients receiving peritoneal dialysis. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, 46(4):713-719

Yilmaz, H., Gursoy, A., Sahin, M., and Demirag, N. G. (2007). Comparison of insulin monotherapy and combination therapy with insulin and metformin or insulin and rosiglitazone or insulin and acarbose in type 2 diabetes. *Acta Diabetologica*, 44(4):187-192

Zhu, X. X., Pan, C. Y., Li, G. W., Shi, H. L., Tian, H., Yang, W. Y., Jiang, J., Sun, X. C., Davies, C., and Chow, W. H. (2003). Addition of rosiglitazone to existing sulfonylurea treatment in chinese patients with type 2 diabetes and exposure to hepatitis b or c. *Diabetes technology and therapeutics*, 5(1):33-42

Ziyadeh, N., McAfee, A. T., Koro, C., Landon, J., and Chan, K. A. (2009). The thiazolidinediones rosiglitazone and pioglitazone and the risk of coronary heart disease: a retrospective cohort study using a us health insurance database. *Clinical therapeutics*, 31(11):2665-2677

Studies included in the systematic review in Chapter 6

Alexeeff, S. E., Baccarelli, A. A., Halonen, J., Coull, B. A., Wright, R. O., Tarantini, L., Bollati, V., Sparrow, D., Vokonas, P., and Schwartz, J. (2013). Association between blood pressure and dna methylation of retrotransposons and pro-inflammatory genes. *International Journal of Epidemiology*, 42(1):270-280

Alkema, L., Kantorova, V., Menozzi, C., and Biddlecom, A. (2013). National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. *The Lancet*, 381(9878):1642-1652

Anderson, B. G. and Bell, M. L. (2009). Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States. *Epidemiology*, 20:205-213

Assiri, A., McGeer, A., Perl, T. M., Price, C. S., Al Rabeeah, A. A., Cummings, D. A., Alabdullatif, Z. N., Assad, M., Almulhim, A., Makhdoom, H., Madani, H., Alhakeem, R., Al-Tawfiq, J. A., Cotten, M., Watson, S. J., Kellam, P., Zumla, A. I., and Memish, Z. A. (2013). Hospital outbreak of middle east respiratory syndrome coronavirus. *New England Journal of Medicine*, 369(5):407-416.

Baccini, M., Biggeri, A., Grillo, P., Consonni, D., and Bertazzi, P. A. (2011a). Health impact assessment of fine particle pollution at the regional level. *American Journal of Epidemiology*, 174(12):1396-1405

Baccini, M., Kosatsky, T., Analitis, A., Anderson, H. R., D'Ovidio, M., Menne, B., Michelozzi, P., Biggeri, A., and the PHEWE Collaborative Group (2011b). Impact of heat on mortality in 15 european cities: attributable deaths under different weather scenarios. *Journal of Epidemiology and Community Health*, 65(1):64-70

Baeten, D., Baraliakos, X., Braun, J., Sieper, J., Emery, P., van der Heijde, D., McInnes, I., van Laar, J. M., Landew, R., Wordsworth, P., Wollenhaupt, J., Kellner, H., Paramarta, J., Wei, J., Brachat, A., Bek, S., Laurent, D., Li, Y., Wang, Y. A., Bertolino, A. P., Gsteiger, S.,Wright, A. M., and Hueber, W. (2013). Anti-interleukin 17a monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *The Lancet*, 382(9906):1705-1713

Baird, D. D., Dunson, D. B., Hill, M. C., Cousins, D., and Schectman, J. M. (2007). Association of physical activity with development of uterine leiomyoma. *American Journal of Epidemiology*, 165(2):157-163

Baird, D. D., Travlos, G., Wilson, R., Dunson, D. B., Hill, M. C., D'Aloisio, A. A., London, S. J., and Schectman, J. M. (2009). Uterine Leiomyomata in Relation to Insulin-like Growth Factor-I, Insulin, and Diabetes. *Epidemiology*, 20:604-610

van Ballegooijen, W. M., van Houdt, R., Bruisten, S. M., Boot, H. J., Coutinho, R. A., and Wallinga, J. (2009). Molecular sequence data of hepatitis B virus and genetic diversity after vaccination. *American Journal of Epidemiology*, 170(12):1455- 1463

Barnett, A. G. (2007). Temperature and cardiovascular deaths in the US elderly: changes over time. *Epidemiology*, 18:369-372

Barnett, A. G., Batra, R., Graves, N., Edgeworth, J., Robotham, J., and Cooper, B. (2009). Using a longitudinal model to estimate the effect of methicillin-resistant Staphylococcus aureus infection on length of stay in an intensive care unit. *American Journal of Epidemiology*, 170(9):1186-1194

te Beest, D. E., Wallinga, J., Donker, T., and van Boven, M. (2013). Estimating the generation interval of influenza a (h1n1) in a range of social settings. *Epidemiology*, 24(2):244-250

Belanger, K., Holford, T. R., Gent, J. F., Hill, M. E., Kezik, J. M., and Leaderer, B. P. (2013). Household levels of nitrogen dioxide and pediatric asthma severity. *Epidemiology* , 24(2):320

Bell, M. L., Ebisu, K., Peng, R. D., and Dominici, F. (2009). Adverse health effects of particulate air pollution: modification by air conditioning. *Epidemiology*, 20:682-686

Bellan, S. E., Fiorella, K. J., Melesse, D. Y., Getz, W. M., Williams, B. G., and Dushoff, J. (2013). Extra-couple fHIVg transmission in sub-saharan Africa: a mathematical modelling study of survey data. *The Lancet*, 381(9877):1561-1569

Best, N. and Hansell, A. L. (2009). Geographic variations in risk: adjusting for unmeasured confounders through joint modeling of multiple diseases. *Epidemiology*, 20:400-410

Block, J. P., Christakis, N. A., OMalley, A. J., and Subramanian, S. V. (2011). Proximity to food establishments and body mass index in the Framingham heart study offspring cohort over 30 years. *American Journal of Epidemiology*, 174(10):1108-1114

Boughey, J. C., Suman, V. J., Mittendorf, E. A., Ahrendt, G. M., Wilke, L. G., Taback, B., Leitch, A. M., Kuerer, H. M., Bowling, M., Flippo-Morton, T. S., et al. (2013). Sentinel lymph node surgery after neoadjuvant chemotherapy in patients with node-positive breast cancer: The acosog z1071 (alliance) clinical trial. *JAMA*, 310(14):1455-1461

Brazier, J. E., Fukuhara, S., Roberts, J., Kharroubi, S., Yamamoto, Y., Ikeda, S., Doherty, J., and Kurokawa, K. (2009). Estimating a preference-based index from the Japanese SF-36. *Journal of Clinical Epidemiology*, 62(12):1323-1331

Breban, R., Riou, J., and Fontanet, A. (2013). Interhuman transmissibility of middle east respiratory syndrome coronavirus: estimation of pandemic risk. *The Lancet*, 382(9893):694-699

Briggs, A. D. M., Mytton, O. T., Kehlbacher, A., Tiffin, R., Rayner, M., and Scarborough, P. (2013). Overall and income specific effect on prevalence of overweight and obesity of 20% sugar sweetened drink tax in uk: econometric and comparative risk assessment modelling study. *BMJ*, 347:f6189

Bryant, J. M., Grogono, D. M., Greaves, D., Foweraker, J., Roddick, I., Inns, T., Reacher, M., Haworth, C. S., Curran, M. D., Harris, S. R., et al. (2013). Whole-genome sequencing to identify transmission of *mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *The Lancet*, 381(9877):1551-1560

Cerda, M., Tracy, M., Messner, S. F., Vlahov, D., Tardiff, K., and Galea, S. (2009). Misdemeanor policing, physical disorder, and gun-related homicide: a spatial analytic test of ”broken-windows” theory. *Epidemiology*, 20:533-541

Chaix, B., Rosvall, M., and Merlo, J. (2007a). Neighborhood socioeconomic deprivation and residential instability: effects on incidence of ischemic heart disease and survival after myocardial infarction. *Epidemiology*, 18:104-111

Chaix, B., Rosvall, M., and Merlo, J. (2007b). Recent increase of neighborhood socioeconomic effects on ischemic heart disease mortality: A multilevel survival analysis of two large swedish cohorts. *American Journal of Epidemiology*, 165(1):22-26

Chandola, T., Clarke, P., Wiggins, R. D., and Bartley, M. (2005). Who you live with and where you live: setting the context for health using multiple membership multilevel models. *Journal of Epidemiology and Community Health*, 59(2):170-175

Chiavegatto Filho, A. D. P., Kawachi, I., Wang, Y. P., Viana, M. C., and Andrade, L. H. S. G. (2013). Does income inequality get under the skin? A multilevel analysis of depression, anxiety and mental disorders in Sao Paulo, Brazil. *Journal of Epidemiology and Community Health*, 67(11):966-972

Chu, H., Gange, S. J., Yamashita, T. E., Hoover, D. R., Chmiel, J. S., Margolick, J. B., and Jacobson, L. P. (2005). Individual variation in CD4 cell count trajectory among human immunodeffciency virus-infected men and women on long-term highly active antiretroviral therapy: an application using a Bayesian random change-point model. *American Journal of Epidemiology*, 162(8):787-797

Cleries, R., Martinez, J. M., Moreno, V., Yasui, Y., Ribes, J., and Borras, J. M. (2013). Predicting the change in breast cancer deaths in spain by 2019: a bayesian approach. *Epidemiology*, 24(3):454-460

Cotten, M., Watson, S. J., Kellam, P., Al-Rabeeah, A. A., Makhdoom, H. Q., Assiri, A., Al-Taw

q, J. A., Alhakeem, R. F., Madani, H., AlRabiah, F. A., Hajjar, S. A., Al-nassir, W. N., Albarrak, A., Flemban, H., Balkhy, H. H., Alsubaie, S., Palser, A. L., Gall, A., Bashford-Rogers, R., Rambaut, A., Zumla, A. I., and Memish, Z. A. (2013). Transmission and evolution of the middle east respiratory syndrome coronavirus in saudi arabia: a descriptive genomic study. *The Lancet*, 382(9909):1993-2002

Curtis, A. B., Worley, S. J., Adamson, P. B., Chung, E. S., Niazi, I., Sherfesee, L., Shinn, T., and St. John Sutton, M. (2013). Biventricular pacing for atrioventricular block and systolic dysfunction. *New England Journal of Medicine*, 368(17):1585-1593.

Degenhardt, L., Whiteford, H. A., Ferrari, A. J., Baxter, A. J., Charlson, F. J., Hall, W. D., Freedman, G., Burstein, R., Johns, N., Engell, R. E., Flaxman, A., Murray, C. J., and Vos, T. (2013). Global burden of disease attributable to illicit drug use and dependence: findings from the global burden of disease study 2010. The Lancet, 382(9904):1564-1574

Di Cesare, M., Bennett, J. E., Best, N., Stevens, G. A., Danaei, G., and Ezzati, M. (2013). The contributions of risk factor trends to cardiometabolic mortality decline in 26 industrialized countries. *International Journal of Epidemiology*, 42(3):838-848

Dominici, F., Peng, R. D., Zeger, S. L., White, R. H., and Samet, J. M. (2007). Particulate air pollution and mortality in the United States: did the risks change from 1987 to 2000? *American Journal of Epidemiology*, 166(8):880-888

Dowd, J. B., Albright, J., Raghunathan, T. E., Schoeni, R. F., LeClere, F., and Kaplan, G. A. (2011). Deeper and wider: income and mortality in the USA over three decades. *International Journal of Epidemiology*, 40(1):183-188

Engel, S. A. M., Erichsen, H. C., Savitz, D. A., Thorp, J., Chanock, S. J., and Olshan, A. F. (2005a). Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology*, 16:469-477

Engel, S. A. M., Olshan, A. F., Savitz, D. A., Thorp, J., Erichsen, H. C., and Chanock, S. J. (2005b). Risk of small-for-gestational age is associated with common anti-inflammatory cytokine polymorphisms. *Epidemiology*, 16:478-486

Feltbower, R. G., Manda, S. O. M., Gilthorpe, M. S., Greaves, M. F., Parslow, R. C., Kinsey, S. E., Bodansky, H. J., and McKinney, P. A. (2005). Detecting small-area similarities in the epidemiology of childhood acute lymphoblastic leukemia and diabetes mellitus, type 1: a Bayesian approach. *American Journal of Epidemiology*, 161(12):1168-1180

Glocker, E.-O., Hennigs, A., Nabavi, M., Schäffer, A. A., Woellner, C., Salzer, U., Pfeifer, D., Veelken, H.,Warnatz, K., Tahami, F., Jamal, S., Manguiat, A., Rezaei, N., Amirzargar, A. A., Plebani, A., Hannesschläger, N., Gross, O., Ruland, J., and Grimbacher, B. (2009). A homozygous CARD9 mutation in a family with susceptibility to fungal infections. *New England Journal of Medicine*, 361(18):1727-1735

Granich, R., Oh, P., Lewis, B., Porco, T., and Flood, J. (2005). Multidrug resistance among persons with tuberculosis in california, 1994-2003. *JAMA*, 293(22):2732-2739

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R. F., Brooker, S., Smith, D. L., Moyeed, R. A., and Snow, R. W. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med*, 6(3):e1000048

Holmes, D. R., Reddy, V. Y., Turi, Z. G., Doshi, S. K., Sievert, H., Buchbinder, M., Mullin, C. M., and Sick, P. (2009). Percutaneous closure of the left atrial appendage versus warfarin therapy for prevention of stroke in patients with atrial fibrillation: a randomised non-inferiority trial. *The Lancet*, 374(9689):534-542

Kelly, C. M., Schootman, M., Baker, E. A., Barnidge, E. K., and Lemes, A. (2007). The association of sidewalk walkability and physical disorder with area-level race and poverty. *Journal of Epidemiology and Community Health*, 61(11):978-983

Kivimäki, M., Lawlor, D. A., Smith, G. D., Eklund, C., Hurme, M., Lehtimäki, T., Viikari, J. S. A., and Raitakari, O. T. (2007). Variants in the CRP gene as a measure of lifelong differences in average C-reactive protein levels. *American Journal of Epidemiology*, 166(7):760-764

von Klot, S., Gryparis, A., Tonne, C., Yanosky, J., Coull, B. A., Goldberg, R. J., Lessard, D., Melly, S. J., Suh, H. H., and Schwartz, J. (2009). Elemental carbon exposure at residence and survival after acute myocardial infarction. *Epidemiology*, 20:547-554

Law, D. C. G., Klebanoff, M. A., Brock, J. W., Dunson, D. B., and Longnecker, M. P. (2005). Maternal serum levels of polychlorinated biphenyls and 1,1-dichloro-2,2-bis(p-chlorophenyl)ethylene (DDE) and time to pregnancy. *American Journal of Epidemiology*, 162(6):523-532

Leyland, A. H. (2005). Socioeconomic gradients in the prevalence of cardiovascular disease in scotland: the roles of composition and context. *Journal of Epidemiology and Community Health*, 59(9):799-803

Lian, M., Schootman, M., Doubeni, C. A., Park, Y., Major, J. M., Torres Stone, R. A., Laiyemo, A. O., Hollenbeck, A. R., Graubard, B. I., and Schatzkin, A. (2011). Geographic variation in colorectal cancer survival and the role of small-area socioeconomic deprivation: A multilevel survival analysis of the nih-aarp diet and health study cohort. *American Journal of Epidemiology*, 174(7):828-838

Liu, D., Shi, W., Shi, Y., Wang, D., Xiao, H., Li, W., Bi, Y., Wu, Y., Li, X., Yan, J., Liu, W., Zhao, G., Yang, W., Wang, Y., Ma, J., Shu, Y., Lei, F., and Gao, G. F. (2013). Origin and diversity of novel avian influenza a fH7N9g viruses causing human infection: phylogenetic, structural, and coalescent analyses. *The Lancet*, 381(9881):1926-1932

Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S. Y. W., Shapiro, B., Pybus, O. G., Allain, J.-P., and Hatzakis, A. (2009). The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med*, 6(12):e1000198

Marciante, K. D., Bis, J. C., Rieder, M. J., Reiner, A. P., Lumley, T., Monks, S. A., Kooperberg, C., Carlson, C., Heckbert, S. R., and Psaty, B. M. (2007). Reninangiotensin system haplotypes and the risk of myocardial infarction and stroke in pharmacologically treated hypertensive patients. *American Journal of Epidemiology*, 166(1):19-27

Matthews, F. E., Arthur, A., Barnes, L. E., Bond, J., Jagger, C., Robinson, L., and Brayne, C. (2013). A two-decade comparison of prevalence of dementia in individuals aged 65 years and older from three geographical areas of england: results of the cognitive function and ageing study i and fIIg. *The Lancet*, 382(9902):1405-1412

Mehtälä, J., Antonio, M., Kaltoft, M. S., OBrien, K. L., and Auranen, K. (2013). Competition between streptococcus pneumoniae strains: implications for vaccine induced replacement in colonization and disease. *Epidemiology*, 24(4):522-529

Morris, R. K., Malin, G. L., Quinlan-Jones, E., Middleton, L. J., Hemming, K., Burke, D., Daniels, J. P., Khan, K. S., Deeks, J., and Kilby, M. D. (2013). Percutaneous vesicoamniotic shunting versus conservative management for fetal lower urinary tract obstruction (pluto): a randomised trial. *The Lancet*, 382(9903):1496-1506

Murray, C. J., Richards, M. A., Newton, J. N., Fenton, K. A., Anderson, H. R., Atkinson, C., Bennett, D., Bernab, E., Blencowe, H., Bourne, R., Braithwaite, T., Brayne, C., Bruce, N. G., Brugha, T. S., Burney, P., Dherani, M., Dolk, H., Edmond, K., Ezzati, M., Flaxman, A. D., Fleming, T. D., Freedman, G., Gunnell, D., Hay, R. J., Hutchings, S. J., Ohno, S. L., Lozano, R., Lyons, R. A., Marcenes, W., Naghavi, M., Newton, C. R., Pearce, N., Pope, D., Rushton, L., Salomon, J. A., Shibuya, K., Vos, T., Wang, H., Williams, H. C., Woolf, A. D., Lopez, A. D., and Davis, A. (2013). fUKg health performance: findings of the global burden of disease study 2010. *The Lancet*, 381(9871):997-1020

Muss, H. B., Berry, D. A., Cirrincione, C. T., Theodoulou, M., Mauer, A. M., Kornblith, A. B., Partridge, A. H., Dressler, L. G., Cohen, H. J., Becker, H. P., Kartcheske, P. A., Wheeler, J. D., Perez, E. A., Wolff, A. C., Gralow, J. R., Burstein, H. J., Mahmood, A. A., Magrinat, G., Parker, B. A., and Hart, R. D. (2009). Adjuvant chemotherapy in older women with early-stage breast cancer. *New England Journal of Medicine*, 360(20):2055-2065

Næss, Ø., Piro, F. N., Nafstad, P., Smith, G. D., and Leyland, A. H. (2007). Air pollution, social deprivation, and mortality: a multilevel cohort study. *Epidemiology*, 18:686-694

Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L., and Samet, J. M. (2005). Seasonal analyses of air pollution and mortality in 100 US cities. *American Journal of Epidemiology*, 161(6):585-594

Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Dewi, M., Temperley, W. H., Williams, T. N., Weatherall, D. J., and Hay, S. I. (2013b). Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *The Lancet*, 381(9861):142-151

Piel, F. B., Hay, S. I., Gupta, S., Weatherall, D. J., and Williams, T. N. (2013a). Global burden of sickle cell anaemia in children under five, 20102050: Modelling based on demographics, excess mortality, and interventions. *PLoS Med*, 10(7):e1001484

Pirkola, S., Sund, R., Sailas, E., and Wahlbeck, K. (2009). Community mental-health services and suicide rate in Finland: a nationwide small-area analysis. *The Lancet*, 373(9658):147-153

Pitzer, V. E., Leung, G. M., and Lipsitch, M. (2007). Estimating variability in the transmission of severe acute respiratory syndrome to household contacts in Hong Kong, China. *American Journal of Epidemiology*, 166(3):355-363

Presanis, A. M., De Angelis, D., Hagy, A., Reed, C., Riley, S., Cooper, B. S., Finelli, L., Biedrzycki, P., Lipsitch, M., and The New York City Swine Flu Investigation Team (2009). The severity of pandemic H1N1 influenza in the United States, from April to July 2009: a Bayesian analysis. *PLoS Med*, 6(12):e1000207

Ramis-Prieto, R., Garcia-Perez, J., Pollan, M., Aragones, N., Perez-Gomez, B., and Lopez-Abente, G. (2007). Modelling of municipal mortality due to haematological neoplasias in Spain. *Journal of Epidemiology and Community Health*, 61(2):165-171

Richardson, G., Bloor, K., Williams, J., Russell, I., Durai, D., Cheung, W. Y., Farrin, A., and Coulton, S. (2009). Cost effectiveness of nurse delivered endoscopy: findings from randomised multi-institution nurse endoscopy trial (minuet). *BMJ*, 338:b270

Rosenbaum, J. E. (2009). Truth or consequences: the intertemporal consistency of adolescent self-report on the Youth Risk Behavior survey. *American Journal of Epidemiology*, 169(11):1388-1397

Rosso, S., Zanetti, R., Sanchez, M. J., Nieto, A., Miranda, A., Mercier, M., Loria, D., Østerlind, A., Greinert, R., Chirlaque, M.-D., Fabbrocini, G., Barbera, C., Sancho-Garnier, H., Lauria, C., Balzi, D., and Zoccola, M. (2007). Is 2,3,5-pyrroletricarboxylic acid in hair a better risk indicator for melanoma than traditional epidemiologic measures for skin phenotype? *American Journal of Epidemiology*, 165(10):1170-1177

Sacerdote, C., Guarrera, S., Smith, G. D., Grioni, S., Krogh, V., Masala, G., Mattiello, A., Palli, D., Panico, S., Tumino, R., Veglia, F., Matullo, G., and Vineis, P. (2007). Lactase persistence and bitter taste response: Instrumental variables and mendelian randomization in epidemiologic studies of dietary factors and cancer risk. *American Journal of Epidemiology*, 166(5):576-581

Salah, A. B., Kamarianakis, Y., Chlif, S., Alaya, N. B., and Prastacos, P. (2007). Zoonotic cutaneous leishmaniasis in central Tunisia: spatio-temporal dynamics. *International Journal of Epidemiology*, 36(5):991-1000

Saurina, C., Bragulat, B., Saez, M., and Lpez-Casasnovas, G. (2013). A conditional model for estimating the increase in suicides associated with the 2008-2010 economic recession in England. *Journal of Epidemiology and Community Health*, 67(9):779-787

Shiri, T., Auranen, K., Nunes, M. C., Adrian, P. V., van Niekerk, N., de Gouveia, L., von Gottberg, A., Klugman, K. P., and Madhi, S. A. (2013). Dynamics of pneumococcal transmission in vaccine-nave children and their hiv-infected or hiv uninfected mothers during the first 2 years of life. *American Journal of Epidemiology*, 178(11):1629-1637

Son, J.-Y., Lee, J.-T., Park, Y. H., and Bell, M. L. (2013). Short-term effects of air pollution on hospital admissions in korea. *Epidemiology*, 24(4):545-554

Song, F., Li, X., Zhang, M., Yao, P., Yang, N., Sun, X., Hu, F. B., and Liu, L. (2009). Association between heme oxygenase-1 gene promoter polymorphisms and type 2 diabetes in a chinese population. *American Journal of Epidemiology*, 170(6):747-756

Sweeting, M. J., Hope, V. D., Hickman, M., Parry, J. V., Ncube, F., Ramsay, M. E., and De Angelis, D. (2009). Hepatitis C infection among injecting drug users in England and Wales (1992-2006): there and back again? *American Journal of Epidemiology*, 170(3):352-360

Tarafder, M. R., Carabin, H., Gyorkos, T. W., and Joseph, L. (2009). Diarrhea and colds in child day care centers: Impact of various numerator and denominator deffinitions of illness episodes. *Epidemiology*, 20:796-799

Tarr, G. A. M., Eickhoff, J. C., Koepke, R., Hopfensperger, D. J., Davis, J. P., and Conway, J. H. (2013). Using a bayesian latent class model to evaluate the utility of investigating persons with negative polymerase chain reaction results for pertussis. *American Journal of Epidemiology*, 178(2):309-318

Tassone, E. C., Waller, L. A., and Casper, M. L. (2009). Small-area racial disparity in stroke mortality: an application of Bayesian spatial hierarchical modeling. Epidemiology, 20:234-241

Welty, L. J. and Zeger, S. L. (2005). Are the acute effects of particulate matter on mortality in the national morbidity, mortality, and air pollution study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology*, 162(1):80-88

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N., Burstein, R., Murray, C. J., and Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The Lancet*, 382(9904):1575-1586

Worby, C. J., Jeyaratnam, D., Robotham, J. V., Kypraios, T., O'Neill, P. D., De Angelis, D., French, G., and Cooper, B. S. (2013). Estimating the effectiveness of isolation and decolonization measures in reducing transmission of methicillin-resistant staphylococcus aureus in hospital general wards. *American Journal of Epidemiology*, 177(11):1306-1313

Xu, H., Short, S. E., and Liu, T. (2013). Dynamic relations between fast-food restaurant and body weight status: a longitudinal and multilevel analysis of Chinese adults. *Journal of Epidemiology and Community Health*, 67(3):271-279

Xu, W. H., Dai, Q., Xiang, Y. B., Long, J. R., Ruan, Z. X., Cheng, J. R., Zheng, W., and Shu, X. O. (2007). Interaction of soy food and tea consumption with CYP19A1 genetic polymorphisms in the development of endometrial cancer. *American Journal of Epidemiology*, 166(12):1420-1430

Yang, G., Wang, Y., Zeng, Y., Gao, G. F., Liang, X., Zhou, M., Wan, X., Yu, S., Jiang, Y., Naghavi, M., et al. (2013). Rapid health transition in china, 1990-2010: findings from the global burden of disease study 2010. The lancet, 381(9882):1987-2015

Yu, H., Alonso, W. J., Feng, L., Tan, Y., Shu, Y., Yang, W., and Viboud, C. (2013). Characterization of regional influenza seasonality patterns in China and implications for vaccination strategies: Spatio-temporal modeling of surveillance data. *PLoS Med*, 10(11):e1001552

Zajacova, A., Dowd, J. B., and Burgard, S. A. (2011). Overweight adults may have the lowest mortalitydo they have the best health? *American Journal of Epidemiology*, 173(4):430-437

Samenvatting in het Nederlands

Bij het interpreteren van onderzoeksresultaten houden onderzoekers op een informele en kwalitatieve manier rekening met resultaten uit eerdere onderzoeken. Dit is in lijn met een natuurlijke neiging om nieuwe informatie te beoordelen in het licht van hetgeen al bekend is. De power priorverdeling, zoals gepresenteerd door Ibrahim en Chen (2000) biedt een kwantitatieve en flexibele aanpak om historische resultaten formeel te incorporeren in de analyse van nieuwe gegevens. In deze procedure wordt gebruik gemaakt van een parameter die het gewicht van de historische onderzoeken representeert. Twee procedures voor het de specificatie van deze informatieve priorverdeling worden besproken en gevalueerd in deze thesis.

En daarvan is de simultane power priorverdeling waarbij de parameter voor het gewicht en de parameter waarin men genteresseerd is (bijvoorbeeld een effectgrootte) tezamen worden geschat uit de historische en nieuwe onderzoeksgegevens. Met deze aanpak hangt de grootte van het gewicht van het historische bewijs af van de mate waarin de historische en nieuwe data op elkaar lijken. Wanneer de historische data in overeenstemming zijn met de nieuwe data, dan krijgt de historische data een groter gewicht dan wanneer de data erg van elkaar verschillen. In deze thesis wordt de wenselijkheid van deze eigenschap ter discussie gesteld, omdat overeenkomsten en verschillen tussen twee datasets toevallige steekproefresultaten kunnen zijn. Een numeriek voorbeeld in het vierde hoofdstuk toont hoe twee steekproeven afkomstig uit dezelfde populatie uiteenlopende schattingen van een parameter kunnen opleveren. Hierdoor zou de grootte van het studiegewicht ernstig worden onderschat. We concluderen dat de grootte van het gewicht niet af zou moeten hangen van de overeenkomsten en verschillen tussen de historische en nieuwe onderzoeksgegevens, maar van overeenkomsten en verschillen in onderzoekskenmerken. Dit betekent dat de gewichten niet zouden moeten worden bepaald door de data, maar door de onderzoekers zelf.

Dit idee wordt gedeeld met andere onderzoekers die de power priorverdeling evalueerden. Zo stellen Neuenschwander et al. (2009) dat het niet aan te raden is om de parameter voor het studiegewicht als een onbekende te beschouwen. Ook Neelon and O'Malley (2009) waarschuwen ervoor dat de power priorverdeling de neiging heeft de historische data te weinig gewicht te geven zelfs wanneer de verschillen tussen de historische en nieuwe gegevens minimaal zijn. De alternatieve procedure waarbij de waarde van het studiegewicht als bekend wordt verondersteld, geeft de onderzoeker meer controle over de invloed van de historische data. De beschikbare literatuur gaat echter niet in op de methoden voor het eliciteren van deze gewichten bij experts, waarschijnlijk omdat het meenemen van expert kennis een ongewenste bron van subjectiviteit introduceert in het analyse proces.

Voor het bepalen van de studiegewichten moet een expert de kwaliteit en relevantie van de beschikbare kennis beoordelen. Dit is iets wat onderzoekers gewend zijn om te doen, zij het impliciet. De formele synthese van onderzoeksresultaten vraagt om een expliciete afweging van keuzes en een onderbouwing hiervan. Steeds moet een onderzoeker zich afvragen welke historische informatie kan worden meegenomen, op welke manier en in welke mate. De subjectiviteit die gepaard gaat met deze afwegingen is onvermijdelijk. Bovendien is de veelgebruikte aanpak waarbij men bij de analyse van onderzoeksgegevens enkel afgaat op de nieuwe gegevens evengoed een subjectieve keuze.

In het tweede hoofdstuk wordt de bruikbaarheid van de power priorverdeling onderzocht voor het schatten van een behandeleffect in een gerandomiseerd onderzoek wanneer de historische informatie afkomstig is uit gerandomiseerde onderzoeken met een net iets andere onderzoeksopzet of gebruikte populatie. Dit hoofdstuk wordt gellustreerd met data afkomstig uit gynaecologische klinische onderzoeken naar de effecten van intra-partum bewaking van de foetus op enkele klinische uitkomsten. Een sensitiviteitsanalyse laat zien dat enkel het ordenen van historische onderzoeken op grond van kwaliteit en relevantie, en het toekennen van gewichten op grond van die ordening, al resulteert in stabiele posterieure schattingen voor het effect van de interventie. Dit resultaat blijkt in dit geval niet af te hangen van de daadwerkelijke grootte van de gekozen gewichten.

Het derde hoofdstuk bouwt voort op dit idee door een panel van experts te vragen om aan een aantal historische onderzoeken een rangordening en bijpassende gewichten toe te kennen. Aan hen werd gevraagd hun keuzes schriftelijk te motiveren en deze onderbouwing werd anoniem met de andere experts gedeeld middels een Delphi techniek. Op deze manier werden de experts aangemoedigd elkaar te overtuigen van hun ideeen om zo gezamenlijk tot een beoordeling te komen van een verzameling historische onderzoeken. Dit onderzoek laat zien dat het in deze context mogelijk is om in een beperkt aantal ronden overeenstemming te bereiken in het panel met betrekking tot de ordening van de onderzoeken en een voldoende mate van convergentie van de studiewichten zelf. Dit resultaat, samen met de positieve resultaten uit het tweede hoofdstuk, toont de aantrekkelijkheid van deze procedure voor het eliciteren van studiegewichten.

Ondanks het feit dat de casussen zoals gepresenteerd in het tweede en derde hoofdstuk vrij specifiek zijn, is het mogelijk om enkele algemene aanbevelingen en conclusies af te leiden.

De gewichten die verkregen zijn met de Delphi methode kunnen direct worden gebruikt voor het specificeren van de power parameter. De variaties in gekozen gewichten tussen de experts kunnen worden gebruikt als leidraad bij het uitvoeren van sensitiviteitsanalyses. Zo kan bijvoorbeeld worden gevalueerd of de gekozen gewichten van de ene experts leiden tot andere posterieure resultaten dan de gewichten van een andere expert.

Onderzoekers wensen de power priorverdeling met vaste gewichten niet altijd te gebruiken in verband met de sterke subjectieve aard ervan. In dat geval is het mogelijk om de informatie die verkregen is van het expert panel te gebruiken voor het specificeren van een hyper priorverdeling voor het studiegewicht in de gezamenlijke power priorverdeling. Hobbs et al (2011) beschrijven de specificatie van deze hyper priorverdeling maar laten deze volledig afhangen van de mate van gelijkheid tussen de historische en nieuwe data. Het gebruik van expert kennis met betrekking tot de mate van gelijkheid tussen de kenmerken van de historische en nieuwe onderzoeken is nog onbesproken in de huidige literatuur. Echter, nader onderzoek is gewenst, omdat een onderzoeker op deze manier enerzijds kan profiteren van eigenschappen van de gezamenlijke power priorverdeling waarbij het specificeren van gewichten niet nodig is, terwijl hij anderzijds gebruik kan maken van de expert kennis voor het kiezen van een verstandige hyper priorverdeling.

Een belangrijk probleem dat is besproken in het derde hoofdstuk is de hoeveelheid werk die de experts moeten verrichten wanneer een groot aantal historische onderzoeken beoordeeld moet worden. Een Delphi benadering zou in dat geval niet haalbaar zijn. Ook zal het met meerdere onderzoeken lastiger zijn om convergentie te bereiken. Toekomstig onderzoek zou zich kunnen richten op het ontwikkelen van een efficintere procedure om grote aantallen gewichten te eliciteren. Wellicht zouden tussentijdse sensitiviteitsanalyses uit kunnen wijzen of het noodzakelijk is om naast een ordening van historische onderzoeken ook daadwerkelijke gewichten te eliciteren.

Niet alleen voor het specificeren van de power priorverdeling met vaste gewichten is de beschikbaarheid van meerdere historische onderzoeken complex, ook voor de specificatie van de simultane power priorverdeling zorgt dit voor moeilijkheden. De gezamenlijke power priorverdeling zoals gepresenteerd door Ibrahim en Chen (2000) kent slechts de mogelijkheid om twee historische onderzoeken te includeren, in andere gevallen zijn de gewichten ongedentificeerd. En manier om toch verschillende historische onderzoeken mee te nemen is door de resultaten van deze historische onderzoeken samen te vatten tot n priorverdeling waar vervolgens slechts n enkel gewicht aan wordt toegekend (zie bijvoorbeeld Welton et al. (2012)). In het vijfde hoofdstuk wordt deze aanpak gebruikt om informatie uit onderzoeken met een verschillende onderzoeksopzet samen te vatten. In dit geval wordt de informatie uit observationele onderzoeken gebruikt als input voor een power priorverdeling, die vervolgens wordt gebruikt voor het schatten van het effect van een interventie in een verzameling gerandomiseerde klinische onderzoeken. Per onderzoeksvraag werden de studiegewichten zodanig gekozen dat de priori precisie gelijk is aan de precisie in de nieuwe data. In navolging van Neelon and O'Malley (2009) worden referentie analyses uitgevoerd met studiegewichten gelijk aan nul en n. Door de resultaten van de verschillende analyses met elkaar te vergelijken kan de sensitiviteit van de posterieure verdeling voor de geschatte parameters voor verschillende priorverdelingen worden gevalueerd. Aangetoond wordt dat de invloed van de priorverdeling substantieel kan zijn, met name wanneer een beperkte hoeveelheid gegevens is verzameld in het nieuwe onderzoek. Hiermee is het belang van sensitiviteitsanalyses om de invloed van de priorverdeling te evalueren eens te meer bewezen.

In het zesde hoofdstuk wordt een systematische review beschreven naar het gebruik van Bayesiaanse data analyse technieken in medisch en epidemiologisch onderzoek. Een belangrijke bevinding is dat in de gevonden onderzoeken sensitiviteitsanalyses met betrekking tot de robuustheid van de posterieure verdeling voor veranderingen in de priorverdeling nauwelijks werden gerapporteerd. Een opmerkelijk resultaat gezien het feit dat deze analyses worden aangeraden in elke richtlijn over het rapporteren over Bayesiaanse analyses. Deze richtlijnen blijken berhaupt nauwelijks te worden gebruikt, waardoor de kwaliteit van de rapportages vaak beperkt is. Met name over het specificeren van de priorverdeling ontbreekt vaak essentile informatie in de rapportages, terwijl deze informatie cruciaal is voor het correct interpreteren van de posterieure resultaten. Ook onderzoekers die gebruik maakten van informatieve priorverdelingen rapporteerden vaak niet duidelijk welke informatie op welke wijze gebruikt werd om de priorverdeling te specificeren. Het nauwgezet volgen van bestaande richtlijnen lijkt in het geval van Bayesiaanse analyses juist extra van belang gezien de aanhoudende controverse aangaande het subjectieve proces van de specificatie van priorverdelingen. Het gebruik van Bayesiaanse technieken voor de synthese van onderzoeksresultaten zou kunnen worden verbeterd door richtlijnen over het rapporteren Bayesiaanse analyses op te nemen in veel gebruikte algemene richtlijnen over rapporteren van data-analyses.

Het merendeel van de gebruikte voorbeelden in deze thesis zijn medisch van aard. Binnen dit vakgebied is de noodzaak van synthese van onderzoeksresultaten enorm toegenomen door de opkomst van evidence based practice. Binnen de sociale wetenschappen zien we vergelijkbare ontwikkelingen. In dit veld is de synthese van onderzoeksresultaten extra complex door de grote heterogeniteit tussen onderzoeken wat betreft de onderzochte populaties, de gebruikte onderzoeksopzetten en de gekozen uitkomsten en metingen daarvan. Deze verschillen bemoeilijken het proces van de specificatie van de priorverdeling: elicitatie procedures zullen meer tijd vragen wanneer er binnen een panel van experts op voorhand al onenigheid is over de relevantie van de te beoordelen onderzoeken. Tegelijkertijd benadrukt deze heterogeniteit juist de noodzaak van het gebruik van geavanceerdere technieken voor de kwantitatieve synthese van bewijzen. Immers, in de aanwezigheid van een complexe set van historische onderzoeksresultaten heeft het weinig waarde om te focussen op de uitkomsten van n enkel nieuw onderzoek. Daarbij heeft een gezamenlijke analyse van onderzoeksresultaten znder aandacht te besteden aan de verschillen tussen de onderzoeken weinig betekenis. Net als in de medische wetenschappen is de evaluatie van relevantie en kwaliteit van onderzoeken onvermijdelijk als het gaat om het vinden van het beste bewijs. Daarmee zijn de resultaten zoals gevonden in deze thesis en daar buiten ook van toepassing op de sociale wetenschappen en kunnen onderzoekers in de sociale wetenschappen hun voordeel doen met methodologische ontwikkelingen uit medisch onderzoek.

Acknowledgements

I would like to take this opportunity to thank my colleagues, friends and family who contributed in any way to the completion of this dissertation.

First of all I would like to thank my supervisors Herbert Hoijtink and Carl Moons for their helpful ideas and advise. A special word of thanks for my daily supervisors Irene Klugkist and Rolf Groenwold for their inspiring discussions, helpful feedback, and for their patience. Kristel Janssen, I did not forget about you, thanks for starting up this project and for staying in touch.

Many thanks to all my co-authors for your valuable contributions to the papers included in this dissertation. Special thanks to Gudrun and Thomas for all the nice moments we spent drinking coffee and discussing our work and other interesting stuff, and to Ming-Hui Chen for inviting Irene and me over to Storrs. I will never forget Thanksgiving dinner at your house and the giant turkey we tried to eat with chopsticks.

I am lucky to have so many nice colleagues around that made and make life at the Uithof inspiring, pleasant and fun. Thanks to all of you and in special to my roommates throughout the years.

I could not have completed this dissertation if I was not surrounded by many sweet people outside the University that supported me: thanks to my friends, my family-in-law and my dear brothers Maarten en Niek for being there for me.

I certainly could not have completed this dissertation if I would not have received so much support from my parents Peter and Nora. Thank you for this and in special for taking care of little Nijs and Mels so often.

Finally, I definitely could not have completed this dissertation without the presence of my favourite colleague, my friend, and dedicated father of my beautiful sons. Thank you Nijs for preparing my lunch every day.