Bayesian Meta-Analysis:
Methods and Applications in Clinical Research

Habilitationsschrift

Mathematisch-Naturwissenschaftliche Fakultät
der Heinrich-Heine Universität Düsseldorf

vorgelegt von

Dr. rer. nat. Pablo Emilio Verde

aus Buenos Aires, Argentinien

June 2015

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

To

Isabel, Lucía, Heide,

to the living example of my mother

and

to the memory of my father.

Acknowledgments

I am very thankful to Christian Ohmann for supporting my work during all these years. Great part of the work presented in this thesis is coming from his research initiative. I thank Christian for his politeness and his strong intellectual sparring. I am very grateful to Martin Lercher for hosting my teaching activities in the Institute of Bioinformatics and for welcoming my research work, without his support and trust this work would not be possible. This work was supported by the German Research Foundation project DFG Oh 39/11-1 and DFG VE 986/1-1.

I have been incredibly fortunate to meet Brad Efron when I was a student. I thank Brad for our long standing friendship and for been a permanent example of in my life. My thanks to Graciela Boente and Martin Grondona for my statistical education in Argentina. I was immensely lucky to learn Bayesian Statistics directly from David Spiegelhalter and Nicky Best. I am indebted to Nicky for her support of our DFG project on Generalized Evidence Synthesis while she was head of the BIAS project, and to David for his help on my teaching Bayesian statistical courses all around the world.

My sincere acknowledgement to my colleges at the Coordination Center for Clinical Trials, in particular to Andreas Vogt for his side by side daily work. I am grateful to Joelle Murray and Sean Fortune for their tireless and professional proof reading of my articles. Of course all remaining typos are my responsibility. I am very thankful to those friends and family that every time we meet have encouraged me to finalize this work, those people are: Dimitris Venizeleas, Ajrun Shakiri, Frank Hoffmeister, Ulrich Langwald and Ricardo Filomena. I would like to thank my uncle Rafael Verde, who always reminded me that my dad would be proud of my work. My thanks to my parents in law Eve and Klaus.

Finally my immense gratitude to my wife Heide for her love, her infinite patience, for her outstanding family management and for her great support! Without her help this work would not be possible. All my gratitude and love go to my daughters Isabel and Lucía for their every day sunshine and to my mother for being a living example.

Contents

Chapter 1

Introduction

”This is the information age...The only trouble is that most of it is misinformation, or its uglier cousin, disinformation. There are many reasons to mis- or disinform the public, and small rewards for accuracy.”
Bradley Efron,
Public Policy and Statistics: Case Studies from RAND, 2000, pag. vii

One of the most important processes in science is the accumulation of information and knowledge. Ideally, for a particular research problem we should have a collection of experiments and studies which indicate the best way to proceed. However, this is not the case in several areas of *empirical research*. Instead, researchers have to face a heterogeneous and fragmented evidence coming from published articles, unpublished reports, databases, etc., that has to be analyzed together.

Meta-Analysis is a branch of statistical techniques that helps researchers to combine results and evidence from a multiplicity of sources in a coherent statistical model. During the last 20 years meta-analysis has been very popular in the *Evidence Base Medicine*, where it injected scientific formality in the evaluation and optimization of medical decision-making (Welton et al. 2012). However meta-analysis has been crossing the borders of different disciplines and it is currently applied to combine information in different areas such as: Genetics and Genomics (Guerra and Goldstein 2010; Veyrieras et al. 2007; Goffinet and Gerber 2000), Social Sciences (Boruch 2005), Economics (Stanley and Doucouliagos 2012), Ecology and Evolution (Koricheva et al. 2013) and Astrophysics (Hogg 2001; Vallee 2002).

This is a work of *Bayesian Meta-Analysis*, this approach of meta-analysis is characterized by the construction of formal probability models to combine multiple sources of information. The *Bayesian*

Meta-Analysis has its roots in the work of Eddy et al. (1992), where the interconnection of each piece of evidence is described by using a *Directed Acyclic Graph* (DAG). Inferential statements about model parameters are based on posterior distributions that are approximated by using Markov Chain Monte Carlo (MCMC) computation techniques. A gentle introduction to Bayesian Meta-Analysis can be found in (Spiegelhalter et al. 2004, Chap. 8).

Aims of this work

The following points summarize the main work in Bayesian Meta-Analysis investigated in this thesis:

- *Generalized Evidence Synthesis* is the extension of the meta-analysis to deal with results coming from different statistical designs (e.g. retrospective, prospective, observational, etc.). In this work we investigated the state of the art in this area by making a review of methods and applications of the published work in the last 20 years.

- *Cross Design Synthesis* is a meta-analysis technique to explore in which extent experimental results can be extrapolated into a new application framework. A new approach for *Cross Design Synthesis* was developed to investigate how to combine evidence from aggregated results (e.g. published papers) with individual participant data (e.g. databases), when the sources of information are coming from different statistical designs.

- *Meta-analysis of diagnostic test* is still an open research topic, in this work we investigated how to deal with the complexities of the evidence of diagnostic test data. We developed a novelty model that realistically include the multiple sources of variability in this type of meta-analysis.

- *Software development* has been a main focus of this work. We investigated how to implement a complex Bayesian Meta-Analysis model into an easy to use statistical software. We developed a package in the R system that simplifies the applications of Bayesian meta-analysis for non-statisticians.

Overview of the chapters

This thesis is a collection of articles in Bayesian meta-analysis, each chapter presents one of those papers. At the beginning of each chapter information about authorship is provided as well as publication status.

In Chapter 2 we present two letters to the editors (Curcio and Verde 2011; Verde and Curcio 2012) which have been motivated by methodological pitfalls in meta-analysis. The case in point was a meta-analysis performed after the safety communication in July 2010 issued by the The Food and Drug Administration. This communication warrants about the increased risk of death with Tygacil (tigecycline) compared to other antibiotics used to treat similar infections.

Researchers may have multiple motivations for combining disparate pieces of evidence in a meta-analysis, such as: generalizing experimental results or increasing the power to detect an effect that a single study is not able to detect. Chapter 3 presents a methodological review of *Generalized Evidence Synthesis* performed by Verde and Ohmann (2014). In this review we cover statistical methods that have been used for the evidence-synthesis of different study-types with the same outcome and similar interventions. For the methodological review a literature retrieval in the area of generalized evidence-synthesis was performed and publications were identified, assessed, grouped and classified. Furthermore real applications of these methods in medicine were identified and described. For these approaches 39 real clinical applications could be identified to save some pages in adobe reader. A new classification of methods is provided, which takes into account: the inferential approach, the bias modeling, the hierarchical structure and the use of graphical modeling. We conclude with a discussion of pros and cons of our approach and give some practical advice.

In Chapter 4 we present the recent work on *Cross Design Synthesis* of Verde et al. (2015). This paper describes a unified modeling framework to combine aggregated data from randomized controlled trials (RCTs) with individual participant data (IPD) from observational studies. Rather than simply pooling the available evidence into an overall treatment effect, adjusted for potential confounding, the intention of this work is to explore treatment effects in specific patient populations reflected by the IPD. In this way, by collecting IPD we can potentially gain new insights from RCTs' results which cannot be seen using only a meta-analysis of RCTs. We present a new Bayesian hierarchical meta-regression model which combines sub-models, representing different types of data, into a coherent analysis. We highlight different types of model's parameters: those which are the focus of inference (e.g. treatment effect in a subgroup of patients) and those which are used to adjust for biases introduced by data collection processes (e.g. internal or external validity). The methods are applied to a case study where RCTs' results, investigating efficacy in the treatment of diabetic foot problems, are extrapolated to groups of patients treated in medical routine and who were enrolled in a prospective cohort study.

Chapter 5 presents a short comment on conflict of evidence which has been my contribution to the discussion of Finegold and Drton (2014). The conflict of evidence is the deconstructionist part of meta-analysis, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. In Verde (2014) I conjectured that a way to perform conflict of evidence in a multi-parameter meta-analysis model was to extend the random effects distribution by using a scale mixture of normal distributions per random effect. I have called this technique *"splitting the studies' weights"* and it is implemented in the R package **bamdit**, which is the topic of Chapter 7.

Meta-analysis of diagnostic test data is another complex type of meta-analysis. It differs from other types of meta-analysis in several aspects: First, the diagnostic summaries that we aim to combine (e.g. sensitivity and specificity) could be interdependent and a marginal combination by pooling these quantities might be misleading. Second, diagnostic studies are usually performed under slightly different diagnostic setups and they can be applied to different patients' populations. Hence, we can expect high heterogeneity between studies' results. In Chapter 6 we present a novelty approach for meta-analysis of diagnostic test proposed by Verde (2010). This approach is based on flexible random-effects distributions based on scale mixture of Normals.

Usually, practitioners conducting meta-analysis are not statistical experts and combining results from diagnostic studies may become a challenge. In Chapter 7 we present the R package **bamidt** (Verde 2015), its name stands for "**Ba**yesian **m**eta-analysis of **di**agnostic **t**est-data". **bamdit** was developed with the aim of simplifying the use of models in meta-analysis, that up to now have demanded great statistical expertise in Bayesian meta-analysis. The package implements a series of innovative statistical techniques including: the Bayesian Summary Receiver Operating Characteristic (BSROC) curve, the use of prior distributions that avoid boundary estimation problems of component of variance and correlation parameters, analysis of conflict of evidence and robust estimation of model parameters. In addition, the package comes with several published examples of meta-analysis that can be used for illustration or further research in this area.

Bibliography

Boruch, R. (2005), “”What is the Campbell Collaboration and how is it helping to identify ”what works”?” *The Evaluation Exchange, Harvard Family Research Project.*, 11.

Curcio, D. and Verde, P. E. (2011), “Comment on: Efficacy and Safety of Tigecycline a Systematic Review and Meta-Analysis,” *J Antimicrob Chemother*, 66, 2893–5.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1992), *Meta-analysis by the confidence profile method: The statistical synthesis of evidence.*, Academic Press, San Diego, CA.

Finegold, M. and Drton, M. (2014), “Robust Bayesian Graphical Modeling Using Dirichlet t -Distributions,” *Bayesian Anal.*, 9, 521–550.

Goffinet, B. and Gerber, S. (2000), “Quantitative Trait Loci: A Meta-Analysis,” *Genetics*, 155, 463–473.

Guerra, R. and Goldstein, D. (2010), *Meta-Analysis and Combining Information in Genetics and Genomics*, CRC Press.

Hogg, D. W. (2001), “A meta-analysis of cosmic star-formation history,” *arXiv preprint astro-ph/0105280*.

Koricheva, J., Gurevitch, J., and Mengersen, K. (2013), *Handbook of Meta-analysis in Ecology and Evolution*, Princeton Universtity Press.

Spiegelhalter, D., Abrams, K. R., and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley & Sons, Ltd. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Stanley, T. D. and Doucouliagos, H. (2012), *Meta-Regression Analysis in Economics and Business*, Routledge Advances in Research Methods.

Vallee, J. P. (2002), “Metastudy of the Spiral Structure of Our Home Galaxy,” *The Astrophysical Journal*.

Verde, P. E. (2010), “An introduction of Bayesian data analysis with R and BUGS: a simple worked example,” *Estadistica*, 62, 21–44.

— (2014), “A comment mentioning possible application in meta-analysis of Dirichlet t-distributions,” *Bayesian Analysis*, 9, 589–590.

— (2015), “bamdit: an R packge for Bayesian meta-analysis of diagnostic test data,” *Journal of Statistical Software*.

Verde, P. E. and Curcio, D. (2012), “Imbalanced Mortality Evidence for Tigecycline: 2011, the Year of the Meta-Analysis,” *Clincial Infectious Disease*, 55, 471–472.

Verde, P. E. and Ohmann, C. (2014), “Combining randomized and non-randomized evidence in clinical research: a review of methods and applications,” *Research Synthesis Methods*, DOI: 10.1002/jrsm.1122.

Verde, P. E., Ohmann, C., Morbach, S., and Icks, A. (2015), “Bayesian evidence synthesis for exploring generalizability of treatment effects: a case study of combining randomized and non-randomized results in diabetes,” *SStatistics in Medicine (under review)*.

Veyrieras, J., Goffinet, B., and Charcosset, A. (2007), “MetaQTL: a package of new computational methdos for the meta-analysis of QTL mapping experiments,” *BMC Bioinformatics*, 8, 1–6.

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

Chapter 2

Meta-analyses of mortality rates of the Tigecycline antibiotic

> ”In moving beyond the confines of classical statistics,
> we are also moving outside its wall of protection”
>
> -Bradly Efron (2010) *Large-Scale Inference*, Prologue, page x.

Contributions of the authors:

As clinical researcher DC brought to discussion the results of the meta-analysis of Yahav et al. PEV designed and wrote the papers, presented the methodological critics and developed the new statistical method.

Journal of Antimicrobial Chemotherapy Advance Access published September 15, 2011

Journal of Antimicrobial Chemotherapy

J Antimicrob Chemother
doi:10.1093/jac/dkr368

Comment on: Efficacy and safety of tigecycline: a systematic review and meta-analysis

Daniel Curcio[1] and Pablo E. Verde[2]*

[1]Instituto Sacre Couer, Infectología Institucional SRL, Buenos Aires, Argentina; [2]Coordination Centre for Clinical Trials, University of Düsseldorf, Düsseldorf, Germany

*Corresponding author. Tel: +49-211-81-04129; Fax: +49-211-81-19702; E-mail: pabloemilio.verde@uni-duesseldorf.de

Keywords: randomized clinical trials, severe infections, test of heterogeneity

Sir,

In the September 2011 issue of the *Journal of Antimicrobial Chemotherapy*, Yahav *et al.*[1] published a systematic review and meta-analysis of 15 randomized clinical trials (RCTs) that compared tigecycline with other antibiotics for the treatment of severe infections. The overall 30 day mortality was estimated to be higher with tigecycline compared with other regimens [relative risk (RR) 1.29, 95% confidence interval (CI) 1.02–1.64]; therefore, the authors recommend that clinicians should avoid tigecycline monotherapy in the treatment of severe infections and reserve it as a last-resort drug.

The authors performed a test of heterogeneity between studies. Given that the test result was not significant at 5%, they decided to pool all the RRs by using a fixed-effect meta-analysis model. Unfortunately, this is a common practice in meta-analysis, which usually leads to very misleading results. First of all, the pooled RR as well as its standard error are sensitive to the estimation of the between-studies standard deviation (SD).[2] SD is difficult to estimate with a small number of studies. On the other hand, it is very well known that the significant test of heterogeneity lacks statistical power to detect values of SD greater than zero.[3] In addition, the statistically non-significant results of this test cannot be interpreted as evidence of the homogeneity of the results among all RCTs included.[4]

The profile likelihood of the SD in a random-effect model is an alternative method to analyse the evidence of heterogeneity in the RCTs included in the review;[3] Figure 1 presents this type of analysis. In the left panel we have the profile likelihood of SD, which summarizes the support from the RCTs for different values of SD. The broken lines are the 95% CI for SD (0–0.538). Clearly, a value of SD=0 has the maximum support; however, values of SD greater than zero (e.g. SD=0.1) might be considered as being reasonably supported by the data of

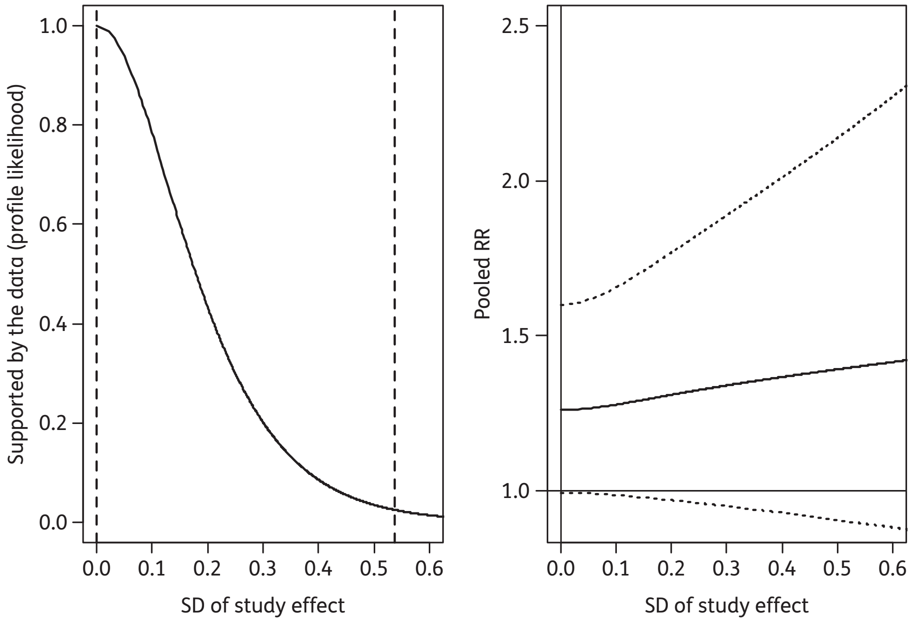


Figure 1. Meta-analysis sensitivity plot. Left panel: profile likelihood of the SD of between-study effects. The *y*-axis represents support from the data of the studies included in the meta-analysis (0=no support; 1=maximum support). The broken lines are the 95% CI of the SD of this variability parameter. Right panel: the *y*-axis is the pooled RR and the *x*-axis is the SD of between-study effects. The broken lines are the 95% CI of the RR for different values of the SD of between-study effects.

the RCTs. In the right panel we show how the pooled RR and its 95% CI change for different values of SD. For example, for SD=0.1 the pooled RR (95% CI) is 1.28 (0.987–1.656), which is not statistically significant. This sensitivity analysis shows that SD=0 is not a robust choice as an estimate, since small non-zero values of SD, which are well supported by the data, can have a strong influence on the conclusions. Therefore, a sharp conclusion based on SD=0 is misleading in this context.

The decision to pool studies with SD=0 is based on the assumption that the studies are identical, which is incorrect, mainly from a clinical point of view. For example, the RCT of hospital-acquired pneumonia presents a total mortality rate of 62.5%, while the mortality rates of the other studies are between 2.1% and 13.2%. That clearly casts doubt on the simplistic assumption of the homogeneity of the studies. In addition, the correlation between the RR of the studies and the total mortality rates in the logarithmic scale is −0.72, which indicates that the meta-analysis should include an adjustment for the total mortality rate.[5]

Lastly, the authors do not present any predictive quantities in the meta-analysis. The predictive summary statistics are considered the most important quantities in a meta-analysis.[6] The main reason is that these quantities are associated with the future use and the potential clinical use of the meta-analysis results. In the context of few therapeutic options for treating infections due to multidrug-resistant pathogens, this is a very important issue to solve in this tigecycline meta-analysis. By using the authors' fixed-effects model, the 95% predictive interval for the RR is 0.971–1.641, which predicts that a future comparative study might have an RR <1. However, six studies included in the meta-analysis cannot be predicted from the model presented by the authors [complicated skin and skin structure infections (n=2), complicated intra-abdominal infections, diabetic foot infection with osteomyelitis, community-acquired pneumonia and methicillin-resistant *Staphylococcus aureus* infections]. This clearly indicates the inconsistency between the data and the model used for the meta-analysis.

In summary, the main conclusion presented by the authors that the overall mortality was higher with tigecycline compared with other regimens is, at least, misleading.

A suitable statistical analysis, which accounts for the complexity of the clinical evidence, should be presented for application of the published results in clinical practice.

Acknowledgements

We thank the anonymous reviewers, who provided valuable comments and suggestions. P.E.V. is grateful to the German Research Foundation (DFG Oh 39/11-1) for support.

Transparency declarations

D. C. is an adviser of Pfizer (formerly Wyeth) Laboratories Argentina for antibiotics and he has participated in several experimental and observational studies with tigecycline (Tygacil®). P. E. V.: none to declare.

References

1 Yahav D, Lador A, Paul M *et al*. Efficacy and safety of tigecycline: a systematic review and meta-analysis. *J Antimicrob Chemother* 2011; **66**: 1963–71.

2 Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. Chichester: John Wiley & Sons, Ltd, 2004.

3 Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998; **17**: 841–56.

4 Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

5 Sharp SJ, Thompson SG. Analyzing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Stat Med* 2000; **19**: 3251–74.

6 Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 137–59.

Clinical Infectious Diseases Advance Access published May 18, 2012

Correspondence

Imbalanced Mortality Evidence for Tigecycline: 2011, the Year of the Meta-analysis

To the Editor—During 2011, 3 high-quality meta-analyses investigated, among other clinical questions, the difference between mortality rates for any cause of 30-day mortality for tigecycline and any other antibiotic [1–3]. Although the meta-analyses of Cai et al [1] and Tasina et al [2] concluded that the evidence of mortality differences should be considered nonconclusive, Yahav et al [3] boldly concluded that the differences between mortality rates were statistically significant. Given that these meta-analyses included almost the same published studies, it is worth asking why they arrived at different conclusions regarding mortality differences.

The source of disagreement between authors is the statistical model they applied. Although the empirical evidence (ie, published studies) can be declared correct, the statistical model is always "wrong." The statistical model is wrong in the sense of its limitations to describe the complexity of the problem at hand. How can we trust the conclusions of a wrong statistical model? A wrong statistical model upgrades to become useful if it is able to predict the published studies included in the systematic review. For example, in a recent publication [4] we pointed out that the model used by Yahav et al was not able to predict 6 of 14 studies, so a bold conclusion from this statistical model is definitively misleading.

Another way to interpret the evidence of mortality in these meta-analyses is by asking whether the extent of difference in mortality rates is related to the underlying risk of the patients in the different trials. If this relationship exists, then it has important implications in the interpretation of the mortality results (eg, by detecting which patients may be at risk under application of tigecycline and which patients may be not).

A natural way to measure underlying risk in a clinical trial is by estimating the mortality rate of the control group. We can assess the relationship between underlying risk and difference in mortality rates using the model proposed by Sharp and Thompson [5]. Figure 1 shows the regression line that summarizes this relationship for the tigecycline meta-analysis published by Yahav et al [3]. The negative slope of this line indicates a decrease in mortality differences as the mortality in the comparator group increases. Moreover, the predictive confidence bounds include all studies used in the meta-analysis, which gives confidence in the conclusions coming from this model.

The message of Figure 1 is that we cannot make a general statement about mortality differences for tigecycline; it may depend on the underlying mortality of the study population. For populations with low mortality rates (left hand side of Figure 1), the model favors the comparator drug, whereas for populations with increased mortality rates, there is no differences between the groups' mortality rates.

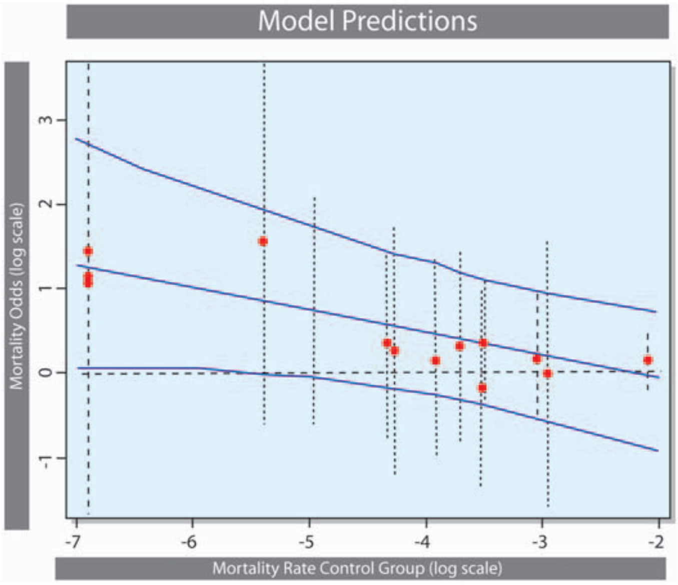


Figure 1. Evidence of mortality in tigecycline compared with any other antibiotic. Each vertical line represents a study result. The points correspond to the observed odds ratio of mortality, and the dashed lines represent the 95% confidence intervals. The meta-regression line is presented by the solid line in the center with its 95% predictive interval.

We hope that our view will motivate readers to think more critically about meta-analysis results in general.

Notes

Financial support. This work was supported by the German Research Foundation (DFG Oh 39/11-1 to P. E. V.).

Potential conflicts of interest. D. C. is an adviser of Pfizer (formerly Wyeth) Laboratories Argentina for antibiotics, and he has participated in several experimental and observational studies with tigecycline (Tygacil). P. E. V. certifies no potential conflicts of interest.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Pablo E. Verde[1] and Daniel Curcio[2]

[1]University of Düsseldorf, Germany; and [2]Instituto Sacre Couer, Infectología Institucional SRL, Buenos Aires, Argentina

References

1. Cai Y, Wang R, Liang B, Bai N, Liu Y. Systematic review and meta-analysis of the effectiveness and safety of tigecycline for treatment of infectious disease. Antimicrob Agents Chemother **2011**; 55:1162–72.
2. Tasina E, Haidich AB, Kokkali S, et al. Efficacy and safety of tigecycline for the treatment of infectious diseases: a meta-analysis. Lancet Infect Dis **2011**; 11:834–44.
3. Yahav D, Lador A, Paul M, et al. Efficacy and safety of tigecycline: a systematic review and meta-analysis. J Antimicrob Chemother **2011**; 66:1963–71.
4. Curcio D, Verde PE. Comment on: Efficacy and safety of tigecycline: a systematic review and meta-analysis. J Antimicrob Chemother **2011**; 66:2893–5.
5. Sharp SJ, Thompson SG. Analyzing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. Stat Med **2000**; 19:3251–74.

Correspondence: Pablo E. Verde, PhD, University of Düsseldorf, Coordination Centre for Clinical Trials, Moorenstr 5, 40225, Duesseldorf, Germany (pabloemilio.verde@uni-duesseldorf.de).

Clinical Infectious Diseases

DOI: 10.1093/cid/cis397

Chapter 3

Review on Generalized Evidence Synthesis

”What a long, strange trip it’s been.”

-Grateful Dead Sentiment, 1977.

Contributions of the authors:

PEV designed and wrote the paper, made the methodological systematic review and developed the new classification of statistical method. CO proposed the research work, performed the systematic review of the clinical applications and provided several inputs into the discussion.

Original Article

Research Synthesis Methods

Received 19 September 2013, Revised 11 April 2014, Accepted 21 April 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1122

Combining randomized and non-randomized evidence in clinical research: a review of methods and applications

Pablo E. Verde* and Christian Ohmann

Researchers may have multiple motivations for combining disparate pieces of evidence in a meta-analysis, such as generalizing experimental results or increasing the power to detect an effect that a single study is not able to detect. However, while in meta-analysis, the main question may be simple, the structure of evidence available to answer it may be complex. As a consequence, combining disparate pieces of evidence becomes a challenge. In this review, we cover statistical methods that have been used for the evidence-synthesis of different study types with the same outcome and similar interventions. For the methodological review, a literature retrieval in the area of generalized evidence-synthesis was performed, and publications were identified, assessed, grouped and classified. Furthermore real applications of these methods in medicine were identified and described. For these approaches, 39 real clinical applications could be identified. A new classification of methods is provided, which takes into account: the inferential approach, the bias modeling, the hierarchical structure, and the use of graphical modeling. We conclude with a discussion of pros and cons of our approach and give some practical advice. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: observational studies; randomized control trials; bias modeling; network meta-analysis; cross-design synthesis; generalized evidence synthesis; hierarchical Bayesian models

1. Introduction

Statistical evidence synthesis is a branch of statistical methods that allows researchers to combine scientific results from multiple pieces of evidence into a single analysis. These techniques are used to extend the scope of a single experiment, by combining results from several experiments. A typical application is the meta-analysis of experiments addressing the same primary research question and using the same statistical design, where the application of simple statistical procedures (e.g., random-effects meta-analysis) is sufficient.

However, while in a meta-analysis, the clinical question may be simple, (e.g., what is the effect of an intervention in a population of interest?), the structure of evidence available to answer it may be complex (e.g., the published results may have different grades of quality), as a consequence, combining disparate pieces of evidence becomes a challenge.

In this review, we cover statistical methods that have been used for the evidence-synthesis of *different study types* with *the same outcome* and *similar interventions*. The study types considered are randomized controlled trials (RCTs) and non-randomized studies, covering studies with non-randomized control groups and studies without a control group (e.g., register and cohort study) (Deeks *et al*., 2003). The methods reviewed are methods used for combining aggregated data and used for combining aggregated with individual data as well.

The main reason for the aforementioned restrictions is the increasing complexity and quantity of methodological research in evidence-synthesis, which requires focusing on a methodological review. However, there is a strong need for such a specific review; for example, to answer the relevant question of how to generalize results from RCTs to clinical practice, that is, how much of the proved *efficacy* can be translated into *effectiveness*.

Coordination Center for Clinical Trials, University of Duesseldorf, Germany

**Correspondence to: Pablo E. Verde, Coordination Center for Clinical Trials University of Duesseldorf, Germany.
E-mail: pabloemilio.verde@uni-duesseldorf.de*

Moreover, researchers may have multiple motivations for combining different study types in a meta-analysis, for example:

- To increase the power to detect an effect that a single source of data is not able to detect.
- To reconstruct evidence that is not directly observable in a single study.
- To learn from the evidence how to improve the statistical design of future studies.
- To make decisions in situations where further experimentation may not be helpful, could not be ethical, or may not be feasible due to time or budget constraints.

However, no study type is free of bias, and the resulting analysis will be a trade-off between extending the inferential scope of a meta-analysis and adjusting the bias that is introduced by combining different study types.

The interest of including non-randomized studies in evidence synthesis has recently been highlighted in a special issue of this journal, where the authors presented the outcomes of a special workshop led by the Non-Randomized Studies Methods Group of the Cochrane Collaboration (Reeves *et al*., 2013). Four discussion papers covered the following: issues in study design and risk of bias by Higgins *et al*. (2013), issues relating to confounding factors when including non-randomized evidence by Valentine and Thompson (2013), issues in selective reporting by Norris *et al*. (2013), applicability of non-randomized evidence as complementary source of evidence by Schünemann *et al*. (2013), and a guideline of checklists for review authors by Wells *et al*. (2013).

Much has been written in evidence synthesis and meta-analysis from many perspectives. An early review of Bayesian meta-analysis methods in tutorial style is presented by Sutton and Abrams (2001). Probably, the most complete review in multi-parameter evidence synthesis is given by Ades and Sutton (2006). Sutton and Higgins (2008) presented an extensive review of methodological developments in meta-analysis. The paper of Higgins *et al*. (2009) concentrates on issues and applications of random-effects meta-analysis. Ioannidis (2010) reviews issues in meta-analysis from the practitioner’s point of view.

This review updates previous methodological reviews (Sutton and Abrams, 2001; Ades and Sutton, 2006; Sutton and Higgins, 2008) in specific topics and includes new methods that were not developed at that time. Furthermore, a new classification of methods was developed and, for the first time, real medical applications of the methods assessed.

We omitted the highly important topic of publication bias, which addresses the problem that studies which claim statistically significant results are more likely to be published than studies with inconclusive results. Useful literature relating to this topic includes the following: Sutton *et al*. (2000), Rothstein *et al*. (2005) and the recent work of Copas (2013).

This paper is organized as follows: Section 2 describes the searching and classification techniques used to identify methodological work and their applications. Results of the methodological work are organized in chronological order and we provide an annotated description of the methods and their applications. Section 3 presents our results, and Section 4 provides a general discussion with some recommendations for practitioners.

2. Methods

2.1. Identification and classification of methodological work

Methodological papers have been previously identified in general reviews such as (Ades and Sutton, 2006) and (Sutton and Higgins, 2008). We use these reviews as a starting point to update the main methodological work and group them into methods, which investigated the combination of different study types in meta-analysis.

A manual search was performed by carefully looking at cross-references and in main applied statistical journals with a focus on applications in life sciences and medicine. Those included the following: Biometrics, Biometrical Journal, Biostatistics, Journal of the Royal Statistical Society series A and C, Research Synthesis Methods, Statistics in Medicine, and Statistical Methods in Medical Research. We also included main methodological journals, which publish applied work, those are the following: Annals of Applied Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society series B, and Statistical Science.

For historical reasons, we start by presenting the Confidence Profile Method (CPM) in Section 3.1. Network meta-analysis is the topic of Section 3.2. The cross-design synthesis (CDS) and related approaches are covered in Section 3.3. Bias modeling of different study types is covered in Section 3.4, and the state of the art of Bayesian hierarchical models (BMHs) is presented in Section 3.5. In each section, we add a subsection with applications in clinical context. Section 3.6 summarizes the inferential approaches and operational characteristics of the statistical methods reviewed.

2.2. Identification of applications in clinical context

Clinical applications were identified from PubMed and within the Web of Science (Version 5.10), by using the following strategy:

- First, we select a key methodological paper in which the proposed method has been originally presented.
- Second, citations of the key methodological paper were identified and classified as follows:

- Methodological reference: In this case, the method is cited in a methodological or discussion context, where the method itself is not applied, but it is used as a reference for methodological extensions or discussion.
- Application in methodological context: This type of application is used for demonstration in a real data problem and used for methodological motivation or to highlight potential benefits in clinical use.
- Application in clinical context: Examples of clinical applications include the use of a method to provide scientific evidence for a clinical problem, the development of guidelines, or systematic reviews.

The citations databases used, with default starting date, were the following: Science Citation Index Expanded – 1945–present, Social Sciences Citation Index – 1956–present, Arts and Humanities Citation Index – 1975–present, Conference Proceedings Citation Index-Science – 1990–present, Conference Proceedings Citation Index-Social Sciences and Humanities – 1990–present, Book Citation Indexâ€“ Science – 2005–present, Book Citation Indexâ €“ Social Sciences & Humanities – 2005–present.

In addition, a PubMed search was performed with different search patterns: ‘Confidence Profile Methods’, ‘Network meta-analysis’, ‘Cross-design Synthesis’, ‘Bayesian hierarchical model’; in combination with ‘different study design’ or ‘meta-analysis’.

3. Results

3.1. *The Confidence Profile Method*

The CPM was introduced by Eddy (1989) as a general statistical framework to combine multiple sources of information in evidence synthesis and further described in a series of tutorial articles (Eddy, *et al.*, 1992; Eddy, 1989; Eddy *et al.*, 1990b; Eddy *et al.*, 1990a; Shachter *et al.*, 1990) and in a book with numerous examples (Eddy *et al.* (1992). The CPM was proposed under the realistic assumption that the empirical evidence used in meta-analysis could be incomplete, indirect, and biased.

The CPM was a vanguard approach, and it has influenced further developments over the last decades, including indirect treatment comparisons and network meta-analysis (NMA) (Section 3.2), direct bias modeling (Section 3.4) and the use of Bayesian graphical models in evidence syntheses (Ades, 2003; Spiegelhalter *et al.*, 2004; Ades and Sutton, 2006) (Section 3.5).

Several important aspects have been introduced in the CPM framework, and we can highlight the following:

- First, the evidence to be analyzed in a systematic review is not considered as a realization of a random sample. As a consequence, statistical techniques with roots in the analysis of a single experimental data could lead to misleading results.
- Second, the problem of analysis of clinical evidence is embedded in a formal probability model with a Bayesian network representation. That allows a pictorial representation of the pieces of evidence, parameters of interest, functional parameters, and bias modeling.
- Third, the analysis of evidence is explicitly subjective. The analyst has to formalize his/her current state of knowledge of the problem at hand and include this aspect into the statistical model. Although statistical computations of the CPM can be done with direct use of the likelihood function or using Bayesian techniques, the interpretation is always subjective.
- The CPM emphasizes a case-specific modeling approach, where variability and bias of multiple sources of evidences have to be assembled in a single model. That contrasts with the statistical procedural approach, such as meta-analysis using fixed or random effects, where one approach applies to every situation.

3.1.1. Type of evidence, bias modeling, and inference. The CPM classified different *types of evidence*, where the *type of evidence* defines the likelihood function for interpretation of experimental results at *face value*. The main source of classification is the experimental design ((Eddy *et al.*, 1992), Chapter 5).

Given that no experimental design is free of bias, Eddy *et al.* (1992) [p. 66–68] classified the propensity of bias of different experimental designs, with two main types of bias:

- *Bias to internal validity*, which is composed of factors that cause the observed results to not reflect the effects of the intervention in the circumstances of investigation. Typical examples of these factors are confounding variables, loss to follow-up, patient-selection bias, and dilution bias.
- *Bias to external validity and comparability*, which are composed of factors that make differences between the circumstances of investigation and the circumstances of interest. Examples of bias to external validity are population bias and intensity bias.

The CPM was not a BHM like those reviewed in Section 3.5. Statistical inference was carried out by direct application of Bayesian methods, that is, by multiplying the likelihood functions of the model parameters by their priors. Multiple parameters were assumed independent a priori and conjugate or Jeffrey’s priors were used for these model parameters. The method allowed to calculate posteriors of functional parameters, for example, parameters of interest after adjusting by bias modeling. Computations were based on Normal approximations of the posterior distribution and were implemented in the book's companion software (FAST*PRO) of Eddy *et al*. (1992).

As a general statistical framework, the CPM was perfectly suited to modern Bayesian computation techniques and software, but it was developed prior to the Markov Chain Monte Carlo (MCMC) revolution in statistics. Clearly, the Bayesian graphical approach was one of the more complex parts of the methodology. Although the diagrams were usually simple in their final form, they were not easy to develop unless the practitioner was skilled in structuring conditional independence statements between model quantities. Probably, these issues have restricted its application. Spiegelhalter *et al*. (2004) (Chapter 8) showed straightforward implementation of CPM's ideas with BUGS software (Lunn *et al*., 2009) including Bayesian graphical models and computations using MCMC.

3.1.2. Applications in clinical context. Eighty-five citations of two key methodological papers from Eddy and another 11 references were identified in PubMed. These papers were evaluated with respect to clinical applications:

1. Web of Science: Citations of Eddy (1989) ($n = 45$)
2. Web of Science: Citations of Eddy *et al*. (1990a) ($n = 40$)
3. PubMed: Search pattern 'Confidence Profile Method' ($n = 11$)

3.1.2.1. Guidelines. The CPM was systematically applied in a series of clinical guidelines developed by the American Urological Association and published between 1987 and 2007 in clinical urological journals. These guidelines cover the management of invasive bladder cancer (Eddy, 1989; Smith *et al*., 1999; Hall *et al*., 2007), ureteral calculi (Segura *et al*., 1997), female stress urinary incontinence (Leach *et al*., 1997), organic erectile dysfunction (Montague *et al*., 1996), prostate cancer (Austenfeld *et al*., 1994), and staghorn calculi (Segura *et al*., 1994).

In these guidelines, CPM was used for evidence combination, including meta-analysis of comparable RCTs, of individual arms of RCTs and of individual arms from all studies regardless of study design. The analyses were performed with the Fast*Pro software (Eddy *et al*., 1992).

Two publications describe a guideline for detecting development dysplasia of the hip in children, published in 2000 (Lehmann *et al*., 2000; Pediatrics, 2000). The method used a combination of expert panel, decision modeling, and evidence synthesis. Summarizing evidence was performed across probabilities by the CPM. The calculation was done with the BUGS software (Lunn *et al*., 2009).

3.1.2.2. Meta-analyses. In 2009, a meta-analysis on ovarian preservation during chemotherapy was published in the Journal of Women's Health (Clowse *et al*., 2009). Two systematic reviews using the CPM to combine evidence of RCTs and cohort studies were published in 2003 in the Journal of Hepatology, one dealing with acute hepatitis C (Licata *et al*., 2003) and the other with chronic hepatitis B (Craxi *et al*., 2003). The probability of sudden death from rupture of intracranial aneurysms was calculated in a meta-analysis and published in 2002 in the Journal Neurosurgery (Huang and van Gelder, 2002). In 1999, several meta-analyses were published, with meta-analytic techniques based on the CPM and using the FAST*PRO software. These studies covered hormone replacement therapy and the risk of colon cancer (Obstetrics and Gynecology, (Nanda *et al*., 1999), treatment of chronic hepatitis C (American Journal of Gastroenterology, (Leach *et al*., 1997) and Journal of Hepatology, (Craxi *et al*., 1999) and prophylactic auxiliary node dissection on breast cancer survival (Annals of Surgical Oncology, (Orr, 1999). Three applications of the CPM are related to meta-analysis in cardiology, one investigating predictors of adverse outcome after coronary interventions (Journal of American College of Cardiology, 1998, (Block *et al*., 1998) and two dedicated to risk stratification after myocardial infarction (Annals of Internal Medicine, 1997, (Peterson *et al*., 1997) American Journal of Cardiology, (Shaw *et al*., 1996).

In another application, CPM was used to derive a summary estimate of relative risk of future fractures from different study types, such as prospective cohort, case-control, and cross-sectional studies (Klotzbuecher *et al*., 2000). A systematic review of efficacy of ketogenic diet for the treatment of refractory epilepsy in children combining uncontrolled retrospective and prospective studies was performed with FAST*PRO (Lefevre and Aronson, 2000). The effect of spinal manipulation on patient's pain and functional outcomes in low back pain was assessed by combining data from 25 controlled trials (Shekelle *et al*., 1992). CPM was used to combine data from uncontrolled non-randomized trials into single best estimates of outcome of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower extremity ischemia (Adar *et al*., 1989). Two meta-analyses, with a reference to CPM but combining only RCTs were performed, one dealing with antibiotics in tube thoracostoma (Evans *et al*., 1995) and the other with manipulation and mobilization of the cervical spine (Hurwitz *et al*., 1996).

3.2. Network meta-analysis

Network meta-analysis (NMA) is a new area in evidence synthesis, where the aim is to combine data from studies reporting randomized results of several treatments to not only make pairwise treatment comparisons, but to also reconstruct comparisons that have not been performed head-to-head in any study before.

Increasing interest of practitioners in this area has been recently surveyed by Abdelhamid *et al*. (2012). They reported that '…many reviewers (76%) accepted that indirect evidence is needed as it may be the only source of information for relative effectiveness of competing interventions, provided that review authors and readers are conscious of its limitations'.

Network meta-analysis has its roots in the Eddy's CPM ((Eddy *et al*., 1992), p. 45), where disparate pieces of evidence are combined to reconstruct evidence, which is not directly observable. Early applications of indirect treatment comparison can be found in Higgins and Whitehead, 1996; Hasselblad, 1998; Dominici *et al*., 1999 and Ades, 2003 (see Section 4).

Meta-analysis, which combines results from a mixture of randomized treatment comparisons, has been called in different ways in the statistical literature: *mixed treatment comparisons* (Lu and Ades, 2004), *network meta-analysis* (Lumley, 2002), and *multiple-treatment meta-analysis* (Salanti *et al*., 2008). Statistical methods are based on the use of generalized linear modeling framework from the Bayesian (Dias *et al*., 2013) and classical perspective (Lu *et al*., 2012; Piepho *et al*., 2012). White *et al*. (2012) showed that NMA models can be estimated by expressing them as multivariate random-effect meta-regression models. Meta-analysis of aggregated and patient individual data has been investigated by exploring treatment by patient-level covariates interactions (Donegan *et al*., 2012; Donegan *et al*., 2013).

At first sight, statistical methods in NMA might be similar to the classical topic of incomplete block designs ((Hinkelmann and Kempthorne, 1994), sec. 9.8), where the number of experimental units in a block is smaller than the number of treatments. However, as pointed out by (Senn *et al*., 2011), while the randomization of treatments in incomplete block designs might be performed within and between blocks and the experimenter controls the distribution of treatments per block, in NMA randomization is only performed within the trial and experimenters do not have any control on the number of treatments per study. These issues make difficult to justify a valid measure of treatment effects at both the level of the study and across studies. Moreover, as usual in meta-analysis, trials are performed by different investigators on different patients and with different protocols. As a consequence, variability of treatment effects might be very different within and between trials. Therefore, modeling between-trial heterogeneity is not straightforward (Lu and Ades, 2009) and remains a modeling issue (Thorlund *et al*., 2013).

Another important issue that might arise in NMA is the lack of agreement between direct treatment comparison and evidence of indirect comparison. This type of conflict of evidence results when treatment differences vary between types of trials. Lumley (2002) called this issue *incoherence* while Lu and Ades (2006) called it *inconsistency*. Different statistical techniques have been proposed to detect and model inconsistency in NMA. Lu and Ades (2006) proposed a factor that measures inconsistency between treatment comparisons, while Dias *et al*. (2010) proposed a node-splitting algorithm to test inconsistency. For multi-arm NMA, Higgins *et al*. (2012) distinguish between two types of inconsistencies: *loop inconsistencies* and *design inconsistencies*. Loop inconsistencies are regarded as a special type of *between-studies heterogeneity* that might affect the magnitude of treatment effect. For example, studies of different comparisons were undertaken in different settings or contexts, and these differences are associated with the magnitude of treatment effect. Design inconsistencies are regarded as a study-level covariate that modifies the effect sizes *within the study*. They proposed an approach to identify inconsistencies by including a full set of design-by-treatment interaction terms in an NMA model. This model handles simultaneously design and loop inconsistencies.

Although, statistical methods of NMA of RCTs are an active area of research, the combination of studies with randomized and non-randomized evidence is a new area of research. We found two recent works on NMA and different study types: Schmitz *et al*., (2013) and Soares *et al*., (2014).

Schmitz *et al*. (2013) proposed three alternative approaches of combining data from different trial designs in NMA: a simple combination of study's data by ignoring the different design types; the usage of observational data as prior information to adjust for bias due to trial design; and a three-level hierarchical model to account for heterogeneity between-trial design. The first approach is used to analyze inconsistencies between direct and indirect treatment comparison. The second one is used to understand the bias that observational data may introduce into the analysis. This is performed with a prior to posterior sensitivity analysis. The third approach, the three-level hierarchical model, is used to combine different study types and to provide overall estimates after accounting for between-study type variability. This model is an application of the *grouped random-effects approach* that is reviewed in Section 3.5.1.

Soares *et al*. (2014) developed a hierarchical Bayesian model to include randomized and non-randomized studies in a NMA. Observational studies are used to explore modeling assumptions in evidence synthesis in the presence of sparse data.

3.2.1. Applications in clinical context.

1. Web of Science: Citation of Schmitz *et al*. (2013) ($n = 1$)
2. Web of Science: Citations of Soares *et al*. (2014) ($n = 0$)
3. PubMed: Search pattern 'network meta-analysis and different study types' ($n = 24$). Two clinical applications were identified.

The work of Schmitz *et al*. (2013) was cited by Mesgarpour *et al*. (2013), who combined 48 studies (34 RCTs and 14 observational) to compare safety of off-label erythropoiesis stimulating agents (ESAs) in critically ill patients. ESAs treatment is compared with other effective interventions, placebo or no treatment by using a three-level hierarchical Bayesian model. The model used by the authors accounted for between-studies variability and between-design variability. In addition, a sensitivity analysis is performed by down-weighting the evidence of observational studies. They also analyzed the robustness of their results by comparing results from models that included all studies, with the results from models that excluded studies with high risk of bias or low quality. The authors concluded that there was no statistical evidence of increase of risk in ill patients treated with ESAs.

Bittl *et al*. (2013) performed a Bayesian cross-design and NMA of 12 studies (four randomized clinical trials and eight observational studies) comparing coronary artery bypass graft with percutaneous coronary intervention and seven studies (two randomized clinical trials and five observational studies) coronary artery bypass graft with medical therapy. Based on an NMA, they arrived to the conclusion that medical therapy is associated with higher 1-year mortality than with the use of percutaneous coronary intervention for patients with unprotected left main coronary artery disease (odds ratio, 3.22; 95% credibility interval, 1.96–5.30).

Jones *et al*. (2013) made a systematic review to compare effectiveness of antiplatelet therapy, medical therapy, exercise, and endovascular and surgical revascularization in patients with peripheral artery disease. A meta-analysis of direct comparison was supplemented with an NMA. Evidences were available from 83 RCTs and four observational studies.

3.3. *Cross-design synthesis*

The CDS was a method designed in 1992 by the US General Accounting Office to combine experimental and non-experimental data. The method is described in Droitcour *et al*. (1993) and in Chelimsky *et al*. (1993).

Cross-design synthesis was developed to adjust the typical patients selection bias of the RCTs and generalize their results to populations that have not been included in RCT experimentation. With this end in mind, historical information coming from registers should be combined with RCT's results. Under the CDS's paradigm, experimental and non-experimental data are viewed as complementary sources of information.

The typical application of the CDS is called the *empty cell problem*, where the results of the RCTs should be extrapolated to a subgroup population where the data are only available in the register. Basically, the logic behind the CDS is a step-wise strategy for evidence synthesis:

1. Assemble the literature on the effectiveness of an intervention and the individual patients data of subgroups of interest.
2. Determine whether bias is relevant in individual studies through expert review and then adjust for this bias (e.g., by using the CPM). Adjust the bias of individual data by covariates adjustment, standardization, propensity scores (Agostino, 1998), etc.
3. Combine the experimental and adjusted non-experimental evidence by assuming that clinical effects are proportional between subgroups.

The reliability of the CDS was criticized in the Lancet by Anonymous (1992) and by Begg (1992) who pointed out that the authors have underestimated the problem of harmonizing results from RCTs and medical databases.

A modern view of the CDS was recently given by Kaizar (2011), where the statistical framework proposed by Imai *et al*. (2008) is used to evaluate the statistical properties of a CDS estimator. Kaizar (2011) evaluated the CDS with an extensive computer simulation experiment and used a real case example regarding the effectiveness of insulin pumps versus glargine insulin injections in the regulation of blood glucose in adolescents with type 1 diabetes. (Kaizar, 2011) concludes that under reasonable data assumptions, the simple CDS estimator has smaller bias and better coverage than commonly used estimates based on randomized or observational studies alone.

Another topic directly related to the aims of the CDS is the assessment of effectiveness, that is, the generalization of RCTs results to clinical practice. RCTs provide the gold standard for proving efficacy of interventions. The reason is high internal validity, allowing causal reasoning. Often, RCTs are performed with highly selected patient populations, excluding women, children, elderly, and patients with comorbidity. As a consequence, generalizability of results from RCTs to these patients is severely limited, and different types of bias, such as selection bias, may occur. In addition, adequate information about the recruitment process is often not provided, making an assessment of generalizability to clinical practice difficult. Recent work in this area is presented by Benson and Hartz (2000), Zimmerman *et al*. (2004), Fortin *et al*. (2006), Prentice *et al*. (2006), Greenhouse *et al*. (2008), Ahern *et al*. (2009), Frangakis (2009), and Cole and Stuart (2010).

Some developments in BHMs have been motivated by the CDS method, for example, those by Nixon and Duffy (2002), Prevost *et al*. (2000), and Peters *et al*. (2005). We review these methods in Section 3.5.

3.3.1. *Applications in clinical context.*

1. Web of Science: Citation of (Droitcour *et al*., 1993) ($n = 9$)
2. Web of Science: Citations of (Chelimsky *et al*., 1993) ($n = 3$)
3. PubMed: Search pattern 'cross-design synthesis' ($n = 9$)

The CDS method was applied by the US General Accounting Office in 1994 to study the effect of 'Breast conservation versus mastectomy: patient survival in day-to-day medical practice in randomized studies'. No other clinical application was found.

3.4. Direct modeling of bias

3.4.1. Classical meta-analysis techniques. Meta-analysis with historical controls is analyzed by Begg and Pilote (1991). A random-effects meta-analysis model is presented in which the baseline effect in each study is random, but the treatment effect is constant. With this model, the appropriate contribution of historical studies can be determined. The method is Bayesian in nature, but estimation of hyper-parameters is performed by Empirical Bayesian techniques. The authors illustrated this method by combining four RCTs with 12 uncontrolled studies to analyze the efficacy of bone-marrow transplantation versus conventional chemotherapy in the treatment of acute non-lymphocytic leukemia. Li and Begg (1994) presented a non-iterative estimator of treatment effects based on this method. They studied theoretical properties and presented results from a simulation experiment which contemplate different random-effects distributions (normal, log-normal, exponential, and uniform). They concluded that both the pooled effects and the between-studies estimators are strongly consistent with desirable heuristic properties.

An early work in combining disparate study designs is presented by Brumback *et al*. (1999). They present a meta-analysis where three case-control, and 28 cohort studies are combined to study the association of prenatal testing via chorionic villus sampling with the occurrence of terminal limb defects. The authors combine two types of sub-models in a single meta-analysis: a fixed effect sub-model with a logistic-regression is used to model the evidence of case-control studies and a random-effects with a Poisson regression with a conjugate Gamma distribution is used to model the evidence of the cohort studies. Inference on the pooled effect parameter is estimated by combining the likelihood of the fixed effect, and the marginal likelihood of the random-effects model. The resulting likelihood depends on the parameters of the Gamma distribution, the authors presented a sensitivity analysis by estimating the pooled effect for different values of these parameters.

3.4.1.1. Applications in clinical context. 1. Web of Science: citations of Begg and Pilote (1991) ($n = 25$). In Web of Science, we found 25 methodological citations of Begg and Pilote (1991). No clinical applications were found using the random-effects model of Begg and Pilote (1991).

3.4.2. Adjustment of likelihoods for study design and quality. Wolpert and Mengersen (2004) presented reductionist and alternative method to the CPM. While CPM constructs a global probabilistic model by using conditional independence between model parameters and pieces of evidence, Wolpert and Mengersen (2004) proposed to directly adjust the likelihood of each study's parameter for its potential bias. The adjustment is done by defining a bias function similarly to the CPM. In a second step, the adjusted likelihoods are combined by a meta-analysis model (e.g., a random-effect model). Computations were based on MCMC, and the authors highlight potential advantages of the method, such as inference of functional parameters and ranking parameters.

They apply this technique to combine case-control studies with cohort studies in order to assess the relationship between environmental exposure to tobacco smoke and lung cancer. The likelihood of each study is adjusted by the propensity that each design has with respect to different types of misclassifications, which includes the following: the bias introduced from the misclassification of people who always smoked to people who have never smoked, bias of misclassification of disease and non-disease, and misclassification of exposure status.

Multiple bias modeling of meta-analysis of retrospective case-control studies is analyzed by Greenland (2005). He presented a Bayesian modeling approach where priors are used to encapsulate external information of different types of bias. He called this sort of sensitivity analysis a meta-sensitivity modeling. He applied this technique to adjust a meta-analysis of 14 case-control studies (12 published and two unpublished) of residential magnetic fields and childhood leukemia. The sources of bias considered in the meta-sensitivity analysis were the following: confounding factors of field exposure and leukemia, sampling and response bias, and measurement errors in magnetic fields.

Another way to adjust likelihoods in Bayesian modeling is by explicitly discounting for study's quality bias using a 'power prior' (Ibrahim and Chen, 2000). The likelihood of low quality studies is raised to a power factor between 0 and 1, where values close to 0 indicate low quality and values close to 1 no bias. Recently, Neuenschwander *et al*. (2009) proposed to scale the power priors to a 'proper power prior' to estimate the discounting factor.

Turner *et al*. (2009) recognize the practical limitations and difficulties of elicitation of bias, and they introduced a comprehensive approach to adjust a classical meta-analysis for multiple sources of bias. The idea is that the pooled treatment effect of the bias-adjusted meta-analysis will reflect a more realistic estimate than the naive meta-analysis.

In Turner's approach, multiple sources of bias are divided into two main types of bias: internal validity bias and external validity bias. Each study included in the meta-analysis is evaluated by a group of assessors, who estimate different types of biases by a score system. External empirical evidence of bias can be included in the analysis (Welton *et al*., 2009), but the method assumes that, in general, it is unrealistic that such evidence exists.

3.4.2.1. Applications in clinical context. 1. Web of Science: citations of Wolpert and Mengersen (2004) ($n = 14$). We found 14 citations of Wolpert and Mengersen (2004) with three clear related clinical applications:

- Bayesian modeling for direct adjustment of likelihoods was applied to the outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis (O'Rourke *et al*., 2007). In this analysis, the likelihood of the log odds ratio is approximated by a normal distribution, the results from observational case-control studies are adjusted to account for an exaggerated precision of treatment effect and for a systematic bias toward overestimation of treatment effects. The adjustment is based on a fixed value, which reflects that observational studies overestimate treatment effect by a median of 30% (Egger *et al*., 2002).
- A similar approach was presented by O'Rourke *et al*. (2009), where safety and efficacy of IV-TPA for ischaemic stroke were analyzed by a cumulative Bayesian meta-analysis based on a Beta-Binomial model. In this case, results from observational studies are adjusted by overestimation of treatment effect by using a fixed approach.
- Another cumulative meta-analysis with bias adjustment for observational studies is presented by O'Rourke and Walsh (2010). A prior distribution for the OR of dead within 1 year after acute stroke was built using a meta-analysis of 26 RCTs comparing stroke unit care versus alternative models of stroke care. The analysis was performed sequentially by starting with the RCTs prior, data from individual observational studies were used to sequentially update outcome knowledge. Again, the likelihood of each observational study was adjusted for overestimation by using a fix value, which reflects that observational studies overestimate treatment effect by a median of 30% (Egger *et al*., 2002).

The assertion in the examples seems to be that the non-randomized studies overestimate the effect size, whereas Deeks *et al*. (2003) clearly demonstrated that non-random allocation can lead to overestimation or underestimation of treatment effects.

Recently, Turner *et al*. (2012) applied their method to adjust a meta-analysis, which included 10 studies comparing routine antenatal anti-D prophylaxis to control. After adjustment for differences in study design and quality, the authors concluded that there is strong evidence in the benefit of routine antenatal anti-D prophylaxis.

3.5. Bayesian hierarchical methods

Bayesian hierarchical modeling techniques have been used to combine studies with different designs during the last two decades. In this section, we consider full BHMs where uncertainty of the hyper parameters are included into the model and where computations based on MCMC are used to estimate posteriors of all parameters in the model in a single modeling step.

3.5.1. The grouped random-effects approach. Combining dissimilar studies in a common meta-analysis was criticized by Larose and Dey (1997). They proposed to group studies with different designs in a common BHM, where each group has its own treatment effect and dispersion parameter. They called this approach 'the grouped random-effects' model. They illustrated their technique with a meta-analysis which combined six single-blind RCTs with nine double-blind RCTs in the study of efficacy of an anti-epileptic drug, progabide. The model is a binomial-normal BHM where a careful sensitivity analysis of the hyper-priors is analyzed. The authors presented four non-informative models for the hyper-priors. Computations were implemented by using Gibbs sampling and the Metropolis method. They concluded that results were insensitive to the hyper-priors specification. Interesting results of this analysis were that open studies were systematically more dispersed than closed studies, and open studies supported the efficacy of progabide, closed studies supported the reverse hypothesis, while the union of the groups supported neither hypothesis. That was a clear warning for meta-analyses that indiscriminately combined studies with different designs.

Prevost *et al*. (2000) presented the first formal Bayesian approach to the cross-design synthesis problem. They propose a three-level hierarchical model, where the first and the second level are used to model the observed evidence and the variability between studies, respectively. A third level is used to model the variability between-study types. This model allows the exchange of information across the study types, with the additional advantages that neither assumes independence between effects in different study types nor equivalence of such effects. In addition, the authors describe a posterior predictive analysis to the 'empty cell' problem, where results of a new RCT or a non-randomized study are predicted from the model. The model is illustrated by combining evidence of RCTs and non-randomized studies, which describe the benefit, in terms of mortality reduction, of using mammography screening in breast cancer for different age groups of women.

Prevost *et al*. (2000) described carefully how priors for hyper-parameters were chosen, and they presented a sensitivity analysis for the priors specification. They concluded that the variability between-study types has the greatest effect on both, the estimate of the overall pooled effect, and the pooled effects within each type of studies.

Another Bayesian development of the cross-design synthesis is presented by Peters *et al*. (2005). The model is motivated by a toxicological application, which investigated the association between exposure to trihalomethanes in drinking water and low birth weight. The available evidence included the study-specific dose-response slope from studies across two disciplines: epidemiological studies with evidence of humans and toxicological studies with evidence of animals. A three-level BHM is developed to account for study type effects, which is similar to the model of Prevost *et al*. (2000). The authors presented a detailed sensitivity analysis by using different sets of prior distributions. They arrived at similar conclusions as Prevost *et al*. (2000), where priors on the between-study types variance component had a main influence in the analysis.

3.5.2. The hierarchical regression modeling approach. Another area is the combination of aggregated and patient individual data, where both types of evidence correspond to different study designs. Methods for combining aggregated and individual patient data have been developed recently under the name of *Hierarchical Related Regression* (HRR) modeling (Jackson *et al*., 2006; Jackson *et al*., 2008). The main idea of HRR is the existence of shared parameters between different data sources that justify merging information in a common model. In HRR, there is an explicit use of graphical models to describe: the probabilistic relationship of multiple sources of information, which bias sub-models are introduced and how share parameters are linked to different data types. Computations are usually implemented in WinBUGS or other MCMC software (e.g., OpenBUGS, JAGS). Recent applications and further development of HRR are presented by Molitor *et al*. (2009) and Jackson *et al*. (2009). Riley *et al*. (2008) and Sutton *et al*. (2008) described similar approaches of combining aggregated and individual data in meta-analysis of randomized trials.

McCarron *et al*. (2010) combined RCTs and non-randomized studies to syntheses evidence of studies comparing treatment for abdominal aortic aneurysms. They developed a BHM, where each arm's outcomes are modeled with binomial distributions, and study effects are modeled with a normal distribution in the logistic scale (i.e., $log(p/(1-p))$). Systematic variability between different study types are modeled by adjusting the study effects with a meta-regression model. The authors proposed to adjust differences in patients' characteristics between study arms. For example, if age is used for adjustment at the study level, the difference of age between study arms is used as covariate. The idea behind this type of adjustment comes from the empirical finding of Deeks *et al*. (2003), which describe that non-randomized trials tend to present unbalance in patients' characteristics between studies arms. The authors argued that covariate adjustment using aggregate study values does not account for covariate imbalances between treatment arms. In a complementary work, McCarron *et al*. (2011) presented an exhaustive simulation experiment to validate the idea of adjustment by differences between arms in patient characteristics.

3.5.3. The hierarchical weighting approach for study design and quality. Complex cost-effectiveness modeling is an area where evidence is usually collected from different study types. Spiegelhalter and Best (2003) embed a generalized evidence synthesis model into a cost-effectiveness model to predict costs and benefits of hip prostheses in different age-sex subgroups. They introduced a BHM for generalized evidence synthesis where multiple sources of evidence could be weighted according to their assumed quality. In this model, the study effect is the sum of two random effects: one describing the study's external bias and the other describing the study's internal bias. The marginal variance of study's effect is expressed as the product of study's quality weight and the variance between studies due to external bias. The quality weights are interpreted as the proportion of between-study variability unrelated to internal bias. This strategy avoids the estimation of the second variance component related to internal bias. For the quality weights, the authors proposed to give fixed values. These values can be obtained from external empirical information or by elicitation from expert opinion. In either case, a sensitivity analysis to a range of assumptions about the quality weights can be carried out. An example of combining one RCT, one register and one case series is used to illustrate this technique. A sensitivity analysis for different quality weight values is presented where the evidence of non-randomized evidence is down-weighted in different ways.

Welton *et al*. (2009) presented a BHM to model meta-analysis or RCTs that may present a high risk of bias. In particular, the authors consider RCTs that may be biased by failure to conceal randomized allocation at the time of patient recruitment. The authors developed a mixed effects model where treatment effects are considered as fixed and bias effect as random. One novelty of this work was to inject empirical bias information into the model by using prior distributions that are estimated from a collection of previously published meta-analysis of RCTs. Although this model is developed only for RCTs, it can be directly applied to combine experimental and non-experimental studies, where the last ones are at high risk of bias.

Meta-analysis of diagnostic tests is an area where RCTs are usually combined with observational studies. The main motivation is to assess diagnostic accuracy in populations that are not contemplated in RCTs. This type of meta-analysis required special techniques to model the correlation between test operating characteristics (e.g., sensitivity and specificity). Verde, (2010) developed a BHM, where random effects follow a bi-variate scale mixture distribution. He gave direct interpretation of the scale weights as measures of model's deviations. A systematic increase of dispersion of retrospective studies was modeled by allowing a meta-regression equation to the scale weights. This technique is illustrated in a meta-analysis of 51 studies, which investigate the accuracy of computer tomography in the diagnoses of appendicitis. The model is implemented in the R package *bamdit* (Bayesian meta-analysis of diagnostic test data) (Verde, 2013), which combines R and JAGS (Just another Gibbs sampling) (Plummer, 2003). The use of scale mixture distributions is a potential modeling tool to handle different study types in meta-analysis of efficacy outcomes as well.

3.5.4. Further hierarchical modeling techniques. Dominici *et al.* (1999) combined results of RCTs with heterogeneous designs to analyze the effectiveness of commonly recommended prophylactic treatments for migraine headaches. They developed a complex BHM to handle a diversity of reporting results (some studies reported results in continuous scores, others reported differences between treatments, others dichotomous outcomes) by using a latent variable approach. Studies presented different type of treatments and indirect comparison was also used to assess treatments that were not compared in the same trial. This work is one of the earliest full implementation of ideas coming from Eddy's CPM by using modern Bayesian modeling and computational techniques (e.g., MCMC). A related work on mixed treatment comparisons was presented by Ades (2003), who extended the ideas of Eddy's CPM on 'chain of evidence' to reconstruct treatment comparisons where no direct comparison evidence was available. The work of Dominici *et al.* (1999) and Ades (2003) represent an early development of mixed treatment comparisons in meta-analysis.

While cross-design synthesis refers to the inclusion in a meta-analysis of studies addressing the same question under different designs, Nixon and Duffy (2002) proposed to combine studies addressing different but clinically related questions. They called this procedure 'the cross-issue synthesis', which was another name for the 'chain of evidence' problem.

The authors build a BHM to estimate the effectiveness of tamoxifen in the treatment of breast cancer for women with mutations in the BRCA1 or BRCA2 gene. One factor affecting the effectiveness of tamoxifen is the estrogen-receptor (ER) concentration of the primary tumor. Women with this gene mutation are typically ER negative, so the effectiveness of tamoxifen is affected by this mutation. They estimate the effectiveness of tamoxifen in BRCA by combining three different study types: preventive trials of tamoxifen, studies of adjuvant tamoxifen and studies reporting relationship between ER gene mutations. The authors used the *grouped random-effect* approach to allow different variability parameters for each study type and functional parameters to reconstruct the conditional probabilities needed in the analysis. This analysis was an example of what we can call today “research synthesis for personalized medicine”.

Meta-analysis is usually a two-step analysis: In the first step, individual studies are selected and summarized and in the second step a meta-analysis model is applied (e.g., a random-effects model). This contrasts to BHM where a single step is used to estimate all parameters simultaneously. The BHM approach has the advantage of contemplating all parameters' variability in a single model and it offers great technological flexibility by using MCMC methods. However, there are situations where the two-step approach is useful: when study-specific analyses are too complex, when there are several models or parameters of interest to consider or when the parameters of interest are complex functions of other study parameters. Recently, Lunn *et al.* (2013) presented a new strategy of meta-analysis, where a Bayesian two-step approach is proposed. The idea is to give a full Bayesian analysis at the level of each study and summarize study results by the posteriors resulting from MCMC. In the second step, parameters' posteriors for each study are combined in a global Bayesian meta-analysis model. The authors illustrate this new technique with two examples: one meta-analysis that studies the effect of taking diuretics on the risk of pre-eclampsia during pregnancy and another complex meta-analysis where studies provide longitudinal measures of abdominal aortic aneurysms data together with the occurrence of clinical events. Clearly, this new meta-analysis approach can be directly used for combining studies of different design, for example individual bias modeling can be applied to each study in the first stage and combination of study results in the second one.

3.5.4.1. Applications in clinical context.

1. Web of Science: Citations of Larose and Dey (1997) ($n = 20$).
2. Web of Science: Citations of Dominici *et al.* (1999) ($n = 20$).
3. Web of Science: Citations of Prevost *et al.* (2000) ($n = 44$).
4. PubMed: Search pattern 'Bayesian hierarchical model' in combination with 'different study design' ($n = 8$) or 'meta-analysis' ($n = 21$).

For Larose and Dey (1997) and Dominici *et al.* (1999) no clinical applications were found. From 44 citations of Prevost *et al.* (2000), two clinical applications were identified: one was a Bayesian meta-analysis from Grines *et al.* (2008), which compared short-term mortality estimates from RCTs and non-RCTs in the intervention of acute myocardial infarction using AngioJet thrombectomy to percutaneous coronary intervention alone. The other was a BHM from Sampath *et al.* (2007), which assessed the efficacy of loop diuretics in acute renal failure in a meta-analysis by combining RCTs and non-RCTs.

3.6. General characteristics of evidence synthesis methods

The previous sections were divided by the proportion of influence that classical and Bayesian methods have on the development of methods for combining different study types. However, no matter which statistical school, these methods have a particular characteristic: the necessity of bias modeling between pieces of evidence, which clearly introduce an overlapping area between techniques. The aim of this section is to provide a more general understanding of how those methods overlap in terms of the statistical philosophy and the bias modeling technique.

Table 1 represents a classification of statistical methods used in research synthesis. Methods are characterized according to the following features:

- Statistical inference: A method is classified as *Classical* or *Bayesian*, where *Bayesian* means that prior distributions for all parameters are given. For example, the commonly used random-effects model, where all parameters are estimated from the data (i.e., Empirical-Bayes estimation), is considered as a *Classical* inferential approach.
- Bias modeling: We classified the bias modeling as *Yes*, if *explicit modeling of bias* is used (e.g., quality weighting and likelihood adjustment).
- Hierarchical modeling: This feature is classified as *Yes*, if the method involves hierarchical parameter structures to model multiple sources of evidence.
- DAG: *Yes* means that the method is based on a Directed Acyclic Graph representation. DAGs representations were promoted in the early days by (Eddy, 1989) and it is interesting to assess if this feature has been used.

Starting at the top of Table 1, we have as a reference the most popular meta-analysis methods: the fixed-effects and the random-effects models. These methods are blind to potential bias, if they are used for combining different study types, their results are prone to a multiplicity of bias.

The Eddy's CPM is represented as a hierarchical Bayesian meta-analysis with the possibility of extensive bias modeling. Eddy's method was an attempt to improve the bias issues of fixed and random-effects models. However, it is interesting to note that the clinical applications we found used the CPM as a Bayesian random-effects meta-analysis without bias modeling. In some cases, the authors mentioned that the CPM could adjust the meta-analysis when different study types are combined, but they did not make bias adjustment themselves. Eddy himself laments that the complexity of his method has limited its use among practitioners (Eddy, 2013).

Following our historical approach, the cross-design synthesis is classified as a sort of classical meta-analysis with explicit modeling of bias. The work of Begg and Pilote (1991) as well as Brumback *et al*. (1999) uses classical statistical methods, with explicit bias modeling in the case of Brumback *et al*. (1999).

The rest of the papers in Table 1 clearly show that during the last 15 years, the Bayesian approach has dominated this area of meta-analysis. The *grouped random effects* approaches did not focus on bias modeling but on variability between study types, while the *direct adjustment of likelihoods* (Wolpert and Mengersen, 2004; Greenland, 2005; Turner *et al*., 2009), *the hierarchical regression* (Jackson *et al*., 2006; McCarron *et al*., 2010), and *weighting approaches* (Spiegelhalter and Best, 2003; Welton *et al*., 2009; Verde, 2010) have enforced bias modeling.

Table 1. General characteristics of evidence synthesis methods used to combine different study types.

Main reference/method	Statistical Inference	Bias modeling	Hierarchical	DAG
Fixed effects meta-analysis	Classical	No	No	No
Random-effects meta-analysis	Classical	No	Yes	No
Confidence profile method	Bayesian	Yes	Yes	Yes
Cross-design synthesis	Classical	Yes	No	No
(Begg and Pilote, 1991)	Classical	No	Yes	No
(Brumback *et al*., 1999)	Classical	Yes	Yes	No
(Wolpert and Mengersen, 2004)	Bayesian	Yes	Yes	No
(Greenland, 2005)	Bayesian	Yes	Yes	No
(Turner *et al*., 2009)	Bayesian	Yes	Yes	Yes
(Welton *et al*., 2009)	Bayesian	Yes	Yes	No
(Larose and Dey, 1997)	Bayesian	No	Yes	No
(Prevost *et al*., 2000)	Bayesian	No	Yes	No
(Peters *et al*., 2005)	Bayesian	No	Yes	No
(Jackson *et al*., 2006)	Bayesian	Yes	Yes	Yes
(Riley *et al*., 2008)	Classical	Yes	Yes	No
(Sutton and Higgins, 2008)	Bayesian	Yes	Yes	No
(McCarron *et al*., 2010)	Bayesian	Yes	Yes	No
(Spiegelhalter and Best, 2003)	Bayesian	Yes	Yes	Yes
(Verde, 2010)	Bayesian	Yes	Yes	No
(Dominici *et al*., 1999)	Bayesian	No	Yes	No
(Schmitz *et al*., 2013)	Bayesian	Yes	Yes	No
(Soares *et al*., 2014)	Bayesian	Yes	Yes	No

The aim of this summary is to show the relative influence of Bayesian, frequentist, and bias modeling upon different methods developed in the last two decades.

The recent work in combining randomized and observational studies in NMA can be clearly classified. Schmitz *et al*. (2013) is a three-level BHM based on the *grouped random effects* approach to model between-study type heterogeneity. The work of Soares *et al*. (2014) is a BHM with extensive bias modeling.

Finally, the use of DAGs in evidence synthesis has been sporadic and more related to the use of the statistical software (e.g., WinBUGS).

4. Discussion

4.1. Classification of statistical approaches

This paper aims to give an overview of different modeling techniques that have been developed to combine different study types in meta-analysis. For historical reasons, we started with the Eddy's confidence profile method and continued with the cross-design synthesis, but the classification between bias modeling and Bayesian hierarchical modeling was less clear.

The classification in Section 3.6 was an attempt to clarify the overlapping areas between those sections. Although, this is only a rough classification some clear patterns emerge: Independently from the inferential approach, bias modeling is promoted almost for every model, the increasing applications of hierarchical Bayesian modeling, with the classical techniques that have been wiped off the play field. The trend of case-specific modeling approach as was originally promoted by the CPM.

4.2. Critique of our review and the methodological impact on clinical applications

In this review, clinical applications mean applications of the methodology to improve diagnosis, prognosis, or treatment for a clinical problem. It can be assumed that the majority of serious clinical applications have been published in at least one of the databases covered in our review. Nevertheless, it was difficult to identify and link statistical methods to clinical applications. Our searching strategy for identification has been weak in many aspects including the following points: First, there is not always one methodological paper, which can be clearly defined as the origin source of a method/technology. Second, even if such a paper exists, it may not be cited in a clinical application. Therefore, our approach to look at citations may not find clinical applications. Third, a broad search in PubMed without specification gives too many publications (e.g., BHM: 1149). A more restricted search, for example, 'Bayesian hierarchical model' + 'meta analysis', may again miss clinical applications.

As a consequence, our strategy (citation of key papers and restricted search) may identify only a subset of the clinical applications. Nevertheless, it is a systematic and reproducible strategy and for the review more than 250 publications, which have been identified according to this strategy, have been evaluated, and only 39 clinical applications were found. However, from a pragmatic point of view, we can at least have a rough estimate of the amount of clinical applications.

Taking a historical perspective, the impact of methodological work in clinical applications can be summarized as follows: early ideas of the Eddy's CPM were adopted by research groups and guidelines were developed, but the method did not spread out in practice. Classical approaches like the one of Begg and Pilote (1991) and the cross-design synthesis were not applied in real clinical context. Adjusted likelihood techniques were applied by a research group, but they did not reach general practice. The potential that BHM has in complex meta-analysis modeling has been established with a large amount of examples, and methodological work but expertise required for their applications remains an issue.

4.3. Relationships with network-meta-analysis

Specific to NMA are inconsistencies resulting from differences in treatment effects across direct and indirect comparisons, which may result in bias. However, if there are sources of bias that effect direct comparisons of studies then the pooled results of NMA incorporating different study types (e.g., RCTs and studies with non-randomized control groups) are affected as well. In addition, if estimation of between-study heterogeneity is of major issue in NMA, the inclusion of different study types can challenge practitioners in this problem. Therefore, biases generated by combining different study types are also relevant for NMA. The recently work of Schmitz *et al*. (2013) and Soares *et al*. (2014) are two examples of the recent trend in this new area of research.

One potential advantage of combining observational and RCTs in NMA is that we might have direct treatment comparisons in observational studies that are not represented in the RCTs. Making available direct comparisons form observational studies might reduce the risk of having inconsistencies in the NMA, at the expenses of introducing bias due to non-randomized treatment assignment.

4.4. Influence of statistical software

The methodological development in this area has been strongly influenced by the statistical software BUGS (Lunn *et al*., 2009) and Bayesian methodological papers published after 2000 used BUGS. Moreover, the published BUGS scripts allow practitioners to use this software in their own applications. This trend contrasts with papers published during the nineties, where the main focus was on methodological research with little chance of using these methods in clinical contexts. We can consider this development as a success of the early ideas of Eddy's CPM.

Compared with other statistical areas, the development of R packages for meta-analysis has been slow and simplistic where most of the R packages for meta-analysis are focused on single study type meta-analysis. There is a lot of work that remains to be done in software development in this area.

4.5. Some practical advice

Combining different study types in a single meta-analysis is motivated by the principle of using all of the available evidence in a meta-analysis. However, we have seen in this review that there are many alternative methods to perform this task. Some of these methods require substantive input from outside the statistical analysis (e.g., the Turner bias model).

Clearly, transparency in the data collection and detailed information on each study included in the review is one of the basic premises in meta-analysis, but combining different study types demand an extra modeling effort. We add the following advice to practitioners in this area:

- Regardless of which meta-analysis approach is used, we should investigate external sources of information that may help in the bias modeling process. We could use this information for prior elicitation of bias (Turner *et al*., 2009), for direct likelihood adjustments (Wolpert and Mengersen, 2004), for meta-regression approaches (McCarron *et al*., 2010), for empirical bias modeling (Welton *et al*., 2009), or for quality weighting (Spiegelhalter and Best, 2003).
- Before combining different study types in a single meta-analysis, we should first make a separate meta-analysis for each study type. Exploring the differences and contradictions between results may help the modeling process. For example, increases of variability between-study types may be resolved by using grouped random-effects techniques (Larose and Dey, 1997; Prevost *et al*., 2000).
- We may ask to which extent does the model fitted predict future results? Model validation in meta-analysis is not very popular, but it should be like any other statistical modeling problem. Bayesian predictive data are conditionally independent from the data used to build the model and can be used for model checking in meta-analysis (Higgins *et al*., 2009; Verde, 2010).
- Can we detect conflict between pieces of evidence? The *conflict assessment* is the *deconstructionist side of evidence synthesis*, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. Conflict assessment of pieces of evidence in meta-analysis is a new area of methodological research. One possibility is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. For example, Verde 2010a applied a scale mixture of multivariate normal and he made conflict diagnostics by direct interpretation of the scale weights. Another alternative is presented by Presanis *et al*. (2013), where the authors described how to generalize the conflict *p*-value proposed by Marshall and Spiegelhalter (2007) to complex evidence modeling.
- Unfortunately, bias modeling cannot be validated, but a sensitivity analysis based on predictive data can be used to understand how conclusions from a meta-analysis are affected by the inclusion of different study types. We should have in mind that usually there is not 'a best model'. Examples of applications such as those described by Spiegelhalter and Best 2003) show that combining disparate evidence ends in a stochastic sensitivity analysis and not to a single best analysis.
- Bayesian hierarchical models have been the most popular approach for combining disparate sources of evidence, but there are a number of issues from the practical perspective, such as when to judge studies or study types 'exchangeable', how to put suitable priors on variance components, which type of sensitivity analysis is particularly relevant, and so on.

5. Acknowledgements

The authors are very grateful to David Spiegelhalter for his comments on a draft version of this paper. We also would like to thank the associate editor for his comments and two reviewers (supplied anonymously), which helped us to improve the quality of this paper. This work was supported by the German Research Foundation project DFG Oh 39/11-1.

References

Abdelhamid A, Loke Y, Parekh-Bhurke S, Chen Y, Sutton A, Eastwood A, Holland R, Song F. 2012. Use of indirect comparison methods in systematic reviews: a survey of Cochrane review authors. *Research Synthesis Methods* **3**: 71–79.

Adar R, Critchfield G, Eddy D. 1989. A confidence profile analysis of the results of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower-extremity ischemia. *Journal of Vascular Surgery* **10**: 57–67.

Ades A. 2003. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine*. **22**: 2995–3016.

Ades A, Sutton A. 2006. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**: 5–35.

Agostino R. 1998. Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomised control group. *Statistics in Medicine* **2281**: 2265–2281.

Ahern J, Hubbard A, Galea S. 2009. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *American Journal of Epidemiology* **169**: 1140–1147.

Anonymous. 1992. Cross design synthesis: a new strategy for studying medical outcomes? *The Lancet* 944–946.

Austenfeld M, Thompson I, Middleton R. 1994. Meta-analysis of the literature: guideline development for prostate cancer treatment. *The Journal of Urology* **152**: 1866–1869.

Begg C. 1992. Book review: cross design synthesis: a new strategy for medical effectiveness research. *Statistics in Medicine* **11**: 1627–1630.

Begg C, Pilote L. 1991. A model for incorporating historical controls into a meta-analysis. *Biometrics* **47**: 899–906.

Benson K, Hartz A. 2000. A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine* **342**: 1878–1886.

Bittl JA, He Y, Jacobs AK, Yancy CW, Normand S-LT. 2013. Bayesian methods affirm the use of percutaneous coronary intervention to improve survival in patients with unprotected left main coronary artery disease. *Circulation* **127**: 2177–2185.

Block P, Peterson E, Krone R, Kesler K, Hannan E, O'Connor G, Detre K. 1998. Identification of variables needed to risk adjust outcomes of coronary interventions: evidence-based guidelines for efficient data collection. *JACC* **32**: 275–82.

Brumback BA, Holmes LB, Ryan LM. 1999. Adverse effects of chorionic villus sampling: a meta-analysis. *Statistics in Medicine* **18**, 2163–2175.

Chelimsky E, Silberman G, Droitcour J. 1993. Cross design synthesis. *The Lancet* **341**, 498.

Clowse M, Behera M, Anders C, Copland S, Coffmann C, Leppert P, Bastian L. 2009. Ovarian preservation by GnRH agonists during chemotherapy: a meta-analysis. *Journal of Women's Health* **18**(3): 311–319.

Cole S, Stuart, E. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology* **172**, 107–115.

Copas J. 2013. A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**: 47–66.

Craxi A, Camma C, Giunta M. 1999. Definition of response to antiviral therapy in chronic hepatitis C. *Journal of Hepatology* **31**: 160–167.

Craxi A, DiBona D, Camma C. 2003. Interferon-alpha for HBeAg-positive chronic hepatitis B. *Journal of Hepatology* **39**, 99 – 105.

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakrovitch C, Song F, Petticrew M, Altman DG. 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment NHS R&D HTA Programme* **7**(27): 1–173.

Dias S, Welton N, Caldwell J, Ades A. 2010. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**: 932–944.

Dias S, Sutton A, Ades A, Welton N. 2013. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* **33**: 607–617.

Dominici F, Parmigiani G, Wolpert R, Hasselblad V. 1999. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *Journal of the American Statistical Association* **94**: 16–28.

Donegan S, Williamson P, D'Alessandro U, Smith CT. 2012. Assessing the consistency assumption by exploring treatment by covariate interaction in mixed treatment comparison meta-analysis: individual patient level covariate versus aggregate trial level covariates. *Statistics in Medicine* **31**: 3840–3857.

Donegan S, Williamson P, D'Alessandro U, Garner P, Smith CT. 2013. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine* **32**: 914–930.

Droitcour J, Silberman G, Chelimsky E. 1993. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care* **9**: 440–9.

Eddy DM. 1989. The confidence profile method - a Bayesian method for assessing health technologies." *Operations Research* **37**: 210–28.

Eddy DM. 2013. "Top 10 projects. The Confidence Profile Method." *Available from*: http://www.davidmeddy.com/Top10projects.htm.

Eddy DM, Hasselblad V, Shachter R. 1990a. A Bayesian method for synthesizing evidence. The confidence profile method." *International Journal of Technology Assessment in Health Care* **6**: 31–55.

Eddy DM, Hasselblad V, Shachter R. 1990b. An introduction to a Bayesian method for meta-analysis: the confidence profile method. *Medical Decision Making* **10**: 15–23.

Eddy DM, Hasselblad V, Shachter R. 1992. Meta-analysis by the confidence profile method: the statistical synthesis of evidence. Academic Press, San Diego, CA.

Egger M, Ebrahim S, Davey Smith G. 2002. Where now for meta-analysis? *International Journal of Epidemiology* **31**: 1–5.

Evans J, Green J, Carlin P, Barrett L. 1995. Meta-analysis of antibiotic in tube thoracostomy. *The American Surgeon* **61**: 215–9.

Fortin M, Dionne J, Pinho G, Gignac J, Lapointe L. 2006. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Annals Of Family Medicine* **4**: 104–108.

Frangakis C. 2009. The calibration of treatment effects from clinical trials to target populations. *Clinical Trials* **6**: 136–140.

Greenhouse J, Kaizar E, Kelleher K, Seltman H, Gardner W. 2008. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Statistics in Medicine* **27**: 1801–1813.

Greenland S. 2005. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**: 267–306.

Grines C, Nelson T, Safian R, Hanzel G, Goldstein J, Dixon S. 2008. A Bayesian meta-analysis comparing AngioJet® thrombectomy to percutaneous coronary intervention alone in acute myocardial infarction. *Journal of Interventional Cardiology* **21**: 459–482.

Hall M, Chang S, Dalbagni G, Pruthi R, Seigne J, Skinner E, Wolf J, Schellhammer P. 2007. Guideline for the management of nonmuscle invasive bladder cancer (stages Ta, T1 and Tis): 2007 update. *The Journal of Urology* **178**: 2314–2330.

Hasselblad V. 1998. Meta-analysis of multi-treatment studies. *Medical Decision Making* **18**: 37–43.

Higgins J, Whitehead A. 1996. Borrowing strength from external trials in a meta-analyses. *Statistics in Medicine* **15**: 2733–2749.

Higgins J, Thompson S, Spiegelhalter D. 2009. A re-evaluation of random-effects meta-analysis. *Journal of Royal Statistical Society: Series A* **172**: 137–159.

Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* **3**: 98–110.

Higgins J, Ramsay C, Reeves B, Deeks J, Shea B, Valentine J, Tugwell P, Wells G. 2013. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 12–25.

Hinkelmann K, Kempthorne O. 1994. Design and analysis of experiments. Volume I: introduction to experimental design. John Wiley & Sons, New York.

Huang J, van Gelder J. 2002. The probability of sudden death from rupture of intracranial aneurysms: a meta-analysis. *Neurosurgery* **51**: 1001–1007.

Hurwitz E, Ake P, Adams A, Meeker W, Shekelle P. 1996. Manipulation and mobilization of the cervical spine. A systematic review of the literature. *Spine (Phila Pa 1976)* **21**(15): 1746–59.

Ibrahim J, Chen M. 2000. Power prior distributions for regression models. *Statistica science* **15**: 46–60.

Imai K, King G, Stuart E a. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**: 481–502.

Ioannidis J. 2010. Meta-research: the art of getting it wrong. *Research Synthesis Methods* **1**: 169–184.

Jackson C, Best N, Richardson S. 2006. Improving ecological inference using individual-level data. *Statistics in Medicine* **25**: 2136–2159.

Jackson C, Best N, Richardson S. 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**: 159–178.

Jackson C, Best N, Richardson S. 2009. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics* **10**: 335–351.

Jones W, Schmit K, Vemulapalli S, Subherwal S, Patel M, Hasselblad V, Heidenfelder B, Chobot M, Posey R, Wing L, Sanders G, Dolor R. 2013 May. Treatment strategies for patients with peripheral artery disease. *Rockville* (*MD*): *Agency for Healthcare Research and Quality* (*US*) Report No.: 13-EHC090-EF.

Kaizar E. 2011. Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine* **30**: 2986–3009.

Klotzbuecher C, Ross P, Landsman P, Abbott T, Berger M. 2000. Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. *Journal of Bone and Mineral Research* **15**(4): 721–39.

Larose D, Dey D. 1997. Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine* **16**: 1817–1829.

Leach GE, Dmochowski RR, Appell RA. 1997. Female stress urinary incontinence clinical guidelines panel summary report on surgical management of female stress urinary incontinence. *Journal of Urology* **158**: 875–880.

Lefevre F, Aronson N. 2000. Ketogenic diet for the treatment of refractory epilepsy in children: a systematic review of efficacy. *Pediatrics* **105**(4): E46.

Lehmann H, Hinton R, Morello P, Santoli J. 2000. Developmental dysplasia of the hip practice guideline: technical report. *Pediatrics* **105**: 896–905.

Li Z, Begg C. 1994. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association* **89**: 1523–1527.

Licata A, DiBona D, Schepis F, Shahied L, Craxi A, Camma C. 2003. When and how to treat acute hepatitis C?" *Journal of Hepatology* **39**: 1056–1062.

Lu G, Ades A. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**. 3105–3124.

Lu G, Ades A. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**(474): 447–459.

Lu G, Ades A. 2009. Modelling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10**: 792–805.

Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. 2012. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Research Synthesis Methods* **2**(1): 43–60.

Lumley T. 2002. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**(16): 2313–2324.

Lunn D, Spiegelhalter D, Thomas A, Best N. 2009. The BUGS project: evolution, critique and future directions. *Statistics in Medicine* **28**(25): 3049–3067.

Lunn D, Barrett J, Sweeting M, Thompson S. 2013. Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C: Applied Statistics* **62**(4): 551–572.

Marshall E, Spiegelhalter D. 2007. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis* **2**: 409–444.

McCarron C, Pullenayegum E, Thabane L, Goerree R, Tarride JE. 2010. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesis evidence from randomized and non-randomized studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Medical Research Methodology* **10**: 64.

McCarron C, Pullenayegum E, Thabane L, Goerree R, Tarride JE. 2011. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *Plos One* **6**(10).

Mesgarpour B, Heidinger B, Schwameis M, Kienbacher C, Walsh C, Schmitz S, Herkner H. 2013. Safety of off-label erythropoiesis stimulating agents in critically ill patients: a meta-analysis. *Intensive Care Medicine* **39**(11): 1896–908.

Molitor J, Jackson C, Best N, Richardson S. 2009. Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: application to low birth weight and water. *Journal of the Royal Statistical Society: Series A* 32–50 **172**(3): 615–637.

Montague D, Barada JH, Belker AM. 1996. Clinical guidelines panel on erectile dysfunction: summary report on the treatment of organic erectile dysfunction. *Journal of Urology* **156**: 2007–2011.

Nanda K, Bastian L, Hasselbad V, Simel D. 1999. Hormone replacement therapy and the risk of colorectal cancer: a meta-analysis. *Obstetrics and Gynecology* **93**: 880–888.

Neuenschwander B, Branson M, Spiegelhalter D. 2009. A note on the power prior. *Statistics in Medicine* **28**: 3562–3566.

Nixon R, Duffy S. 2002. Cross-issue synthesis: potential application to breast cancer, tamoxifen and genetic susceptibility. *Journal of Cancer Epidemiology and Prevention* **7**: 205–212.

Norris S, Moher D, Reeves B, Shea B, Loke Y, Garner S, Anderson L, Tugwell P, Wells G. 2013. Issues relating to selective reporting when including non-randomized studies in systematic reviews on the effects of healthcare interventions. *Research Synthesis Methods* **4**: 36–47.

O'Rourke K, Walsh C. 2010. Impact of stroke units on mortality: a Bayesian analysis. *European Journal of Neurology* **17**: 247–251.

O'Rourke K, Walsh C, Hutchinson M. 2007. Outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis: a Bayesian analysis. *Journal of Neurology* **254**: 1547–1554.

O'Rourke K, Walsh C, Kelly P. 2009. Safety and efficacy of IV-TPA for ischaemic stroke in clinical practice - a Bayesian analysis. *Cerebrovascular Diseases* **28**: 572–581.

Orr R. 1999. The impact of prophylactic axillary node dissection on breast cancer survival - a Bayesian meta-analysis. *Annals of Surgical Oncology* **6**(1): 109–116.

Committee on quality improvement. 2000. Clinical practice guideline: early detection of developmental dysplasia of the hip. *Pediatrics* **105**.

Peters J, Rushton L, Sutton A, Jones D, Abrams K, Mugglestone M. 2005. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. *Journal of the Royal Statistical Society, Series C* **54**: 159–172.

Peterson E, Shaw L, Califf R. 1997. Risk stratification after myocardial infarction. *Annals of Internal Medicine* **126**: 561–582.

Piepho H, Williams E, Madden L. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* **68**: 1269–1277.

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using gibbs sampling JAGS: Just Another Gibbs Sampler. *Proceedings of DSC*.

Prentice R, Langer R, Stefanick M, Howard B, Pettinger M, Anderson G, Barad D, Curb J, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J. 2006. Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American Journal of Epidemiology* **163**: 589–99.

Presanis AM, Ohlssen D, Spiegelhalter D, De Angelis D. 2013. Conflict diagnostic in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* **28**(3): 376–397.

Prevost T, Abrams K, Jones D. 2000. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* **19**: 3359–3376.

Reeves B, Higgins J, Ramsay C, Shea B, Tugwell P, Wells G. 2013. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 1–11.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. 2008. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* **27**: 1870–1893.

Rothstein HR, Sutton AJ, Borenstein M. 2005. Publication bias in meta-analysis: prevention, assessment and adjustment. Wiley, Chichester.

Salanti G, Higgins JP, Ades A, Ioannidis JP. 2008. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* **17**: 279–301.

Sampath S, Moran JL, Graham PL, Rockliff S, Bersten AD, Abrams KR. 2007. The efficacy of loop diuretics in acute renal failure: assessment using Bayesian evidence synthesis techniques. *Critical Care Medicine* **35**: 2516–2524.

Schmitz S, Adams R, Walsh C. 2013. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine* **32**: 2935–2949.

Schünemann H, Tugwell P, Reeves B, Akl E, Santesso N, Spencer F, Shea B, Wells G, Helfand M. 2013. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 49–62.

Segura J, Preminger G, Assimos D, Dretler S, Kahn R, Lingeman J, Macaluso J, McCullough D. 1994. Nephrolithiasis clinical guidelines panel summary report on the management of staghorn calculi. *The Journal of Urology* **151**: 1648–1651.

Segura J, Preminger G, Assimos D, Dretler S, Kahn R, Lingeman J, Macaluso J. 1997. Uteral stones clinical guidelines panel summary report on the management of ureteral calculi. *Journal of Urology* **158**: 1915–1921.

Senn S, Gavini F, Magrez D, Scheen A. 2011. Issues in performing a network meta-analysis. *Statistical Methods in Medical Research* **22**(2): 169–189.

Shachter RD, Eddy DM, Hasselblad V. 1990. An influence diagram approach to medical technology assessment. In Influence Diagrams, Belief Nets, and Decision Analysis, Oliver RM, Smith JQ (eds.). Wiley, Chichester; 321–350.

Shaw L, Peterson E, Kesler K, Hasselbad V, Califf R. 1996. A metaanalysis of predischarge risk stratification after acute myocardial infarction with stress electrocardiographic, myocardial perfusion, and ventricular function imaging. *American Journal of Cardiology* **78**: 1327–1337.

Shekelle P, Adams A, Chassin M, Hurwitz E, Brook R. 1992. Spinal manipulation for low-back pain. *Ann* **117**(7): 590–8.

Smith JA, Labasky RF, Cockett ATK, Fracchia JA, Montie JE, Rowland RG. 1999. Bladder cancer clinical guidelines panel summary report on the management of nonmuscle invasive bladder cancer (stages Ta, T1 and TIS). *The Journal of Urology* **162**: 1697–1701.

Soares MO, Dumville JC, Ades AE, Welton NJ. 2014. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **177**: 259–279.

Spiegelhalter D, Best N. 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* **22**: 3687–3709.

Spiegelhalter D, Abrams KR, Myles JP. 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Sutton AJ, Abrams KR. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10**: 277–303.

Sutton A, Higgins J. 2008. Recent developments in meta-analysis. *Statistics in Medicine* **27**: 625–50.

Sutton A, Song F, Gilbody S, Abrams K. 2000. Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research* **9**: 421–445.

Sutton A, Kendrick D, Coupland C. 2008. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* **27**: 651–669.

Thorlund K, Thabane L, Mills EJ. 2013. Modelling heterogeneity variance in multiple treatment comparison meta-analysis. Are informative priors the better solution? *BMC Medical Research Methodology* **13**: 2.

Turner R, Spiegelhalter D, Smith G, Thompson S. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**: 21–47.

Turner R, Lloyd J, Anumba D, Smith G, Spiegelhalter D, Squires H, Stevens J, Sweeting M, Urbaniak S, Webster R, Thompson S. 2012. Routine antenatal anti-dD prophylaxis in women who are RhD negative: meta-analyses adjusted for differences in study design and quality. *PLoS ONE 7* **2**: e30711.

Valentine J, Thompson S. 2013. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 26–35.

Verde PE. 2010. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine* **29**: 3088–3102.

Verde PE. 2013. Bamdit: Bayesian meta-analysis of diagnostic test data, r package version 1.1.

Wells GA, Shea B, Higgins J, Sterne J, Tugwell P, Reeves BC. 2013. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Research Synthesis Methods* **4**: 63–77.

Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**: 119–136.

White IR, Barrett JK, Jackson D, Higgins JPT. 2012. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* **3**: 111–125.

Wolpert RL, Mengersen KL. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science* **19**: 450–471.

Zimmerman M, Chelminski I, Posternak M. 2004. Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *The Journal of Nervous and Mental Disease* **192**: 87–94.

Chapter 4

Bayesian Cross-Design Synthesis

”Taking a model too seriously is really just another way of not taking it seriously at all.”

- Andrew Gelman, 30 November 2009.

Submission history:

This paper was submitted to Statistics in Medicine journal for publication in July 2014. The authors received a positive report from the associate editor and two anonymous reviewers in October 2014. The version of the paper presented in this chapter corresponds to the revised and uploaded one to Statistics in Medicine in April 2015. This paper was accepted for publication in June 2015.

Contributions of the authors:

PEV designed and wrote the paper, proposed the new Bayesian hierarchical model, wrote the computer programs and performed the statistical analysis. CO proposed the research work, performed the systematic review of the clinical papers and provided several inputs into the statistical model. SM provided medical background in diabetes and collected the data for the cohort used in the paper. AI provided epidemiological expertise and provided several inputs in the introduction and discussion sections.

Research Article

Statistics in Medicine

Received 10 July 2014, Accepted 25 October 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6809

Bayesian evidence synthesis for exploring generalizability of treatment effects: a case study of combining randomized and non-randomized results in diabetes

Pablo E. Verde,[a*†] Christian Ohmann,[a] Stephan Morbach[b] and Andrea Icks[c]

In this paper, we present a unified modeling framework to combine aggregated data from randomized controlled trials (RCTs) with individual participant data (IPD) from observational studies. Rather than simply pooling the available evidence into an overall treatment effect, adjusted for potential confounding, the intention of this work is to explore treatment effects in specific patient populations reflected by the IPD. In this way, by collecting IPD, we can potentially gain new insights from RCTs' results, which cannot be seen using only a meta-analysis of RCTs. We present a new Bayesian hierarchical meta-regression model, which combines submodels, representing different types of data into a coherent analysis. Predictors of baseline risk are estimated from the individual data. Simultaneously, a bivariate random effects distribution of baseline risk and treatment effects is estimated from the combined individual and aggregate data. Therefore, given a subgroup of interest, the estimated treatment effect can be calculated through its correlation with baseline risk. We highlight different types of model parameters: those that are the focus of inference (e.g., treatment effect in a subgroup of patients) and those that are used to adjust for biases introduced by data collection processes (e.g., internal or external validity). The model is applied to a case study where RCTs' results, investigating efficacy in the treatment of diabetic foot problems, are extrapolated to groups of patients treated in medical routine and who were enrolled in a prospective cohort study. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: cross-design synthesis; Bayesian hierarchical models; conflict of evidence; bias modeling

1. Introduction

After reviewing and analyzing experimental evidence of randomized clinical trials (RCTs), researchers are usually interested in assessing if these results can be extended to clinical practice. Although high quality RCTs are the gold standard for efficacy research, the context of experimentation is usually different from the context of application, which limits their external validity in clinical practice. For example, a major hurdle in the generalizability of trial results is the presence of a potential effect modification such as comorbidity. In this case, the magnitude of the treatment effect may vary according to the presence of coexisting diseases ([1, 2]).

The topic of this article is to explore the generalization of RCTs' results to groups of patients that may be excluded from RCTs but are treated in medical routine care (e.g., patients with severe comorbidities). Typically, information about patients treated in routine care is available from registers, observational studies (e.g., cohort studies), or other sources. The value of observational data might be attenuated when there is no effect modification. In such a case, we could expect similar treatment effect, although the groups

[a] *Coordination Center for Clinical Trials, University of Duesseldorf, Duesseldorf, Germany*
[b] *Department of Diabetes and Angiology, Marienkrankenhaus, Hamburg, Germany*
[c] *Department of Public Health, University of Duesseldorf, Duesseldorf, Germany*
[*] *Correspondence to: Pablo E. Verde, Coordination Center for Clinical Trials, University of Duesseldorf, Moorenstr. 5, 40225 Duesseldorf, Germany.*
[†] *E-mail: pabloemilio.verde@uni-duesseldorf.de*

of interest have not been targeted by trials. However, as pointed out earlier, the presence of coexisting diseases could influence treatment effects.

Hence, in order to understand to what extent a new group of patients may benefit from a new treatment, we have to combine evidence from different study types, *randomized,* and *non-randomized*, and given that in practice, it is still difficult to have access to individual participant data (IPD) from RCTs, we also have to combine different data types, aggregated data (AD) results with IPD.

One simple approach to extrapolate results from a random effects meta-analysis of RCTs is to use the resulting posterior predictive distribution of treatment effect as an informative prior when analyzing the IPD. However, this approach could introduce a bias if the RCTs' populations are very different from IPD, or it could provide a weak information if the meta-analysis is based on a small number of studies. For these reasons, in this paper, we propose a unified framework in order to combine different data types simultaneously into a single model.

A new hierarchical meta-regression model is presented, which combines results from different study types and different data types. The model is built piece by piece by highlighting the data collection process (e.g., randomized and non-randomized) and the type of data of each piece of evidence (e.g., aggregated or individual). In this approach, experimental and non-experimental data are viewed as complementary sources of evidence, and the model can be used to understand to what extent it is possible to generalize RCTs' results to medical routine.

However, different study types are prone to different types of biases, and results might have different grades of quality. A great part of the work described in Section 3 is devoted to bias modeling issues. The external validity bias of RCTs is modeled by exploring the relationship between baseline risk and treatment effect. This model component is similar to the models presented by McIntosh [3], Thompson *et al.* [4], Sharp and Thompson [5], and Arends *et al.* [6], recently extended by Guolo [7, 8] and Ghidey [9], and applied in clinical context by Verde and Curcio [10]. The internal validity bias of RCTs is adjusted by a weighting approach, which penalizes unusual results by combining different scale normal distributions [11], and this approach is extended to account for the quality limitation of observational data [12]. The uncontrolled patient selection of observational evidence is modeled with a bias component, which relates individual patient characteristics to baseline risk.

In order to apply the model presented in this paper, the basic data requirement is that the same outcome variable is available across different study types. Hence, there are parameters (e.g., event rates) that are common to different study types, we called them *shared parameters*. In Section 3,we use these parameters to connect evidence between different study types.

Several statistical techniques have been developed to combine aggregated and individual-level data. Jackson *et al.* [13, 14] introduced the hierarchical related regression approach to combine observational aggregated and individual data with the aim of increasing statistical power and reducing ecological bias in epidemiology studies. Jackson *et al.* [15] combined different types of data with different covariates but with the same outcome variable. Statistical methods to combine aggregated and IPD in meta-analysis of randomized trials have been presented in Riley *et al.* [16] and Sutton *et al.* [17], where the authors explore the advantages of having individual data in both treatment groups.

A first heuristic attempt to extrapolate RCTs' results by using medical routine data was the cross-design synthesis [18], where the prediction of treatment effect on patients excluded from the RCTs was called the *empty cell problem*. Kaizar [19] evaluated statistical properties of the cross-design synthesis estimator by using an extensive simulation experiment. Verde and Ohmann [20] have reviewed and classified statistical techniques that have been used to combine randomized and non-randomized evidence during the last two decades. The model presented in this paper can be viewed as a combination of a hierarchical regression model with a weighting approach, which accounts for a multiplicity of biases. In this approach, the *empty cell problem* could be handled as a prediction problem in regression analysis.

This paper is organized as follows. In Section 2, we present a case study that investigates treatment efficacy in diabetic foot problems and its extrapolation to patients enrolled in a prospective cohort study. Statistical methods are described in Section 3. In Section 4, we present the statistical analysis and its results. These results should not be considered as a direct contribution to the treatment of diabetic patients, which would require further research. Finally, in Section 5, we give a summary and a brief discussion of the methods presented in this paper.

2. Description of the case study

2.1. The medical problem: diabetes and diabetic foot

Foot ulcers and lower extremity amputations are among the most significant complications of diabetes, and they both have a high risk of recurrence. Moreover, amputations cause higher health costs in patients with diabetes compared with non-diabetic patients [21], and those affected carry an increased risk of mortality [22, 23].

Evidence from RCTs showed that adjunctive therapies result in clinical efficacy and cost efficiency [24–26] compared with standard care healing for foot ulcers and amputations. The question is whether results available from RCTs, which have been performed in selected populations, can be generalized to other patient populations.

2.2. Description of the aggregated data

The AD used in this paper correspond to RCTs resulting from a systematic review [27]. Our clinical question was the effectiveness of adjunctive treatments in managing diabetic foot problems when the outcome variable was minor amputation. With this aim, we selected RCTs with the following criteria: (1) studies with patients with diabetic foot ulcer; (2) studies where the outcome variable of investigation was amputation; and (3) studies where the experimental group was treated with routine care and an adjunctive therapy for diabetic foot problems.

Adjunctive therapies differ in various ways as follows: Negative-pressure wound therapy is a widely used low-cost treatment; Hyperbaric Oxygen Therapy requires specific facilities and equipment and is usually performed in several sessions. Dalteparin is a drug therapy applied subcutaneously, which modulates blood coagulation; granulocyte colony-stimulating factor modulates the immune response to infection, and Human Epidermal Growth Factor is applied as a growth factor. In spite of these differences, our clinical question was about the general effectiveness of adjunctive treatments.

Table I presents results of the identified RCTs, including type of adjunctive therapy, number of patients in the control and treatment group and the number of minor amputations in the control and treatment group, where minor amputations range from toe amputations to amputations of the foot at the ankle joint. In addition, we have the follow-up time in days and two patient characteristics as follows: (1) if the study population includes patients with peripheral artery disease (*PAD*) and (2) patients' ulcer severity characterized by the Wagner score. A detailed description of patients and study characteristics of these studies are presented in the supplementary material of [28].

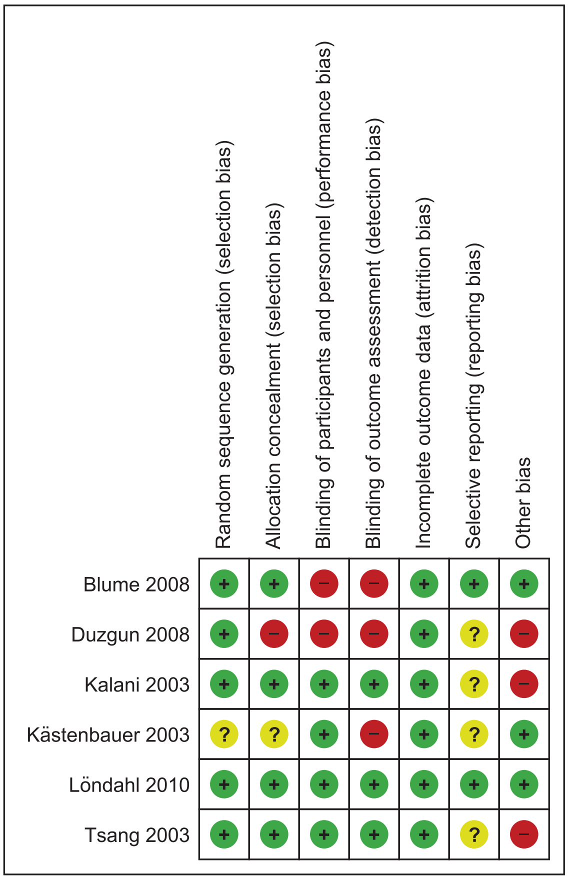
We performed an assessment of risk of bias of the studies included in this case study. This assessment was performed using the risk of bias tool in the Review Manager software (version 5.3.5). Results are summarized in Figures 1 and 2. The risk of bias summarized in Figure 1 shows that the study by Duzgun *et al.* 2008 has a high risk of bias. We found three possible factors that influence bias in this study: (1) selection bias due to allocation concealment; (2) further possible bias includes the higher prevalence of male patients, of obese patients, and of smokers in the treatment group; and (3) as mentioned by the authors in the discussion, bias introduced by not distinguishing between different types of foot ulcers.

Table I. Description of the randomized controlled trials by author and type of adjuvant therapy.

	Name	Adjunctive therapy	*n.ct*	*amp.ct*	*n.tr*	*amp.tr*	*f.up*	*PAD*	*Wagner*
1	Blume *et al*. 2008	NPWT	166	17	169	7	112	no	2 to 3
2	Duzgun *et al*. 2008	Hyperbaric Oxygen	50	24	50	4	30	no	2 to 3
3	Kaestenbauer *et al*. 2003	G-CSF	17	1	20	1	10	no	2 to 3
4	Kalani *et al*. 2003	Dalteparin	42	4	43	2	180	yes	1 to 2
5	Loendahl *et al*. 2010	Hyperbaric Oxygen	45	4	49	4	365	yes	1 to 3
6	Tsang *et al*. 2003	hEGF	19	2	21	2	84	no	1 to 2

The variables are as follows; *n.ct* is the number of patients in the control group, and *amp.ct* is the number of minor amputations in the control group. *n.tr* is the number of patients in the treatment group, and *amp.tr* is the number of minor amputations in the experimental group; *f.up* is the follow-up period in days. *PAD*, if yes, patients with peripheral artery disease included, and *Wagner* is the range of Wagner score of patients included in the randomized controlled trial.

NPWT, negative-pressure wound therapy; G-CSF, granulocyte colony-stimulating factor; hEGF, Human Epidermal Growth Factor.



	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants and personnel (performance bias)	Blinding of outcome assessment (detection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Other bias
Blume 2008	+	+	−	−	+	+	+
Duzgun 2008	+	−	−	−	+	?	−
Kalani 2003	+	+	+	+	+	?	−
Kästenbauer 2003	?	?	+	−	+	?	+
Löndahl 2010	+	+	+	+	+	+	+
Tsang 2003	+	+	+	+	+	?	−

Figure 1. Risk of bias summary: review authors' judgments about each risk of bias item for each study included.

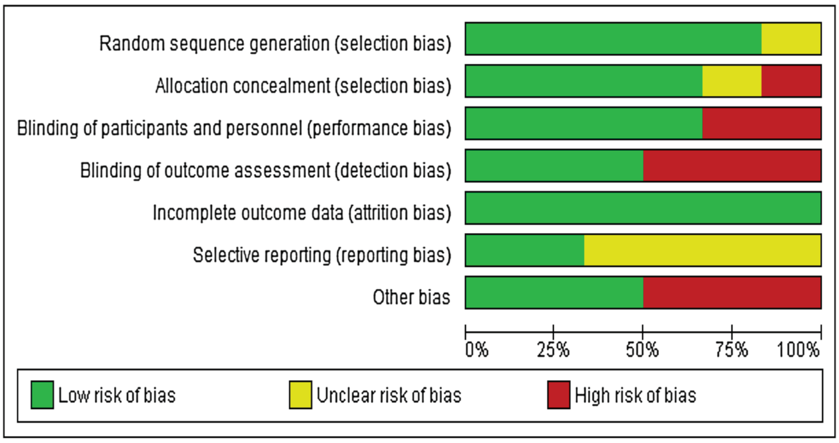


Figure 2. Risk of bias graph: review authors' judgments about each risk of bias item presented as percentages across all included studies.

The overall assessment of the six studies in Figure 2 shows that the bias domains "blinding of outcome assessment" and "other biases"could be important factors of bias (50 % of the RCTs scored at high risk).

2.3. *Description of the individual participant data*

The source of evidence for IPD used in this paper corresponds to the prospective cohort study of Morbach *et al.* [29]. The purpose of this study was to investigate risk factors associated with amputation as well as with mortality during a long-term period of at least 10 years, in a target population of participants presenting diabetic foot ulcers. The authors found that PAD, age, and being on dialysis were the most important risk factors for long-term prediction of major amputation.

We assess the quality of this study using the Newcastle–Ottawa assessment scale for cohort studies [30], with the following domain's results: *selection* scored four stars out of four and *outcome domain* three stars out of three. The *comparability domain* was not applicable in this study. In summary, this is a good observational study, which is somewhat representative of the target population.

The number of participants in this cohort was 260; at inclusion, the mean age was 68.92 ± 10.9 with a diabetes duration in years of 15.87 ± 10.59. Fifty-nine percent of the participants were male; 88% had type 2 diabetes, and 59 % of the participants were smokers or former smokers. Neuropathy and PAD were present at study initiation in 86% and 57% of the participants, respectively. Histories of a coronary event or stroke were reported by 20% and 21% of the subjects, respectively, without major differences according to sex. Ulcer severity was measured with the Wagner score from 1 to 5, where 1 indicates the lowest severity. The distribution of the Wagner score from 1 to 5 was: 34%, 20%, 27%, 17% and 2%, respectively. Further details on risk factors and comorbidities of the cohort are given in Section 4.

As described in Table I, the therapeutic evidence of minor amputation applies to RCTs with follow-up periods of less than a year. Therefore, we concentrated the analysis of IPD of this cohort study on the first year of follow-up.

3. Bayesian evidence synthesis modeling

In this section, we present a new Bayesian hierarchical meta-regression model, which combines submodels, representing different types of data, into a coherent analysis. In Sections 3.1 and 3.2,we describe the submodel used to combine RCT's AD. This is a bivariate random effects meta-analysis model, which accounts for two types of biases as follows: (1) external validity bias due to variation in the RCTs' baseline risk and (2) internal validity bias due to quality issues.

In Section 3.3, we introduce the submodel corresponding to IPD. This model accounts for observational bias, and it is linked with the AD submodel through two shared parameters. In Section 3.3.1, we present a conditional model that naturally predicts treatment effects in subgroups of patients by using a regression approach. Further modeling details are presented in Section 3.4.

3.1. Data model for aggregated and individual data

Consider a meta-analysis of N **randomized studies**, where $y_{0,i}^{AD}$ denotes the number of events in the control group of study i ($i = 1, \dots, N$) arising from $n_{0,i}^{AD}$ subjects and $y_{1,i}^{AD}$ and $n_{1,i}^{AD}$ denote the equivalent quantities in the treatment group. The upper script AD is used to highlight that the results have been aggregated at the level of the study and we do not have access to IPD.

The outcome variables $y_{0,i}^{AD}$ and $y_{1,i}^{AD}$ are modeled with two binomial distributions as follows:

$$y_{0,i}^{AD} \sim \mathtt{Binomial}\left(p_{0,i}^{AD}, n_{0,i}^{AD}\right) \quad \text{and} \quad y_{1,i}^{AD} \sim \mathtt{Binomial}\left(p_{1,i}^{AD}, n_{1,i}^{AD}\right), \tag{1}$$

where $p_{0,i}^{AD}$ and $p_{1,i}^{AD}$ are the event rates for each group.

In addition, suppose we have evidence of participants treated as controls from an **observational** study ($i = N+1$), with the same outcome variable as in the randomized studies, with the individual participant outcome variable $y_{0,N+1,j}^{IPD}$ (for $j = 1, \dots, M$), and with several individual characteristics or risk factors $x_{j,1}, x_{j,2}, \dots, x_{j,p}$. The individual outcome variable $y_{0,N+1,j}^{IPD}$ is modeled with a Bernoulli distribution with the following:

$$p_{0,N+1,j}^{IPD} = \mathtt{Pr}\left(y_{0,N+1,j}^{IPD} = 1\right). \tag{2}$$

In order to simplify the notation in the following sections, we call $y_0^{AD} = \left(y_{0,1}^{AD}, \dots, y_{0,N}^{AD}\right)$, $y_1^{AD} = \left(y_{1,1}^{AD}, \dots, y_{1,N}^{AD}\right)$ and $y_0^{IPD} = \left(y_{0,N+1,1}^{IPD}, \dots, y_{0,N+1,M}^{IPD}\right)$ the respective vectors of data.

3.2. Random effects model for aggregated experimental evidence

For $i = 1, \dots, N$, we model between-studies variability with the following random components:

$$h\left(p_{0,i}^{AD}\right) = \theta_{1,i} \quad \text{and} \quad g\left(p_{1,i}^{AD}\right) - g\left(p_{0,i}^{AD}\right) = \theta_{2,i}, \tag{3}$$

where $\theta_{1,i}$ represents an explicit adjustment for a potential **external validity bias** and $\theta_{2,i}$ represents the relative treatment effect (i.e., relative to the control treatment). The random effect $\theta_{1,i}$ summarizes the number of patients' characteristics and study design features that may influence the treatment effect $\theta_{2,i}$, we called $\theta_{1,i}$ **the baseline risk effect** of study i.

The function $g(\cdot)$ corresponds to the link function, which defines the scale where the treatment effect is defined. The function $h(\cdot)$ represents the scale where a linear relationship between underlying risk and treatment effect is modeled. Two forms of link functions are used in this work: the logistic link $logit(p) = \log(p/(1-p))$, which represents the odds ratio in the logarithmic scale and the complementary log–log link function $\log(-\log(1-p)) = \log(H)$ where H represents the cumulative hazard up to the mean follow-up.

The effects $\theta_{1,i}$ and $\theta_{2,i}$ are modeled as exchangeable between studies, and they follow a *scale-mixture of bivariate normal* distributions with mean and variance as follows:

$$E\left[\begin{pmatrix}\theta_{1,i}\\ \theta_{2,i}\end{pmatrix}\right] = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad var\left[\begin{pmatrix}\theta_{1,i}\\ \theta_{2,i}\end{pmatrix}\right] = \frac{1}{w_i}\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix} \tag{4}$$

and scale mixing density

$$w_i \sim p(w_i). \tag{5}$$

The motivation for including $\theta_{1,i}$ into the model is twofold: On the one hand, participants' characteristics are aggregated at the study level, and they cannot be used to directly adjust treatment effect. On the other hand, study level factors are usually known (e.g., length of the follow-up period), but the amount of studies N could be too small to make a direct adjustment useful.

The inclusion of the random weights w_i into the model is similar to the bivariate random effect meta-analysis of Verde [11], where $p(w_i)$ allows for a great flexibility to model the marginal distribution of $\theta_{1,i}$ and $\theta_{2,i}$. Two important cases are as follows: $w_i \sim \texttt{Gamma}(\alpha, \beta)$ with $\alpha = \beta = \nu/2$, which corresponds to a marginal bivariate t-distribution with known degrees of freedom ν and $p(w_i = 1) = 1$, which corresponds to a bivariate normal distribution.

Another important aspect of w_i is its interpretation as **estimated bias correction**. A priori all studies included in the review have a mean of $E(w_i) = 1$. We can expect that studies, which are unusually heterogeneous, will have posteriors means substantially less than 1. If the model is not corrected by the influence of unusual study results, then the meta-analysis may produce biased results.

Unusual results could be produced by factors that may affect the quality of the study, such as dilution of the treatment effect, confounding factors, loss to follow-up, and others. For that reason, the studies' weights w_i can be interpreted as an adjustment of studies' **internal validity bias**.

The present model is a simplification of the complexity involved in modeling the between-study heterogeneity. There are several alternatives to extend the distribution of the random effects $\theta_{1,i}$ and $\theta_{2,i}$ including an asymmetric bivariate distribution or adding a regression equation that explains systematic changes of the studies weights $w_1, \dots, w_N$ [11].

3.3. Combining observational individual data with aggregated experimental data

The fact that the evidence concerning individual participants is non-experimental has to be handled with care. Clearly, data resulting from different study types might have different grades of quality. Therefore, the potential bias introduced by combining different study types has to be explicitly modeled. In this section, we consider a simple formulation that can be used as a first modeling step.

In order to combine results from different study types, we assume that participants treated as controls may have similar results across study types. Therefore, the event rate of the control group is a parameter defined across study types. Hence, the starting point to *connect* evidence across different study types is the combination of the event rate parameters p_0^{IPD} and $p_0^{AD} = h^{-1}(\mu_1)$ and the marginal variance $var(\theta_1) = \sigma_1^2$. We called μ_1 and σ_1^2 *shared parameters* to highlight that they are commonly estimated across different study types.

Now, let us suppose that because of uncontrolled patient selection and quality limitations, the cohort suffers from an internal bias ϕ that is modeled as follows:

$$\phi \sim \texttt{Normal}\left(\mu_\phi, \sigma_\phi^2\right). \tag{6}$$

Hence, the cohort has a baseline risk random effect represented by the following:

$$h\left(p^{IPD}_{0,N+1}\right) = \theta_{1,N+1} + \phi \tag{7}$$

$$= \theta^{IPD}_{1,N+1}. \tag{8}$$

The cohort effect $\theta^{IPD}_{1,N+1}$ combines the effect of a hypothetical experimental control group $\theta_{1,N+1}$ with an *intrinsic observational bias* ϕ and has a biased mean of

$$E\left(\theta^{IPD}_{1,N+1}\right) = \mu_1 + \mu_\phi \tag{9}$$

$$= \mu^{IPD}_1 \tag{10}$$

and a variance inflation of

$$var\left(\theta^{IPD}_{1,N+1}\right) = \sigma^2_1 + \sigma^2_\phi. \tag{11}$$

Now, by taking

$$w_{N+1} = \frac{\sigma^2_1}{\sigma^2_1 + \sigma^2_\phi}, \tag{12}$$

we have

$$var\left(\theta^{IPD}_{1,N+1}\right) = \frac{\sigma^2_1}{w_{N+1}} \tag{13}$$

$$= \sigma^{IPD\,2}_1 \tag{14}$$

in this way the variance of ϕ is modeled by using the shared parameter σ_1 and the weight component w_{N+1}, which are estimated across study types. Note that we assumed that $\theta_{1,N+1}$ is independent of the observational bias ϕ. In section 5, we briefly discuss this point.

The individual patients' characteristics $x_1, x_2, \ldots, x_p$ are used to reduce the influence of the patient selection bias μ_ϕ as follows:

$$E\left(\theta^{IPD}_{1,N+1} | x_{1,j}, \ldots x_{p,j}\right) = \mu_1 + \mu_\phi + \beta_1 x_{1,j} + \ldots + \beta_p x_{p,j}. \tag{15}$$

If the participants' characteristics correctly adjust for selection bias and there is no other known systematic bias, then we expect that the posterior of μ_ϕ will be concentrated around zero. Hence, the parameter μ_1 can be estimated by using evidences from different study types. In other words, experimental and non-experimental data are not in conflict, and we expect that the pooled amputation rate of the RCTs is similar to the base line risk of the cohort after adjusting by patient risk factors. In Section 3.4.1, we describe a procedure to check if experimental and non-experimental data are in conflict.

In our approach, we leave μ_ϕ as a free parameter, where its prior is updated by using the contribution of IPD. However, depending on the context of application, other authors have given a fixed value for μ_ϕ [12, 13].

The weight w_{N+1} represents an *adjustment for the quality* of observational evidence. An immediate question this approach raises is as follows: How much contribution does a non-randomized study have compared with a randomized trial? In this work, we follow a data driven approach by approximating the joint posterior of $w_1, \ldots, w_N$ and w_{N+1}. If the posterior of w_{N+1} is concentrated at 1, then empirically the observational evidence is not in conflict with the RCTs' evidence. In this approach, randomized and non-randomized evidences are partially exchangeable after learning from the posterior distribution of $w_1, \ldots, w_N$ and w_{N+1}.

If the posterior of w_{N+1} is concentrated at lower positive values, then the observational data strongly deviate from the RCTs' evidence. In this case, RCTs' results cannot be generalized to the observational data. One first step to harmonize this heterogeneity is to search for individual data covariates that may correct this deviation.

Clearly, if there is prior evidence that the observational data are of poor quality, then a deterministic approach can be applied by giving a fixed value to $w_{N+1} = k$, which penalizes evidence with lower quality. This was applied by Spiegelhalter and Best [12]. However, a deterministic penalization could be too arbitrary in practice. For these reasons, we propose an alternative procedure in this paper. Suppose that a priori we assume that $w_{N+1} \sim \texttt{Gamma}(a_{N+1}, b_{N+1})$, with the prior mean as follows:

$$E(w_{N+1}) = \frac{a_{N+1}}{b_{N+1}} = k, \tag{16}$$

and we are working with a model with v degrees of freedom. Now, by taking

$$a_{N+1} = v/2 \quad \text{and} \quad b_{N+1} = (v + \delta)/2, \tag{17}$$

we can elicit δ by $\delta = v(1-k)/k$.

We use this procedure in Section 4 to perform a sensitivity analysis of the predictive posteriors of subgroups of patients. For example, we fixed a priori a low value of $E(w_{N+1}) = 0.4$, and we checked if this value influenced the results. This procedure can be extended to any piece of evidence that we are combining in a meta-analysis.

3.3.1. The conditional model and the treatment effect of a subgroup of patients. The random effect model of $(\theta_{i,1}, \theta_{i,2})$ $(i = 1, \ldots, N)$ of Section 3.2 and the model of $\theta_{N+1,1}$ of Section 3.3 are equivalent to

$$\theta_{i,1} \sim \texttt{Normal}\left(\mu_1, \sigma_1^2/w_i\right), \tag{18}$$

$$\theta_{i,2}|\theta_{i,1} \sim \texttt{Normal}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_{i,1} - \mu_1), \left(1-\rho^2\right)\sigma_2^2/w_i\right) \tag{19}$$

and

$$\theta_{N+1,1} \sim \texttt{Normal}\left(\mu_1 + \mu_\phi + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma_1^2/w_{N+1}\right). \tag{20}$$

The conditional mean of $\theta_{i,2}|\theta_{i,1}$ is used to extrapolate a treatment effect for a particular value of $\theta_{i,1}$. In this work, we approximate the posterior conditional mean of $\theta_{i,2}|\theta_{i,1}$ by using Markov chain Monte Carlo (MCMC) computations and by specifying a suitable range of values of $\theta_{i,1}$ (Section 3.4.3).

We note that, if the posterior distribution of ρ is centered at zero, the treatment effect is summarized by the posterior distribution of μ_2. In such a case, we should be careful with the interpretation of results, and we should analyze whether there are other study level characteristics that can be useful to adjust for external validity by using for example a meta-regression approach (Section 3.4.2).

One interesting aspect of this conditional model is that it gives *a solution* to the *empty cell problem* in cross-design synthesis [18]. The idea here is to combine the distribution of (19) with (20). Hence, inference of treatment effect of a particular subgroup of patients characterized by the risk factor x_k $(k = 1, \ldots, p)$ and baseline risk at $\theta_1^{x_k} = \mu_1 + \mu_\phi + \beta_k$, is based on the posterior distribution of the functional parameter:

$$\eta_k = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\mu_\phi + \beta_k). \tag{21}$$

The baseline risk of the group $\theta_1^{x_k}$ gives the location on the axes of θ_1 where the treatment effect η_k is predicted. For details in the implementation, see the Bayesian analysis using gibbs sampling (BUGS) script in the appendix.

3.3.2. Priors for hyperparameters. The formulation of the model for aggregate data is completed by specifying the priors for the hyperparameters $\mu_1, \mu_2, \mu_\phi, \sigma_1, \sigma_2$, and ρ. We assume that parameters are independent and we use the following set of priors:

$$\mu_1 \sim \texttt{Logistic}(m_1, v_1), \quad \mu_2 \sim \texttt{Logistic}(m_2, v_2), \quad \mu_\phi \sim \texttt{Logistic}(m_3, v_3) \tag{22}$$

and

$$\sigma_1 \sim \texttt{Uniform}(0, u_1), \quad \sigma_2 \sim \texttt{Uniform}(0, u_2). \tag{23}$$

The correlation parameter ρ is transformed by using the Fisher transformation,

$$z = \mathtt{logit}\left(\frac{\rho+1}{2}\right)$$

and a Normal prior is used for z as follows:

$$z \sim \mathtt{Normal}(m_z, v_z). \tag{24}$$

Using independent priors that constrain $\sigma_1 > 0$, $\sigma_2 > 0$ and $|\rho| < 1$ guarantee that in each MCMC iteration, the variance–covariance matrix of the random effects θ_1 and θ_2 is positive definite. For implementation details, see the BUGS script in the appendix.

An alternative prior for the variance matrix is the inverse Wishart distribution. However, this distribution is less flexible than working directly with each parameter's prior. In addition, an inverse Wishart prior implies that σ_1^2 and σ_2^2 follow inverse Gamma priors, which might lead to biased results when a small number of studies are included in the analysis [31].

The values of the constants $m_1, v_1, m_2, v_2, m_3, v_3, u_1, u_2, m_z$, and v_z have to be given. They can be used to include valid prior information that might be empirically available, or they could be the result of expert elicitation. If such information is not available, we recommend to set these parameters to values that represent weakly informative priors. In this work, we use $m_1 = m_2 = m_3 = m_z = 0$, $v_1 = v_2 = v_3 = 4$, and $v_z = 1$ as weakly informative prior setup. These values are fairly conservative in the sense that they induce priors for the event rates $p_{0,i}^{AD}$ and $p_{1,i}^{AD}$, which have a U-shaped form like non-informative Jeffreys priors for probabilities. If we wish to work with uniform priors in the probability scale, we should take $v_1 = v_2 = v_3 = 1$. Our priors' setup gives locally uniform distributions for μ_1, μ_2, and μ_ϕ, uniforms for σ_1 and σ_2, and a symmetric distribution for ρ centered at 0.

It is well-known that for a small number of studies, the posteriors of σ_1 and σ_2 are sensitive to the priors, so different values of u_1 and u_2 should be used for a sensitivity analysis. In our experience, the most difficult parameter to estimate in this model is ρ. This is a critical parameter because it makes the difference between adjusting or not for external validity bias. Therefore, we recommend to perform *a prior to posterior sensitivity analysis* by giving different values for m_z and v_z to understand their effect in the model.

3.3.3. Priors for regression parameters. The prior distribution of the regression coefficients $\beta_1, \ldots, \beta_k$ encapsulates a variable selection procedure, and depending on the context of application, some priors' configuration may be more suitable than others. Basically, if regression parameters are *independent* a priori, no variable selection is performed. When these coefficients are modeled as *exchangeable* with a prior mean of 0 and unknown variance σ_β^2, then a shrinkage effect toward 0 is produced. This effect regularizes the model by penalizing the inclusion of uninteresting variables. In addition, the probability distribution used increases the penalty according to the heaviness of its tail. In this way, coefficients with posteriors far from 0 are considered statistically relevant for the model.

In general, we have to scale the covariates in order to make the assumption of exchangeability reasonable. In this application, every covariate is binary, and scaling covariates is therefore not necessary.

There is an extensive number of possibilities to handle variable selection in regression. For a recent review on this topic, see [32]. In order to be flexible when dealing with this model component, we apply a normal-gamma prior distribution to the coefficients $\beta_1, \ldots, \beta_p$ with the following parametric form [33]:

$$\beta_k \sim \mathtt{Normal}\left(0, \lambda_k \sigma_\beta^2\right), \quad \sigma_\beta \sim \mathtt{Uniform}(0, u_3) \tag{25}$$

and

$$\lambda_k \sim \mathtt{Gamma}(a, b). \tag{26}$$

When $a = b = 1$, the marginal distribution of β_k is a double exponential prior (i.e., a least absolute shrinkage and selection operator (LASSO) penalization); taking $a = b = \nu/2$ gives a t-distribution with ν degrees of freedom, and making $\lambda_1 = \cdots = \lambda_p = 1$ gives a Normal prior with unknown variance σ_β^2 (i.e., a Ridge penalization). We chose to work with $u_3 = 5$, which leaves the priors for β_k locally uninformative. We applied these different configurations to explore which regression models are more suitable for linking risk effects and events. However, as pointed out by one reviewer, in the context of our application, there are no practical reasons for decreasing the amount of covariates. Therefore, a LASSO

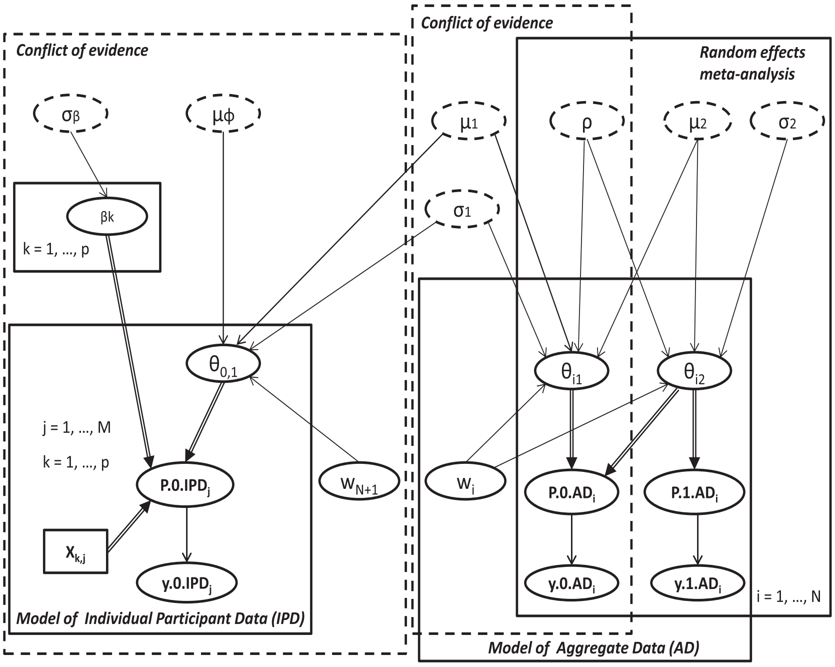


Figure 3. Directed acyclic graph for the model which combines aggregated data (AD) and individual participants data (IPD). Ellipses represent random variables and dashed ellipses indicate hyper-prior parameters. A double-lined arrow indicates a functional relationship between variables and a single-lined arrow a stochastic relationship. Two submodels are represented as follows: on the left hand side a model, which links individual risk factors to the outcome variable y_0^{ID}. On the right hand side a model for aggregated results $\left(y_0^{AD}, y_1^{AD}\right)$ adjusted by RCTs' external validity θ_1. Both submodels include a bias adjustment by internal validity w_i. Frames with dashed lines show the models used to assess conflict of evidence. The frame with the solid line displays a random effects meta-analysis model without external validity adjustment.

penalization seems unnecessary. Instead, a ridge regression could be preferred as this approach allows to correct a maximum of potential confounder effects.

3.4. Further modeling topics and statistical computations

3.4.1. Assessing conflict of evidence with a directed acyclic graph. An important issue when combining different sources of information is the potential inconsistency between pieces of evidence. The dashed frames in the directed acyclic graph (DAG) presented in Figure 3 show two important submodels that could be in conflict in our analysis: the left frame corresponds to the submodel with IPD only and the right frame corresponds to the submodel with AD only for the control group in the RCTs.

We assess a potential conflict of evidence between IPD and AD by setting $\mu_\phi = 0$ and $\rho = 0$ as constant nodes, and by splitting the node μ_1 into two nodes μ_1^{AD} the pooled mean of the control group of the RCTs and β_0 the intercept of the regression model of the individual participants data. In this way, the posteriors of μ_1^{AD} and β_0 are calculated independently of each other and without the influence of the treatment effect μ_2.

By comparing the resulting posteriors of μ_1^{AD} and β_0, we can assess if the RCTs' data are in conflict with the individual patient data. Substantial deviation between posteriors indicates inconsistency between study types. Formally, a Bayesian p-value can be calculated in order to measure the departure from the null hypothesis of no conflict. For technical details on calculating Bayesian p-values in conflict of evidence in hierarchical models, see [34] and more recently [35]. In this paper, we made an informal analysis by visualizing the posteriors of model parameters.

3.4.2. Inclusion of study level covariates. The extent to which observed study level characteristics influence treatment effect can be investigated with a meta-regression by including these covariates into the conditional mean of the model (19). In this work, two study level covariates presented in Table I are analyzed as follows. The first one is the length of follow-up in months, and the second one is the grouping factor, which indicates if the trial includes patients with *PAD*.

3.4.3. Software implementation and computations. The Bayesian hierarchical model presented in this work can be fitted by using MCMC simulations. Samples from the full posterior distribution of model parameters can be generated using the statistical software `WinBUGS` [36].

All calculations were implemented in `R` [37] and by calling `WinBUGS` from `R` using the package `R2WinBUGS` [38]. Results are based on two parallel MCMC simulations with 50,000 iterations, taking the first half of the iterations as burn-in period. Convergence was investigated using the R package `coda` [39]. The `BUGS` script used in the statistical analysis is part of the supplementary material of this paper and can be applied to similar types of statistical analysis.

4. Statistical analysis and results

4.1. Preliminary analysis for aggregated and individual patient data

4.1.1. Analysis of link functions. The modeling aspect of deciding, which link functions to use, was analyzed by comparing the posteriors of the hyperparameters under different combinations of link functions. We analyzed the following combinations of link functions: (1) logit and complementary log–log function for $g(\cdot)$ and (2) logit, complementary log–log and log for $h(\cdot)$.

The four combinations of link functions show an important overlapping of posteriors of μ_1, μ_2, σ_1, σ_2, and ρ. These results indicate that conclusions of the analysis are not sensitive to the choice of link functions. Therefore, to make results easy to interpret, we decided to use the logit link for $g(\cdot)$ and $h(\cdot)$ for further analyses.

4.1.2. Analysis of structural distribution. We started by comparing results with the bivariate normal distribution and a scale mixture distribution. We decided to work with $\nu = 4$ degrees of freedom, which is a value that is small enough to penalize studies with unusual results.

The mixture model clearly identified the study by *Duzgun et al.* (2008) as an outlier, with a posterior mean weight $E(w_2|y_0^{AD}, y_1^{AD}) = 0.5$. Interestingly, the risk of bias assessment in Section 2.2 pointed out that this study was at high risk for internal validity.

The mixture model automatically corrected the influence of unusual study results. In our case study, this correction strongly influences the posterior distribution of ρ. For the Normal model, the posterior mean was -0.559 and had a 95% credibility interval [-0.961, 0.013], while the corresponding results for the scale mixture had a posterior mean of -0.338 and a 95% credibility interval of [-0.822, 0.137]. Further analyses in this section are based on the scale mixture of normal distribution for random effects.

4.1.3. Analysis of covariates at study level. In order to analyze the influence on treatment effect based on study level characteristics, we performed a meta-regression analysis with the length of the follow-up period and the presence of patients with peripheral artery disease (PAD) as covariates. The posterior distribution of the regression coefficients has the following mean and 95 % credibility intervals: for follow-up 0.1 [-2.3, 2.5] and PAD 0.01 [-1.7, 1.6]. These results clearly show that study level covariates do not help to explain systematic variability between studies.

4.1.4. Regression analysis of participant individual data. The aim of this analysis is to find out which individual risk factors are associated with individual baseline risk. We analyzed three different Bayesian variable selection procedures as follows:

- No penalty for including covariates in the model. This procedure is achieved by using independent priors for regression coefficients.
- The use of *ridge penalization*, which corresponds to exchangeable regression coefficients with normal priors.
- The use of *lasso penalization*, which is equivalent to using exchangeable regression coefficients with double exponential priors.

Figure 4 summarizes the regression results by comparing coefficients' posteriors between the independent coefficients and the exchangeable ridge regression model. An important shrinkage effect can be observed by modeling coefficients as exchangeable. This effect adjusts the posterior distribution of the regression coefficients by pooling their values toward zero and by reducing the width of their posterior intervals.

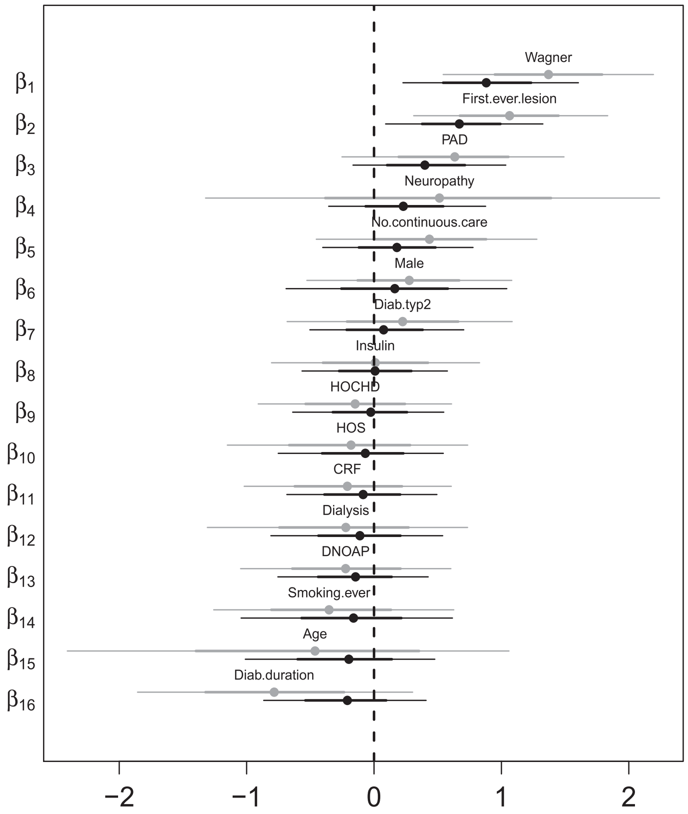


Figure 4. Caterpillar plot of posteriors of regression coefficients: For each regression coefficient β_i two 95% credibility interval are displayed, the upper grey lines correspond to the model with independent coefficients and the lower black lines correspond to the model with exchangeable coefficients. The most relevant risk factors identified in this analysis were: the classification of Wagner score (1, 2, 3 vs. 4), the first ever lesion (no/yes) and moderate effect of patent ductus arteriosus. PAD, peripheral artery disease .

The most important individual risk factors identified in this analysis were as follows: the classification of Wagner score (1, 2, 3 vs. 4), the first ever lesion (no/yes) and moderate effect of PAD.

We did not find any practical differences between ridge and LASSO results. Hence, we decided to use the former to combine individual and AD in the following Sections 4.2 and 4.3. Further results concerning the difference between ridge and LASSO results are presented in [28].

4.2. *Conflicts of evidence between randomized clinical trial results and observational individual data*

4.2.1. Independent analysis of aggregated data and individual participant data. Before combining disparate data together, an important question must be answered: *Are these pieces of evidence in conflict?* In this section, we analyze independently the AD from the IPD by following the recommendations presented in the discussion section of Verde and Ohmann [20].

In particular, we assess conflict of evidence by comparing the pooled amputation rate in the logistic scale of the RCT's control group μ_1^{AD} (Figure 3 dashed frame on the right) with β_0 the intercept of the regression model of the individual participants data (Figure 3 dashed frame on the left). As stated in Section 3.4.1, we calculate the posterior distribution of μ_1^{AD}, without the influence of treatment effect μ_2; we proceed by setting the correlation coefficient $\rho = 0$ and modeling the random effects θ_1 and θ_2 independently. The posterior distribution of μ_1^{AD} is compared with the posterior distribution of β_0.

For the AD, the posterior marginal mean of μ_1^{AD} is -1.914 with 95% posterior interval of (-2.714, -0.924), which completely overlaps the posterior of β_0, which has a mean equal to -2.016 and posterior 95% interval of (-3.155, -0.935). The left panel of Figure 5 shows the overlaps of the posterior distributions of these parameters, which clearly indicate no conflict of evidence between AD and IPD.

4.2.2. Analysis by combining aggregated data and individual participant data. In addition to the previous procedure, we estimate the posterior distribution of μ_1 and μ_ϕ by combining both AD and IPD with

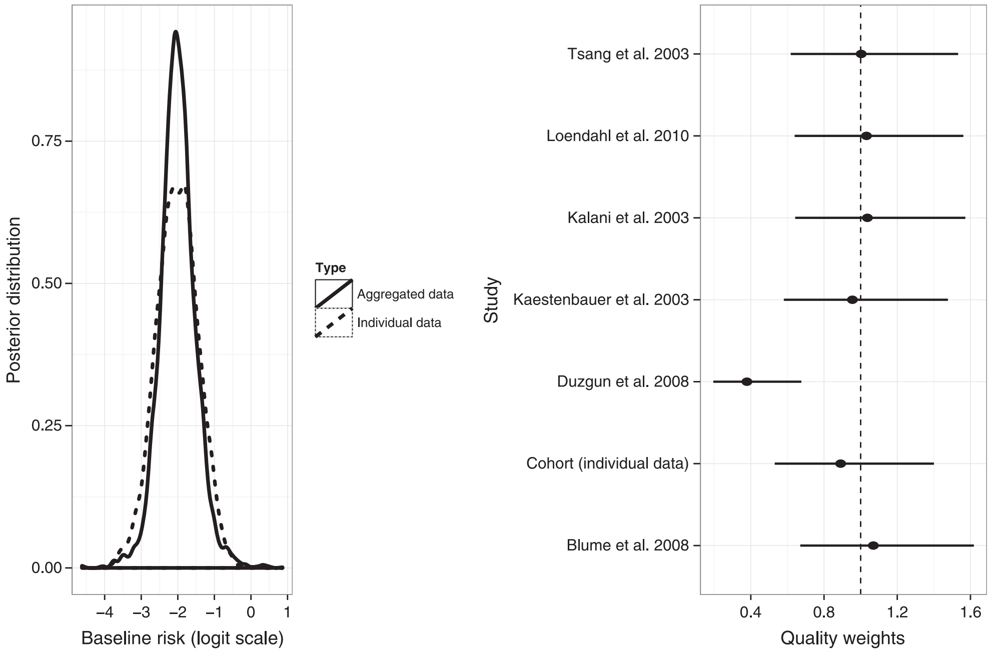


Figure 5. Left panel: Assessment of conflict of evidence by comparing posteriors of baseline risk for aggregated and individual data. Right panel: Posteriors of quality weights, the distributions overlapping the the vertical line at one correspond to studies without discounting.

the model presented in Section 3.3. The resulting posterior mean of μ_1 was -2.05 with 95% credibility interval of (-2.96, -1.12), and the posterior mean of μ_ϕ was -0.94 with 95% posterior interval of (-3.87, 1.98), which indicates that in our example, no additional intercept μ_ϕ is needed when observational data are combined with RCT data. These results are in the same line as the previous section. Therefore, we decided to set $\mu_\phi = 0$ for the subsequent analyses.

We gave equal prior weights to each RCT ($E(w_i) = 1$ for $i = 1, \ldots, N$) and to the cohort study ($E(w_{N+1}) = 1$). The right panel of Figure 5 shows the posteriors for the weights of each piece of evidence, we can see that the study by *Duzgun et al. (2008)* is down-weighted compared with the others RCTs.

Interestingly, the posterior distribution of the cohort's weight is centered at one, which shows concordance with the RCTs' results. We can see that in this case, experimental and observational data are not in conflict, which empirically validates their combination. In the right panel of Figure 5, we can also see that a prior mean of $k = 0.4$ or lower should be necessary for an additional penalization of the observational data. This result is used in Section 4.3.1 for sensitivity analysis of the predictive treatment analysis.

4.3. Results of combining aggregated and individual data

The results of the following sections are based on the full model, which combines the model of aggregated and individual patient data.

4.3.1. Extrapolation treatment effects for subgroups of patients. The aim of this section is to quantify the treatment effect for different groups of patients and the influence of different types of bias. The following scenarios are analyzed:

- Extrapolation of treatment effect by adjusting to external bias and taking the baseline risk of the cohort as the target population.
- We extend the extrapolation of the treatment effects to subgroups of patients. These include (1) patients with Wagner's score of 4, (2) patients with a previous lesion, and (3) patients with *PAD*.

  As depicted in Figure 4, the first two groups have a clear risk factor of amputation while the third group has a moderate risk factor.

Figure 6 summarizes the results of the hierarchical meta-regression model. The solid lines correspond to the posterior median and 95% credibility interval. The model shows that an increase in baseline risk corresponds to an increase in treatment effect (lower values on the vertical axis correspond to an increase in

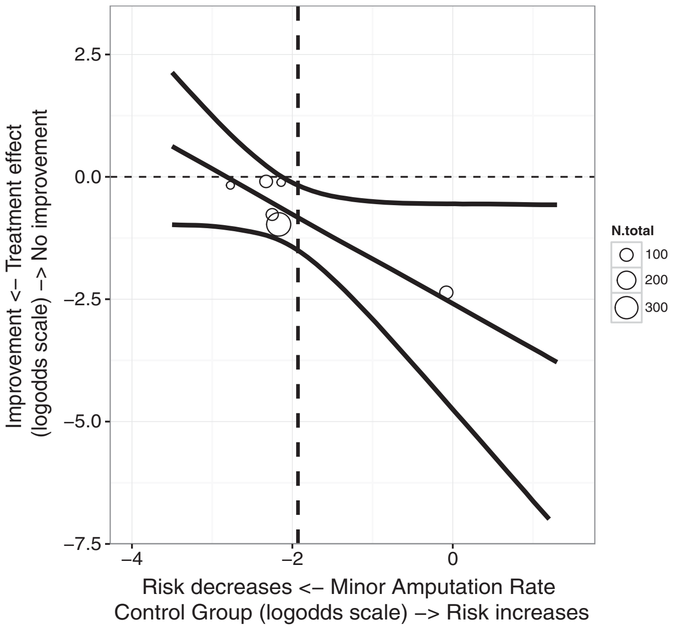


Figure 6. Summary results of generalizing treatment effects in logistic scale: randomized controlled trials' results are displayed as circles. The fitted hierarchical meta-regression model is summarized as follows: The black lines are the posterior median and 95% credibility interval intervals for the distribution of treatment effect given a range of values of the baseline risk. The vertical dashed line corresponds to the baseline risk of the cohort study.

treatment effect). The vertical dashed line marks the baseline risk of the cohort study, and the intersection with the solid lines corresponds to the prediction of treatment effect.

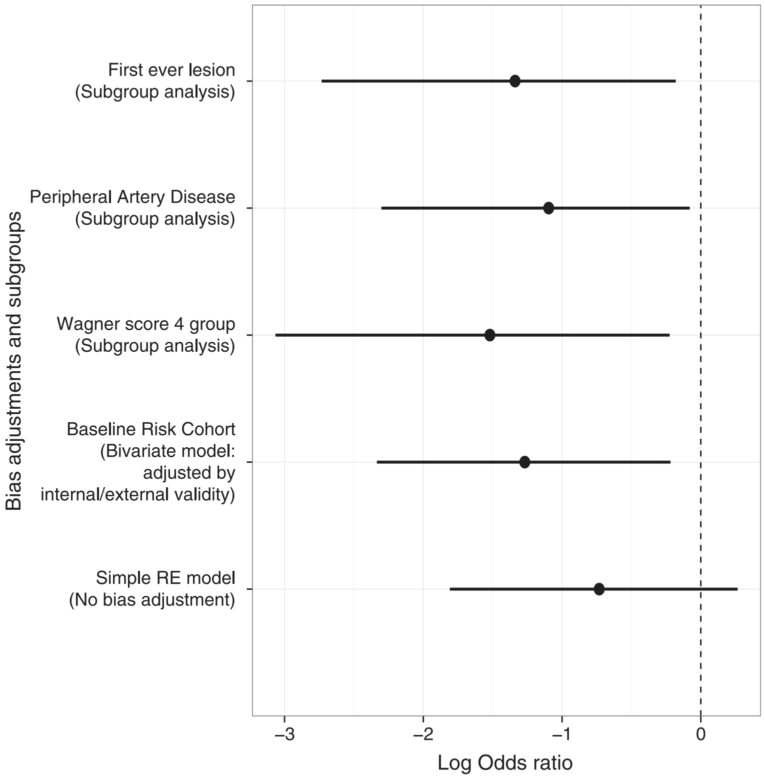


Figure 7. Summary results of generalizing treatment effects. From the bottom to the top the posterior means and 95% credibility intervals of the following cases are presented: no bias adjustment, extrapolation to the cohort baseline risk, and three subgroups of patients.

With more details, Figure 7 compares the 95% posterior intervals of the treatment effect estimated as follows: (1) a simple random effects model without bias correction, (2) the treatment effect of the cohort after adjusting for risk effects, and (3) the previously mentioned subgroups of patients. From the bottom to the top, we have the following results:

- The simple random effects model reveal a trend but not a definitive treatment effect. This result contrasted with the treatment effect adjusted by external validity bias and evaluated at the baseline risk of the cohort. This extrapolation gives a log odds ratio of -1.262 (-2.334, -0.218), which is a clear positive treatment effect. In this way, we expect that patients similar to those participating in the cohort would have a positive treatment effect.
- The next three lines correspond to the three subgroups of patients with Wagner's score 4, PAD positive, and a first ever lesion. Clearly, there is an increase in variability because of the sparsity of the data, but the trend is a reduction of the odds of amputation.
- In particular, the subgroup of patients with Wagner's score 4 has a log odds ratio of -1.5 (-3.0, -0.22), and the group of patients with a first ever lesion has a log odds of -1.3[-2.7, -0.18].
- Figure 8 shows the posterior predictive distributions for two subgroups of patients. The left panel corresponds to patients with Wagner score 4 and the right panel to the group of patients with a lesion at the time of enrollment in the cohort. Hence, if the treatment is applied to these subgroups, the predictive median and 95% posterior intervals for the number of amputations are as follows: (a) 3(0, 11) for the subgroup of patients with Wagner score equal to 4; and b) 6 (1, 219) for the subgroup of patients with first ever lesion. The observed numbers of amputations in each group were: 21 out of 49; and 31 out of 114, respectively. These results show that there is a potential advantage in using adjunctive therapies in these subgroups of patients. We performed a sensitivity analysis of the previous results by using a prior quality weight of 0.4 for the cohort study. Posterior predictive results were very stable with 4(0,11) and 10(2, 24) for each subgroup, respectively.

The results of this section showed the advantage of combining aggregated experimental data with observational individual data and the possibility to target subgroups of patients for further investigation.

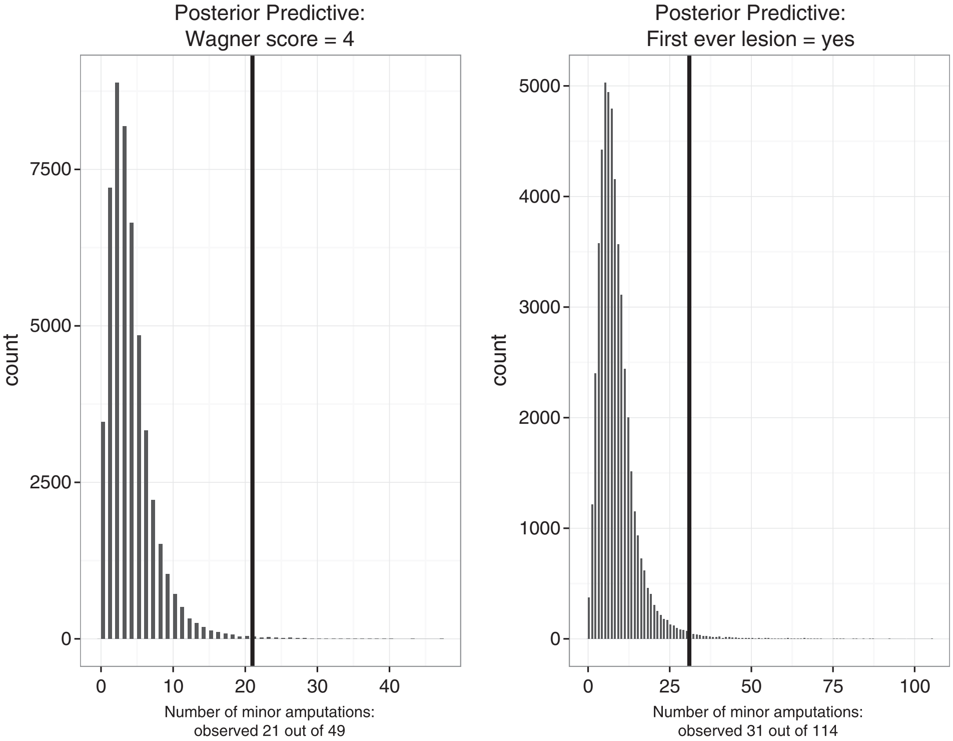


Figure 8. Posterior predictive results. Left panel: Subgroup of patients with Wagner score 4. The vertical line indicates 21 observed amputations out of 49 patients. Right panel: results for the group of patients which had a foot lesion when they enter the cohort, the vertical line indicates 31 amputations out of 114 patients.

If the AD can be adjusted by their internal and external validity bias, then as pictured in Figure 6, we can start linking AD with IPD evidence. Figure 7 highlighted the fact that by collecting IPD we can potentially gain new insights from RCTs' results which cannot be seen using a simple random effects model. Finally, the posterior predictions at the subgroup level displayed in Figure 8 clearly show that these insights can help to understand effects in unstudied patient groups.

5. Conclusions

In this article, we developed a new statistical framework for generalizing RCTs' results in clinical practice. This framework is based on a Bayesian evidence synthesis model, which combines aggregated experimental data with observational individual data of patients treated in medical routine care.

The proposed model can be divided into two submodels. The first one adjusts RCTs' treatment effects with their baseline risk, and the second one links individual risk factors with baseline risk. By combining these submodels, we can predict treatment effect for a particular subgroup of patients that could be underrepresented in the RCTs. In addition, a mechanism for adjusting each piece of evidence by its internal validity is integrated into this approach. In a way, this model can be viewed as a formal Bayesian approach of the heuristic cross-design synthesis method, and it is flexible enough to deal with multiple sources of bias which are usually present in these types of data. In our approach, the *empty cell problem* is naturally handled as a prediction problem in regression analysis.

Our motivation to develop this model was to assess to what extent results from RCTs, which showed a trend in efficacy of diabetic foot problems, can be extrapolated in medical routine care. With this end in mind, we analyzed AD from six RCTs with IPD from a cohort study. The analysis showed that the assessment of treatment effect in a new group of patients is possible, but the validity of these predictions may depend on the quality of the observational data. This type of analysis should be taken as hypothesis generating, for example, by targeting particular groups of patients that may benefit most from a new treatment, or for further investigation rather than as confirmatory results.

Our case study presents a strong agreement between results of RCTs' control groups and the cohort study. The demographic profiles of the patients participating in the RCTs were similar to those enrolled in the cohort. Neither lack of quality nor the patient selection in the cohort were at issue. However, in other applications we could expect that observational bias may be more predominant. We expect that the statistical tools presented in Section 3 could be useful in such a case.

A common way to model the bias introduced by combining different study types is by including a variance component which measures this feature. This is known as *grouped-random effects* in evidence synthesis literature [20]. This approach requires a large number of studies in order to assess between study type variability. In our application, with a few numbers of RCTs and a single observational study, the *grouped-random* effects method is not suitable. The model presented in this paper is an alternative for the *grouped-random* effects approach when few studies of different design have to be combined in a meta-analysis.

We conclude with some remarks on the limitations of the case study and the modeling approach presented in Section 3, and we point out areas that require further research as follows:

- The case study presented in this paper is a real one, but it is quite limited. In order to fully demonstrate the value of the proposed methods, we should apply them to examples where less similarities exist between the different sources of evidence. We are currently working in this line of research, and we expect to include not only therapeutic but also diagnostic and prognostic studies. As mentioned by one reviewer, effect modification and study quality will have a stronger influence in such scenarios, and the availability of IPD and modeling covariates in these data will become more relevant as is currently the case.
- We had the simplistic assumption that $\theta_{1,N+1}$ is independent of the observational bias ϕ and after performing a conflict of evidence analysis, we set $\mu_\phi = 0$. As mentioned by one of our reviewers, one may be surprised to see a positive correlation between these two parameters. Moreover, the bias direction in observational studies is uncertain. Deeks *et al*. [40] showed that non-random allocation can lead to overestimation or underestimation of treatment effects. In this regard, our case study is very limited. In order to further investigate these issues, we should combine several observational studies and RCTs with a *grouped-random effects* approach.
- We proposed to use the weights $w_1, \dots, w_{N+1}$ as a device to construct heavy-tailed distributions for the random effects. If the data allow us to learn from the posteriors of $w_1, \dots, w_{N+1}$, then, technically

speaking, these weights are measures of studies' misfit. In this paper, we argue that in the context of evidence synthesis these weights are linked with the study's internal validity. However, this is a conjecture that requires further research. In meta-analysis of diagnostic test data, Verde [11] found that observational studies tend to have over-dispersion, that is, systematically lower values of w_is.

- As pointed out by one reviewer, one line of research could be to assess risk of bias and quality assessment by existing tools and compare the resulting quality with the weights distributions obtained by our model. Further references regarding this line of research include issues in study design and risk of bias by Higgins *et al*. [41], issues relating to confounding factors when including non-randomized evidence by Valentine and Thompson [42], issues in selective reporting by Norris *et al.* [43], applicability of non-randomized evidence as a complementary source of evidence by Schuenemann [44], and a guideline of checklists for review authors by Wells [45].
- The method has limits when the correlation between θ_1 and θ_2 is low or their variances are low. In these cases, the predicted treatment effect would be unstable. A line of further research could be to use a meta-regression with common covariates between RCTs and observational data that may help to link treatment effect and individual participant characteristics. We are currently working on this problem as well.
- We have presented a multiplicity of biases. The question is as follows: how much does each bias affect the posterior distribution of any parameter of interest? The technical problem we have is the interconnection of all parameters in the model, which makes it difficult to isolate the influence of different types of bias in posteriors of a parameter of interest. To answer this question, we can use the DAG representation of the full statistical model presented in Figure 3. The advantage of having a DAG is that it allows us to calculate the influence of each component by restricting its influence during the simulation process. This is called *cutting feedback in a DAG* and is implemented in the statistical software *WinBUGS* and *OpenBUGS* [36] with the `cut` function. One further topic of research in our work is to use the cut function in order to automatically assess bias influence in the model.
- Finally, our work is just a first step in the complex problem of generalizing evidence. We did not cover important aspects such as modeling several observational studies together with RCTs' evidence, observational studies with different sets of risk factors, if the inclusion of individual data reduces ecological bias in meta-analysis, and the application of our approach to further clinical problems.

As this last section shows, we left a number of methodological questions unresolved. We hope that our work motivates statisticians to investigate further advantages and discovers new insights in combining aggregate and individual data in medical research.

Appendix: R script, BUGS model and further numerical results

In this appendix, we provide the R script and the main BUGS model used in Section 4.3 to combine IPD and AD from RCTs' results. At the end of the appendix, Table A.1 presents detailed results of the analysis reported in Section 4.3.

```
# Model for combining aggregated and individual data .........................
cat(
  "model
{
  # Model for aggregated data ...................................................
  for(i in 1: (N-1))
  {
  y.0[i] ~ dbin(p.0[i], n.0[i])
  y.1[i] ~ dbin(p.1[i], n.1[i])

  logit(p.0[i]) <- theta.1[i]
  logit(p.1[i]) <- theta.2[i] + logit(p.0[i])

  theta.1[i] ~ dnorm(mu.1, pre.theta.1[i])       # Shared random-effect
  theta.2[i] ~ dnorm(mu.2.1[i], pre.theta.2.1[i])

  lambda[i] ~ dchisqr(df)
  s[i] <- df/lambda[i]

  pre.theta.1[i] <- inv.sigma2.1 / s[i]
```

Table A.1. Summary results for the model used in Section 4.3 for $\mu_\phi = 0$.

Block	Parameters	mean	sd	2.5%	50%	97.5%
Hyperparameters	mu.1	-2.15	0.41	-2.97	-2.15	-1.31
	mu.obs	-2.15	0.41	-2.97	-2.15	-1.31
	mu.2	-0.74	0.53	-1.76	-0.74	0.35
	sigma.1	0.79	0.39	0.27	0.71	1.78
	sigma.2	1.33	0.97	0.09	1.06	3.67
	rho	-0.50	0.39	-0.98	-0.56	0.42
Conditional model of AD	beta.0	-0.74	0.53	-1.76	-0.74	0.35
	beta.1	-0.73	0.53	-1.76	-0.74	0.31
Regression coefficients of IPD	beta.first.ever.lesion[2]	0.75	0.32	0.16	0.74	1.42
	beta.pad[2]	0.47	0.31	-0.12	0.46	1.12
	beta.w[2]	0.92	0.36	0.25	0.91	1.65
Random effects	theta.1[1]	-2.19	0.24	-2.68	-2.18	-1.73
	theta.1[2]	-0.26	0.30	-0.87	-0.25	0.32
	theta.1[3]	-2.47	0.66	-3.96	-2.40	-1.36
	theta.1[4]	-2.25	0.43	-3.19	-2.23	-1.46
	theta.1[5]	-2.27	0.43	-3.20	-2.24	-1.50
	theta.1[6]	-2.16	0.56	-3.35	-2.13	-1.14
	theta.1[7]	-2.58	0.60	-3.89	-2.54	-1.53
	theta.2[1]	-0.89	0.41	-1.72	-0.88	-0.12
	theta.2[2]	-2.28	0.61	-3.53	-2.25	-1.16
	theta.2[3]	-0.54	0.85	-2.15	-0.58	1.28
	theta.2[4]	-0.75	0.64	-2.03	-0.75	0.53
	theta.2[5]	-0.44	0.61	-1.57	-0.47	0.84
	theta.2[6]	-0.55	0.75	-1.94	-0.58	1.05
	theta.2[7]	-0.43	1.06	-2.42	-0.49	1.86
Weights	weight[1]	1.23	0.76	0.22	1.07	3.12
	weight[2]	0.51	0.45	0.04	0.38	1.73
	weight[3]	1.11	0.72	0.18	0.96	2.91
	weight[4]	1.19	0.75	0.21	1.04	3.05
	weight[5]	1.19	0.75	0.21	1.03	3.06
	weight[6]	1.16	0.74	0.20	1.00	3.01
	weight[7]	1.05	0.71	0.16	0.89	2.83
Subgroups	mu.t.w	-1.40	0.72	-2.96	-1.35	-0.12
	mu.t.f	-1.28	0.65	-2.69	-1.24	-0.09
	mu.t.pda	-1.08	0.60	-2.35	-1.05	0.06
	pr.t.w	0.08	0.06	0.02	0.06	0.22
	pr.t.f	0.07	0.05	0.02	0.06	0.19
	pr.t.pda	0.07	0.04	0.02	0.06	0.15
	y.new.w	3.86	3.31	0.00	3.00	12.00
	y.new.f	8.35	6.00	1.00	7.00	23.00
	y.new.pda	9.84	6.23	2.00	9.00	24.00

Parameters are in BUGS notation and results organized in blocks of parameters, from the top to the bottom: Hyperparameters, intercept and slope of the conditional model for aggregated data (AD), regression coefficients of the main risk factors for individual participant data (IPD), random-effects, studies' weights and treatment effects for subgroups of patients.

```
pre.theta.2[i] <- inv.sigma2.2 / s[i]

#Conditional precision
pre.theta.2.1[i] <- pre.theta.1[i] / (1-rho*rho)

#Conditional mean
mu.2.1[i] <- mu.2 + rho * sigma.2 / sigma.1 * (theta.1[i] - mu.1)
}
```

```
# DF
df <- 4

# Priors marginal model ...

mu.1 ~ dlogis(0, 0.25)                        # Shared node
mu.2 ~ dlogis(0, 0.25)

inv.sigma2.1 <- 1/(sigma.1*sigma.1)
inv.sigma2.2 <- 1/(sigma.2*sigma.2)
sigma.1 ~ dunif(0, 4)                         # Shared node
sigma.2 ~ dunif(0, 4)

# Correlation
z ~ dnorm(0, 0.25)
rho <-  2*exp(z)/(1+exp(z)) - 1

# Model for individual data ............................................
# Random effect for the cohort study ...................................

mu.2.1[N] <- mu.2 + rho * sigma.2 / sigma.1 * (theta.1[N] - mu.obs)

theta.1[N] ~ dnorm(mu.obs, pre.theta.1[N])
theta.2[N] ~ dnorm(mu.2.1[N], pre.theta.2.1[N])

# Introduced mu.phi

# mu.phi  ~ dlogis(0, 0.25) # non-informative
# mu.phi ~ dnorm(0, 1)      # informative
  mu.phi <- 0               # non-conflict = 0;
                              or constant bias mu.phi > or < 0.
mu.obs <- mu.1 + mu.phi

# Structure of w[N].....................................................
# Case: same as RCTs
a <- df/2
b <- df/2

#Case: penalization at E(w[N])= k = 0.4
# b <- (df + 6)/2

lambda[N] ~ dgamma(a, b)
s[N]  <- 1/lambda[N]

# Case 2: full penalization
# s[N] <- 10

pre.theta.1[N]  <- inv.sigma2.1 / s[N]
pre.theta.2[N]  <- inv.sigma2.2 / s[N]
pre.theta.2.1[N] <- pre.theta.1[N] / (1-rho*rho)

# Weights
for(i in 1:N){weight[i] <- 1/s[i]}

# Regression model ......................................................
for( i in 1:M ){
y.0.pid[i] ~ dbern(p0.pid[i])

logit(p0.pid[i]) <- theta.1[N]                  # Shared random-effect
+ beta.w[w[i]]
+ beta.pad[pad[i]]
+ beta.neuropathy1[neuropathy1[i]]
+ beta.first.ever.lesion[first.ever.lesion[i]]
+ beta.no.continuous.care[no.continuous.care[i]]
+ beta.male[male[i]]
+ beta.diab.typ2[diab.typ2[i]]
+ beta.insulin[insulin[i]]
+ beta.HOCHD[HOCHD[i]]
+ beta.HOS[HOS[i]]
+ beta.CRF[CRF[i]]
+ beta.dialysis[dialysis[i]]
+ beta.DNOAP[DNOAP[i]]
+ beta.smoking.ever[smoking.ever[i]]
+ beta.age[age[i]]
+ beta.diabdur[diabdur[i]]
```

```
  }

  # Setup ...
  beta.w[1] <- 0
  beta.pad[1] <- 0
  beta.neuropathy1[1] <- 0
  beta.first.ever.lesion[1] <- 0
  beta.no.continuous.care[1] <- 0
  beta.male[1] <- 0
  beta.diab.typ2[1] <- 0
  beta.insulin[1] <- 0
  beta.HOCHD[1] <- 0
  beta.HOS[1] <- 0
  beta.CRF[1] <- 0
  beta.dialysis[1] <- 0
  beta.DNOAP[1] <- 0
  beta.smoking.ever[1] <- 0
  beta.age[1] <- 0
  beta.diabdur[1] <- 0

  # Priors ...
  #beta.0 ~ dnorm(0, 1)
  beta.w[2] ~ dnorm(0, pre.beta)
  beta.pad[2] ~ dnorm(0, pre.beta)
  beta.neuropathy1[2] ~ dnorm(0, pre.beta)
  beta.first.ever.lesion[2] ~ dnorm(0, pre.beta)
  beta.no.continuous.care[2] ~ dnorm(0, pre.beta)
  beta.male[2] ~ dnorm(0, pre.beta)
  beta.diab.typ2[2] ~ dnorm(0, pre.beta)
  beta.insulin[2] ~ dnorm(0, pre.beta)
  beta.HOCHD[2] ~ dnorm(0, pre.beta)
  beta.HOS[2] ~ dnorm(0, pre.beta)
  beta.CRF[2] ~ dnorm(0, pre.beta)
  beta.dialysis[2] ~ dnorm(0, pre.beta)
  beta.DNOAP[2] ~ dnorm(0, pre.beta)
  beta.smoking.ever[2] ~ dnorm(0, pre.beta)
  beta.age[2] ~ dnorm(0, pre.beta)
  beta.diabdur[2] ~ dnorm(0, pre.beta)

  pre.beta <- 1/(sigma.beta*sigma.beta)
  sigma.beta ~ dunif(0, 2)

  # Further parameters of interest ............................................

  # Functional parameters for plotting the conditional model: (mu.2 | mu.1)...
  beta.0 <- mu.2
  beta.1 <- rho * sigma.2/sigma.1

  # Mean controls wagner/first lesion/pad group...
  mu.c.w <- mu.1 + mu.phi + beta.w[2]                  # Location on the x-axis
                                                         of Wagner
  mu.c.f <- mu.1 + mu.phi + beta.first.ever.lesion[2]  # Location on the x-axis
                                                         of first ever lesion
  mu.c.pda <- mu.1 + mu.phi + beta.pad[2]              # Location on the x-axis
                                                         of pad

  # Mean treatment groups...
  mu.t.w <- mu.2 + beta.1 * beta.w[2]
  mu.t.f <- mu.2 + beta.1 * beta.first.ever.lesion[2]
  mu.t.pda <- mu.2 + beta.1 * beta.pad[2]

  # Predictive probability of amputation if treatment applyed ...
  pr.t.w <- exp(mu.c.w + mu.t.w)/(1+exp(mu.c.w + mu.t.w))
  pr.t.f <- exp(mu.c.f + mu.t.f)/(1+exp(mu.c.f + mu.t.f))
  pr.t.pda <- exp(mu.c.pda + mu.t.pda)/(1+exp(mu.c.pda + mu.t.pda))

  # Predictive number of amputations per group ...
  y.new.w ~ dbin(pr.t.w, 49)
  y.new.f ~ dbin(pr.t.f, 114)
  y.new.pda ~ dbin(pr.t.pda, 148)

}
  ", file = "Full_new.bug")

mfull.new <- bugs(data.full.new, inits=NULL, par.full.new,
"Full_new.bug",
n.chains = 2,
n.iter = 50000,
n.thin=1,
bugs.directory = bugsdir,
n.burnin=floor(25000),
working.directory = getwd(),
clearWD = FALSE, debug = F)
```

Acknowledgements

We are indebted to Nicky Best who supported our DFG project on generalized evidence synthesis while she was head of the BIAS project at the Imperial College in London. This work was presented by Pablo Emilio Verde in the centenary celebration of the MRC Biostatistics Unit in Cambridge in 2014, who is grateful to Chris Jackson for a fruitful exchange of ideas and to Andrea Riebler and Sandro Gsteiger for their enthusiastic interest in our work. We also thank the associate editor for his comments and two reviewers, who helped us to greatly improve the quality of this paper. This work was supported by the German Research Foundation projects DFG Oh 39/11-1 and DFG VE 986/1-1.

References

1. Lugtenberg M, Burgers J, Clancy C, Westert G, Schneider EC. Current guidelines have limited applicability to patients with comorbid conditions: A systematic analysis of evidence-based guidelines. *PLoS One* 2011; **6**(10):1–7.
2. van Weel C, Schellevis FG. Comorbidity and guidelines: conflicting interests. *Lancet* 2006; **367**(9510):550–551.
3. McIntosh M. The population risk as an explanatory variable in research synthesis of clinical trials. *Statistics in Medicine* 1996; **15**(16):1713–1728.
4. Thompson S, Smith T, Sharp S. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741–2758.
5. Sharp S, Thompson S. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Statistics in Medicine* 2000; **19**(23):3251–3274.
6. Arends L, Hoes A, Lubsen J, Grobbee D, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497–3518.
7. Guolo A. Flexibly modeling the baseline risk in meta-analysis. *Statistics in Medicine* 2013; **32**(1):40–50.
8. Guolo A. The SIMEX approach to measurement error correction in meta-analysis with baseline risk as covariate. *Statistics in Medicine* 2014; **33**(12):2062–2076.
9. Ghidey W, Lesaffre E, Stijnen T. Semi-parametric modelling of the distribution of the baseline risk in meta-analysis. *Statistics in Medicine* 2007; **26**(30):5434–5444.
10. Verde PE, Curcio D. Imbalance mortality evidence for tigecycline. *Clinical Infectious Diseases* 2012; **55**(3):471–472.
11. Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Statistics in Medicine* 2010; **29**(30):3088–3102.
12. Spiegelhalter D, Best N. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* 2003; **22**:3687–3709.
13. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006; **25**(1):2136–2159.
14. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; **171**(1):159–178.
15. Jackson C, Best N, Richardson S. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics* 2009; **10**(1):335–351.
16. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* 2008; **27**(11):1870–1893.
17. Sutton A, Kendrick D, Coupland C. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* 2008; **27**(5):651–669.
18. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care* 1993-01; **9**(3):440–449.
19. Kaizar E. Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine* 2011-11; **30**(25):2986–3009.
20. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* 2015; **6**(1):45–62.
21. Hoffmann F, Claessen H, Morbach S, Waldeyer R, Glaeske G, Icks A. Impact of diabetes on costs before and after major lower extremity amputations in Germany. *Journal of Diabetes and its Complications* 2013; **27**(5):467–472.
22. Schofield C, Libby G, Brennan G, MacAlpine R, Morris A, Leese G, Collaboration D. Mortality and hospitalization in patients after amputation. *Diabetes Care* 2006; **29**:2252–2256.
23. Icks A, Scheer M, Morbach S, Genz J, Haastert B, Giani G, Glaeske G, Hoffmann F. Time-dependent impact of diabetes on mortality in patients after major lower extremity amputation: Survival in a population-based 5-year cohort in Germany. *Diabetes Care* 2011; **34**(6):1350–1354.

24. Apelqvist J, Bakker K, Van Houtum W, Schaper NC. Practical guidelines on the management and prevention of the diabetic foot. *Diabetes/Metabolism research and reviews* 2008; **24 (Suppl 1)**:181–187.
25. Chuck A, Hailey D, Jacobs P, Perry D. Cost-effectiveness and budget impact of adjunctive hyperbaric oxygen therapy for diabetic foot ulcers. *International Journal of Technology Assessment in Health Care* 2008; **24**:178–183.
26. Habacher W, Rakovac I, Goerzer E, Haas W, Gfrerer J, Wach P, Pieber TR. A model to analyse costs and benefit of intensified diabetic foot care in Austria. *Journal of Evaluation in Clinical Practice* 2007; **13**:906–912.
27. Centre for Clinical Practice at NICE (UK and others). Clinical guideline 119. Diabetic foot problems: Inpatient management of diabetic foot problems, National Institute for Health and Clinical Excellence, London, 2011.
28. Verde PE. A comment mentioning possible applications in meta-analysis of dirichlet t-distributions. *Bayesian Analysis* 2014; **9**(3):589–590.
29. Morbach S, Furchert H, Groeblinghoff U, Hoffmeier H, Kersten K, Klauke GT, Klemp U, Roden T, Icks A, Haastert B, Rümenapf G, Abbas ZG, Bharara M, Armstrong DG. Long-term prognosis of diabetic foot patients and their limbs: Amputation and death over the course of a decade. *Diabetes Care* 2012; **35**(10):2021–2027.
30. Wells GA, Shea B, OConnell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2008. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
31. Gelman A. Prior distribution for variance parameters in hierachical models. *Bayesian Analysis* 2006; **1**:1–6.
32. Rockova V, Lesaffre E, Luime J, Loewenberg B. Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine* 2012; **31**(11-12):1221–1237.
33. Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 03* 2010; **5**(1):171–188.
34. Marshall EC, Spiegelhalter DJ. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis* 2007; **2**:409–444.
35. Presanis AM, Ohlssen D, Spiegelhalter D, De Angelis D. Conflict diagnostic in directed acyclic graphs, with applications in bayesian evidence synthesis. *Statistical Science* 2013; **28**(3):376–397.
36. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* 2009; **28**(25):3049–3067.
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2013.
38. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 2005; **12**(3):1–16.
39. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* 2006; **6**(1):7–11.
40. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakrovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. *Health Technology Assessment NHS R&D HTA Programme* 2003; **7**(27):1–173.
41. Higgins J, Ramsay C, Reeves B, Deeks J, Shea B, Valentine J, Tugwell P, Wells G. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**(1):12–25.
42. Valentine J, Thompson S. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**(1):26–35.
43. Norris S, Moher D, Reeves B, Shea B, Loke Y, Garner S, Anderson, Tugwell P, Wells G. Issues relating to selective reporting when including non-randomized studies in systematic reviews on the effects of healthcare interventions. *Research Synthesis Methods* 2013; **4**(1):36–47.
44. Schuenemann HJ, Tugwell P, Reeves B, Akl EA, Santesso N, Spencer FA, Shea B, Wells G, Helfand M. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**(1):49–62.
45. Wells GA, Shea B, Higgins J, Sterne J, Tugwell P, Reeves BC. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Research Synthesis Methods* 2013; **4**(1):63–77.

Chapter 5

A Comment on Conflict of Evidence

> ”There must be, he thought, some key, some crack in this mystery he could use to achieve an answer”.
>
> -P.C. Doherty, Crown in Darkness.

Comment by Pablo E. Verde[1]

My congratulations to the authors for this interesting paper. I found the extension of the classical t-distribution by using a scale mixture of normal distributions per coordinate quite useful in practice and the Dirichlet t-distribution an elegant approach. I would like to make the following practical comments:

Statistical inference of the parameter α in the Dirichlet t-distribution looks challenging. The authors Michael Finegold and Mathias Drton applied two strategies: one by fixing α to different values and another one by applying a Gamma prior distribution with parameters equal to 1, which gives a prior $E(\alpha) = 1$. In applications, I would recommend to make a prior to posterior analysis of this parameter in order to understand if we could learn something about α from the data at hand. The same strategy should be applied to the degrees of freedom parameter ν.

In my work in multi-parameters meta-analysis (Verde 2010; Verde and Sykosch 2011) I found that the single component scale mixture is useful enough for outliers' identification and for down-weighting pieces of evidence with unusual results. However, the introduction of the Dirichlet t-distribution opens an interesting possibility in the detection of conflict of evidence in meta-analysis and in the detection of structural outliers in Bayesian hierarchical modeling.

The conflict assessment is the deconstructionist side of meta-analysis, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. One possibility for this type of analysis is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. For example in Verde et al. (2014), we applied a scale mixture of multivariate normal distributions in a meta-analysis combining randomized and non-randomized evidence and we made conflict diagnostics by direct interpretation of the scale weights. Another alternative is presented by Presanis et al. (2013), where the authors described how to generalize the conflict p-value proposed by Marshall and Spiegelhalter (2007) to complex evidence modeling. In summary, by using a Dirichlet t-distribution conflict of evidence can be generalized and performed for each parameter in a multi-parameter meta-analysis.

References

Marshall, E. C. and Spiegelhalter, D. J. (2007). "Identifying outliers in Bayesian hierarchical models: a simulation-based approach." *Bayesian Analysis*, 2: 409–444.

Presanis, A. M., Ohlssen, D., Spiegelhalter, D., and Angelis, D. D. (2013). "Conflict diagnostic in directed acyclic graphs, with applications in Bayesian evidence synthesis." *Statistical Science*, 28: 376–397.

Verde, P. E. (2010). "Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine*, 30(29): 3088–3102.

[1] Coordination Center for Clinical Trials, University of Duesseldorf, Germany, pabloemilio.verde@hhu.de

Verde, P. E., Ohmann, C., Icks, A., and Morbach, S. (2014). "Bayesian evidence synthesis and combining randomized and nonrandomized results: a case study in diabetes." *Statistics in Medicine*, (under review).

Verde, P. E. and Sykosch, A. (2011). "bamdit: Bayesian meta-analysis of diagnostic test data." *CRAN: R package version 1.1-1*.

Chapter 6

A Bayesian Model for Meta-Analysis of Diagnostic Test Data

> ”Why does he insist that we must have a diagnosis? Some things are not meant to be known by man.”
>
> -Susanna Gregory, An Unholy Alliance.

Research Article

Statistics in Medicine

Received 21 September 2009, Accepted 22 July 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4055

Meta-analysis of diagnostic test data: A bivariate Bayesian modeling approach

Pablo E. Verde*†

In the last decades, the amount of published results on clinical diagnostic tests has expanded very rapidly. The counterpart to this development has been the formal evaluation and synthesis of diagnostic results. However, published results present substantial heterogeneity and they can be regarded as so far removed from the classical domain of meta-analysis, that they can provide a rather severe test of classical statistical methods. Recently, bivariate random effects meta-analytic methods, which model the pairs of sensitivities and specificities, have been presented from the classical point of view. In this work a bivariate Bayesian modeling approach is presented. This approach substantially extends the scope of classical bivariate methods by allowing the structural distribution of the random effects to depend on multiple sources of variability. Meta-analysis is summarized by the predictive posterior distributions for sensitivity and specificity. This new approach allows, also, to perform substantial model checking, model diagnostic and model selection. Statistical computations are implemented in the public domain statistical software (WinBUGS and R) and illustrated with real data examples. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: meta-analysis; diagnostic studies; posterior prediction; hierarchical models; Bayesian modeling; BUGS

1. Introduction

The first crucial information in the presence of illness is a medical diagnosis. How good or bad a diagnosis is performed may directly influence the quality of the health care. Accurate evaluation of diagnostic tests contributes to the prevention of unjustified treatment, as well as unnecessary health costs. In the last decades, the amount of published results on clinical diagnostic tests, their systematic reviews and meta-analysis has expanded very rapidly [1].

In this paper we concentrate on the most common meta-analysis of binary diagnostic outcome, where results for the ith study ($i=1,\dots,N$) are summarized in a 2×2 table as follows:

		Patient status	
		With disease	Without disease
Test	+	tp_i	fp_i
outcome	−	fn_i	tn_i
Sum:		$n_{i,1}$	$n_{i,2}$

where tp_i and fn_i are the number of patients with positive and negative diagnostic results in the group with disease and fp_i and tn_i are the number of patients with positive and negative test results in the group without disease, respectively. The total number of patients with disease is $n_{i,1}=tp_i+fn_i$ and the total number of patients without disease is $n_{i,2}=fp_i+tn_i$. Common summary statistics describing test accuracy can be estimated for each study, the most commonly used are the empirical *true positive rate* or *sensitivity* and the empirical *true negative rate* or *specificity*,

$$\widehat{\mathrm{TPR}}_i=\frac{tp_i}{n_{i,1}},\quad \widehat{\mathrm{TNR}}_i=\frac{tn_i}{n_{i,2}} \tag{1}$$

Coordination Center for Clinical Trials, University of Düsseldorf, Moorenstr. 5, D-40225, Germany

*Correspondence to: Pablo E. Verde, Coordination Center for Clinical Trials, University of Düsseldorf, Moorenstr. 5, D-40225, Germany.

†E-mail: pabloemilio.verde@uni-duesseldorf.de

and their complementary empirical rates, *the false positive rate* ($\widehat{\text{FPR}}$) and *the false negative rate* ($\widehat{\text{FNR}}$),

$$\widehat{\text{FPR}}_i = \frac{fp_i}{n_{i,2}}, \quad \widehat{\text{FNR}}_i = \frac{fn_i}{n_{i,1}}. \tag{2}$$

The main question we investigate in this paper is: How can we combine and summarize these types of diagnostic information?

To start answering this question, we should note that meta-analysis of diagnostic test data differs from other types of meta-analysis, in at least three aspects: First, summaries describing the test accuracy (e.g. sensitivity and specificity) are usually interdependent and a marginal combination by averaging or pooling these quantities might be misleading [2, 3]. Second, the context where diagnostic studies have been performed can be very different in terms of diagnostic setup, study design, population characteristics or study quality [4–6]. Third, it usually involves a small sample of studies.

These issues have contributed to making this area of meta-analysis a very active area of methodological research. On the one hand, methods for searching and assessing the quality of published studies have been established [7] and on the other hand a large amount of diverse statistical methods have been developed. Moses *et al.* [8] introduced a simple fixed-effects meta-regression model, which summarizes study results by an imaginary curve, called the Summary Receiving Operation (SROC) curve. This method has been extensively used in the medical literature and has been recently recommended by the Cochrane Diagnostic Test Accuracy Working Group [9]. Well-known limitations of the SROC curve have motivated alternative meta-regression models [10], the development of a full Bayesian model, called the hierarchical SROC curve (HSROC) [11, 12], and its empirical Bayesian version [13]. Further research work under the SROC curve has been done, which includes a better understanding on its properties [14, 15], meta-analysis of diagnostic test with imperfect reference standard [16] and publication bias [17].

As an alternative to the SROC curve approach, several authors have proposed to jointly model sensitivity and specificity with bivariate random effects models ([18–21], (Arends, unpublished)). Relationships between SROC and bivariate approaches have been investigated [22, 23]. More recently alternative parametrization for the bivariate approach have been proposed [24, 25], recommendations of summarizing a meta-analysis by ROC curves or not has been presented [26], Bayesian computations based on integrated nested Laplace approximations have been investigated [27, 28] and the inclusion of individual patient data with bivariate random effects model has been analyzed [29].

The aim of this paper is to present a flexible class of new statistical models to deal with meta-analysis of diagnostic test data. We construct a hierarchical Bayesian model that realistically reflects the underlying complexities of these types of data. This model can be regarded as a general version of the Bayesian bivariate random effects meta-analysis, where heterogeneity between studies is modeled by a rich class of structural distributions based on scale bivariate normals.

The remainder of this paper is organized as follows: In Section 2 we briefly introduce our running examples, in Section 3 we present the bivariate Bayesian model for meta-analysis of diagnostic test data, in Section 4 this model is applied to our running examples and results compared with other approaches are reported. Finally, in Section 5 we discuss and outline some areas for further work.

2. Examples

Figure 1 displays the pairs ($\widehat{\text{TPR}}_i, \widehat{\text{FPR}}_i$) of our two running examples. The left panel presents results of 52 studies reporting the accuracy of computer tomography (CT) scans in the diagnosis of appendicitis [30] calculated from Table I.

This disease is one of the most common acute surgical events [31], where the traditional clinical examination delivers low diagnostic performance [32]. Therefore, a new diagnostic technology could reduce the risk of postoperative complications and save health-care resources [33]. One of the main research interests of this systematic review was to give overall measurements of diagnostic accuracy of CT technology. Another one was to explore study characteristics or published information that may influence diagnostic results. Standardized data extraction forms were used to collect the papers' results and to assess the quality information [34]. Table II gives some variables describing study characteristics: (*Country, Type of hospital*), patients characteristics (*Inclusion criteria, Children included*), study quality (*Design*) and diagnostic setup (*Contrast medium, Localization*). More details regarding the list of databases and searching templates are described in the original technical report [30]. In Section 4 we analyze these explanatory variables with a meta-regression model to understand how the published information may influence diagnostic results.

The right panel of Figure 1 shows diagnostic results of 10 studies reporting the accuracy of magnetic resonance imaging (MRI) for the diagnosis of lymph node metastasis in women with cervical cancer (Table III of Scheidler *et al.* [35]). This systematic review has been used in the methodology literature on meta-analysis of diagnostic test by several authors (Rutter and Gatsonis [11], Walter [14, 15], Macaskill [13], Reistma *et al.* [18], Gatsonis and Paliwal [3], Chu and Cole [20], Chu and Gou [23], Chu *et al.* [25] and Martino and Rue [28]) and gives a good example to compare

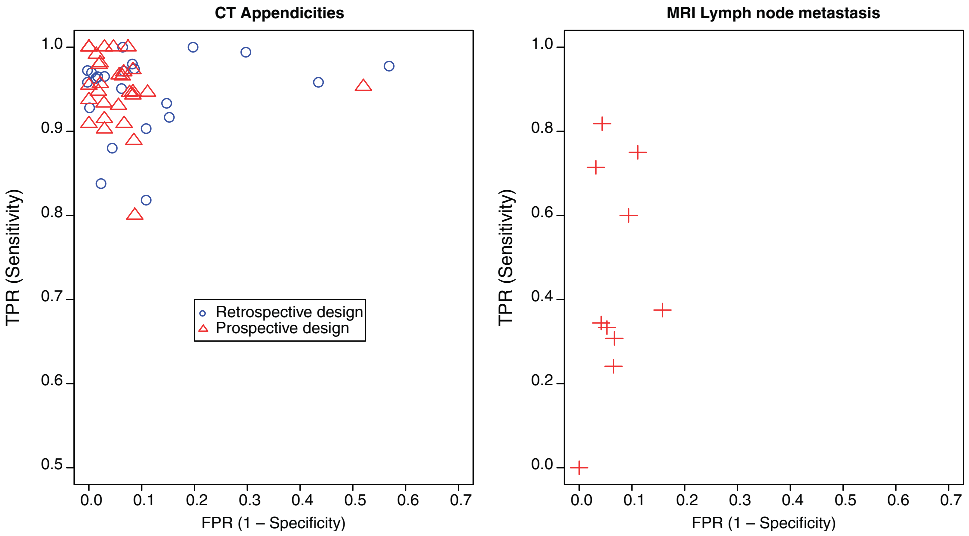


Figure 1. Two examples of meta-analysis of diagnostic test data. *Left panel*: 52 studies, CT scans in the diagnosis of appendicitis. *Right panel*: 10 studies, magnetic resonance imaging (MRI) for the diagnosis of lymph node metastasis in women with cervical cancer.

our approach to other meta-analytic methods whenever possible. More details on these two examples are described in Section 4.

3. A Bayesian framework for meta-analysis of diagnostic test

In Section 1 we highlighted issues involved in the analysis and synthesis of diagnostic test results. In this section we introduce a novel Bayesian modeling framework, which incorporate these data complexities in a very practical way.

3.1. Data model and structural distributions

Following the notation of the introduction, let tp_i and fp_i be the true positive and false positive results for study i $(i=1,\ldots,N)$. Conditioning on $n_{i,1}$ and on $n_{i,2}$ our *data model* consists of two binomial distribution with

$$tp_i \sim \mathrm{Bin}(\mathrm{TPR}_i, n_{i,1}), \qquad fp_i \sim \mathrm{Bin}(\mathrm{FPR}_i, n_{i,2}), \tag{3}$$

where TPR_i and FPR_i are the probabilities to observe a positive test result in the disease and non-disease population respectively. The N pairs of probabilities TPR_i and FPR_i are transformed by a link function $g(\cdot)$ into a scale where they are defined in the range $(-\infty,\infty)$. The canonical link function for binomial data is the logit link function $(g(p)=\log(p/(1-p)))$, but other alternative links, e.g. the complementary log–log $((g(p)=\log(-1\log(1-p))))$ link function can be used. Choosing a suitable link function for the data at hand is a modeling problem that will be illustrated in Section 4. In general, the link function should be chosen to give a parsimonious model to fit the data.

We model the variability between studies by defining *the study accuracy effects*, which are differences and sums of the rates in the $g(\cdot)$ scale:

$$D_i = g(\mathrm{TPR}_i) - g(\mathrm{FPR}_i), \qquad S_i = g(\mathrm{TPR}_i) + g(\mathrm{FPR}_i), \tag{4}$$

where D_i and S_i are modeled with a *scale mixture of bivariate Normal distributions*

$$(D_i, S_i) \sim N(\mu, \Psi_i), \qquad i=1,2,\ldots,N, \tag{5}$$

$$\Psi_i = w_i \times \Lambda, \tag{6}$$

$$w_i \sim p(w_i) \tag{7}$$

with Λ the precision matrix, i.e. the inverse of the covariance matrix Σ and $p(w_i)$ a scale mixing density. Modeling (D_i, S_i) is similar to direct modeling $(g(\mathrm{TPR}_i), g(\mathrm{FPR}_i))$; however, the linear transformation should leave (D_i, S_i) roughly independent making our inference less sensitive to the prior distribution of Λ.

Table I. Cross-classified tables for 52 studies reporting diagnostic results of CT scans used to diagnose appendicitis: id refers to study identification number; tp_i and fp_i are the number of patients with positive diagnostic results in the group with disease; fn_i and tn_i are the number of patients with negative test results in the group without disease; (R) and (P) indicate retrospective and prospective study design respectively. Study numbers 31 and 32 correspond to two different papers with the same data, both papers were included in the review, but only one is used for the statistical analysis.

Study	*tp*	*fp*	*fn*	*tn*
(R)1	87	4	2	3
(R)2	111	1	4	30
(R)3	184	7	8	9
(R)4	168	3	1	7
(P)5	89	3	5	33
(R)6	21	1	0	14
(R)7	125	3	0	12
(P)8	40	4	5	43
(P)9	40	0	4	8
(R)10	104	3	4	185
(R)11	34	4	1	54
(P)12	29	4	1	66
(P)13	28	5	1	74
(P)14	67	7	5	118
(P)15	30	3	3	42
(P)16	36	1	1	11
(P)17	131	12	4	170
(R)18	23	0	1	76
(P)19	4	2	0	25
(R)20	31	2	6	76
(P)21	110	4	5	181
(P)22	37	2	4	66
(P)23	18	1	1	55
(R)24	35	0	1	36
(P)25	28	6	7	63
(R)26	9	2	2	16
(R)27	38	8	1	82
(R)28	64	1	2	128
(R)29	103	1	8	252
(P)30	88	3	5	24
(R)31	137	8	5	402
(R)32	137	8	5	402
(P)33	32	2	0	66
(P)34	114	3	1	211
(P)35	52	1	1	46
(P)36	17	0	0	18
(P)37	56	2	0	41
(R)38	49	4	1	43
(R)39	58	6	3	87
(P)40	21	0	1	34
(R)41	33	11	3	60
(R)42	44	3	6	61
(P)43	43	5	4	166
(P)44	4	0	0	22
(R)45	28	8	3	64
(P)46	30	0	2	25
(P)47	47	1	1	51
(P)48	183	26	9	24
(P)49	28	2	2	68
(P)50	33	3	2	33
(P)51	35	1	2	12
(R)52	42	3	3	17

In this context, D_i is the study effect associated with diagnostic discriminatory power and S_i is the effect associated with diagnostic threshold value. Positive values of D_i indicate the power discrimination of the diagnostic procedure, whereas positive values of S_i indicate good sensitivity at the expense of increase of the FPR_i.

The scale mixing density $p(w_i)$ introduces great flexibility to model the marginal distribution of study effects (D_i, S_i), some particular cases with heavy tails are the bivariate t-distribution which corresponds to $w_i \sim \Gamma(\nu/2, \nu/2)$ with known

Table II. List of covariates describing study characteristics,patients characteristics, study quality and diagnostic setup.

Notation	Variable name	Label	Value description
x_1	Country	EU and others/U.S.A.	0/1
x_2	Type of hospital	University/others	0/1
x_3	Inclusion criteria	Suspected/appendectomy	0/1
x_4	Other CT findings included	No/yes	0/1
x_5	Study design	Prospective/retrospective	0/1
x_6	Contrast medium	No/yes	0/1
x_7	Localization	One area/more than one area	0/1
x_8	Children included	No/yes	0/1

Table III. Notation and parameter names for the Bayesian bivariate model based on scale mixture of Normals.

Notation	Parameter
tp_i	Frequency of true positive patients
fp_i	Frequency of false positive patients
$n_{i,1}$	Total number of disease patients in the study
$n_{i,2}$	Total number of non disease patients in the study
TPR_i	True positive rate of study i
FPR_i	False positive rate of study i
D_i	Study accuracy effect (difference of the link function of TPR and FPR)
S_i	Study threshold effect (sum of the link function of TPR and FPR)
w_i	Study mixture weight
μ_D	Mean of D_i
μ_S	Mean of S_i
Λ	Precision matrix of (D_i, S_i)
σ_D^2	Variance of D_i
σ_S^2	Variance of S_i
$\sigma_{D,S}$	Covariance (D_i, S_i)
v	Degrees of freedom

degrees of freedom $v>2$, the Cauchy distribution with $v=1$ and the Double Exponential distribution with $w_i \sim \text{Exp}(1)$ (see Carlin and Louis [36], p. 184). The bivariate Normal corresponds to $w_i = 1$ when $p(w_i)$ is degenerated at one. In this work, we pay particular attention to the bivariate t-distribution, which is extended to include uncertainty on v and to explain systematic variability (see Section 3.5). Table III summarizes the notation involved in our model.

3.2. Priors

One important component of a Bayesian statistical model is the use of prior distributions that can be applied as a starting point of analysis in similar modeling situations, with this aim in this work we apply weakly informative priors. The use of weakly informative priors introduces numerical stabilization by discarding unrealistic parameter values, while still being vague enough to be used as a default in routine applied work. We also use priors that are conditionally conjugate to the hyper-parameters. These priors can be interpreted in terms of equivalent data, which can simplify elicitation of their parameters in a particular context where more substantial information exists.

The following priors are given for hyper-parameters μ_D, μ_S, Λ and v:

- The parameters μ_D and μ_S are odds ratios on the scale defined by $g(\cdot)$. We can expect that in a typical application where a logit link is used these parameters are in the range of -5 to 5. Then we use independent Normal priors with 0 mean and precision 0.25, these priors cover μ_D and μ_S within these range with 99 per cent probability and they give low probability to very extreme values.
- We represent vague information of the precision matrix Λ with a Wishart distribution with identity scale matrix and three degrees of freedom. This prior gives a uniform distribution for the correlation between (D_i, S_i) and two χ_3^2 for the precision parameters.

- For the degrees of freedom parameter we use $\nu \sim \text{Exp}(1)$. This prior distribution favors lower values of ν and strongly penalizes extreme values of (D_i, S_i).

3.3. Interpretation of the mixture weights

The use of the mixture weights w_i is in principle a mathematical device to construct heavy tails distributions and a data augmentation technique that is useful for computations. However, one crucial issue in meta-analysis is the ability of identifying studies that may influence results and are not simple to find *a priori*. For this reason we are going to give a direct interpretation to the weights w_i.

In this work we use the posterior distribution of w_i as an indication of model misfit or an identification of studies with unusual heterogeneity. *A priori* all studies included in the review have a prior mean of w_i equal to 1, studies which are unusual heterogeneous will have posteriors with values substantially less than 1, say less than 0.7.

To gain more insight into the interpretation of weights w_i we can see that under the normal distribution, where $w_i = 1$ $(i = 1, \ldots, N)$, studies are considered as exchangeable, i.e. the study label carries no relevant modeling information. However, if the w_is are unknown then studies are modeled as partial-exchangeable given what we can learn from w_i.

The presence of weights w_is with posterior distributions not concentrated at 1 indicates lack of exchangeability or misfit with respect to the normal distribution, which can be further investigated by using meta-regression (Section 3.4) or structural dispersion modeling (Section 3.5).

3.4. Meta-regression

A meta-regression model can be useful to analyze the impact of published information, like study characteristics or population differences, into diagnostic accuracy. However, it is worth mentioning some limitations of meta-regression methods: results are susceptible to aggregation or ecological bias, which occurs when study results and published populations' summaries do not directly reflect the relationship between patients' characteristics and patients' diagnostic outcomes. In addition, the published available data for analysis may be limited. In synthesis, meta-regression analysis should be interpreted as an explorative approach where results may be useful to suggest further investigation.

It is easy to include a regression structure to analyze systematic influence of variables in diagnostic results. We write $(\mu_{i,D}, \mu_{i,S})$, the fixed effects of the model, as a system of two regression equations,

$$\mu_{i,D} = \alpha_0 + \alpha_1 x_{i,1} + \cdots + \alpha_p x_{i,p}, \tag{8}$$

$$\mu_{i,S} = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}, \tag{9}$$

where each equation depends on a known p-dimensional vector of covariates $(x_{i,1}, \ldots, x_{i,p})$ and $2(p+1)$ unknown regression coefficients $(\alpha_0, \alpha_1, \ldots, \alpha_p)$ and $(\beta_0, \beta_1, \ldots, \beta_p)$.

Prior distributions for the regression coefficients are essential ingredients in any Bayesian analysis, they encapsulate a variable selection strategy with an implicit regularization technique. For example, modeling α_j and β_j as exchangeable with normal priors results in a Bayesian version of ridge regression and using Laplace priors results in a Bayesian version of the Lasso regression [37]. As suggested by the referee, we model α_j and β_j as exchangeable with a Gaussian distribution,

$$\alpha_0, \ldots, \alpha_p, \beta_0, \ldots, \beta_p \sim N(0, \phi). \tag{10}$$

For the precision parameter ϕ we use a uniform prior between 0 and 10 on the standard deviation $1/\sqrt{(\phi)}$.

3.5. Combining studies with different designs and structural dispersion

Usually, meta-analysis includes studies with different design (e.g. retrospective and prospective designs), which is also known as *generalized evidence synthesis*. Including different study designs in a meta-analysis may extend the inferential scope, e.g. the spectrum of the population under study at the cost of increasing the data heterogeneity.

Given the uncontrolled context where the data of retrospective studies are obtained, we may expect that retrospective studies present substantially more variability than prospective ones. One way to quantify this feature is by adding a systematic structure to the weights w_i as follows:

$$w_i \sim \text{Gamma}(\rho, \mu_{\text{w}_\text{i}}), \tag{11}$$

$$\log(\mu_{w_i}) = \gamma_0 + \gamma_1 x_i, \tag{12}$$

where x_i is an indicator variable with

$$x_i = \begin{cases} 0 & \text{Prospectivedesign,} \\ 1 & \text{Restrospectivedesign.} \end{cases}$$

As in the previous sections we use weakly informative priors for $(\rho, \gamma_0, \gamma_1)$. We give an exponential prior for shape parameter $\rho \sim \text{Exp}(1)$. Alternatively, we may set *a priori* $\rho = 1$ and model w_i as exponential with parameter μ_{w_i}. Given that we do not expect a design effect greater than 5 on the logarithmic scale we set two normal priors for the regression parameters $\gamma_0 \sim \text{N}(0, 0.25)$ and $\gamma_1 \sim \text{N}(0, 0.25)$.

The regression parameter γ_1 accesses the variability introduced by studies with retrospective design, a posterior of $\exp(\gamma_1)$ concentrated on values greater than 1 indicates an increase in variability. Another way to access an increase in variability is presented by Verde [38]. This technique consists of modeling the variance matrices of the random effects for each study design separately and calculating a measure of excess of variability, e.g. the ratio of the generalized variances. In our applications, we found that modeling directly w_i gives a better model fitness than fitting variance matrices separately.

3.6. Summary quantities of interest

We report for the parameters in the model $(\mu_D, \mu_S, \sigma_D^2, \sigma_S^2, \rho_{D,S}, v)$ posteriors means and percentiles (2.5 per cent, 50 per cent, 97.5 per cent) for the marginal posterior distributions.

The overall diagnostic accuracy is summarized by the posterior distribution of the functional parameters:

$$\text{Sensitivity(pooled)} = g^{-1}[(\mu_D + \mu_S)/2], \quad \text{Specificity(pooled)} = 1 - g^{-1}[(\mu_S - \mu_D)/2], \tag{13}$$

which give an internal meta-analytic summary, and by their marginal *predictive posteriors* $p(\text{Sensitivity(predicted)}|Data)$ and $p(\text{Specificity(predicted)}|Data)$, which predict results of a future study. These predictive summaries are the most important and completed statistical inference that can be drawn from the meta-analysis [39].

For the meta-regression and structural dispersion components we report numerical summaries as above and a forest-plot for the regression coefficients based on the percentiles (2.5 per cent, 50 per cent, 97.5 per cent) of their posteriors.

3.7. Model checking

As usual in practice, we cannot guarantee that a fitted model is correct. The basic approach used here for model checking is to simulate predictive values from the fitted model and comparing these quantities with the observed ones. This technique has been extensively used in Bayesian data analysis [40–44].

In this paper we recommend the building of a scatter plot by simulating predictive values $(\text{TPR}^*, \text{FPR}^*)$ from $p(\text{TPR}^{\text{pred}}, \text{FPR}^{\text{pred}}|Data)$ and by comparing these pairs with the estimated values directly calculated from the diagnostic tables $(\widehat{\text{TPR}}_i, \widehat{\text{FPR}}_i)$. Although this method is quite simple, this visual devise usually spots out deficits of the fitted model, see in Section 4. Then, the model can be updated by changing the link function, using different structural distributions or by adding covariates. This modeling process is monitored by reporting the DIC (Deviance Information Criterion) [45] a measure based on a trade off between goodness of fit and model complexity. Models with smaller DIC are better supported by the data in the sense of short-term predictions.

3.8. Statistical computations and software implementation

All these marginal posteriors and predictive distributions, which are presented in this section are not analytically tractable. We based our inference on MCMC techniques implemented in the WinBUGS package [46] and linked to R [47] with the R2WinBUGS package [48]. More computational details are given in the Appendix.

4. Data analysis

In the examples presented in this section we used the following computational setup: Calculations are based on five chains with random starting values and with 20 000 replications. The last 10 000 iterations are used for analysis. Convergence checking was carefully performed by visual analysis of trace plots and empirical autocorrelation functions, with the B-G-R diagnostic test [49, 50] and by the effective sample size (*ESS*) [51]. Graphical and numerical summaries presented in this section are based on a single chain of length 10 000.

Table IV. Summary results of two fitted models for MR data. Posterior distributions are based on a single chain of length 20 000 with the first 10 000 iterations discarded.

Structural distribution	Link function	Parameter	Mean	2.5 per cent	50 per cent	97.5 per cent
Normal	Logistic	μ_D	2.462	1.646	2.450	3.329
		μ_S	−3.115	−3.925	−3.118	−2.284
		σ^2_D	0.836	0.148	0.617	2.836
		σ^2_S	0.822	0.139	0.575	2.939
		$\rho_{D,S}$	0.321	−0.614	0.411	0.909
		Sensitivity (pooled)	0.421	0.274	0.418	0.590
		Specificity (pooled)	0.941	0.909	0.942	0.964
		Sensitivity (predicted)	0.428	0.122	0.414	0.800
		Specificity (predicted)	0.935	0.844	0.942	0.981
		DIC	83.4			
Scale mixture	Logistic	μ_D	2.089	1.376	2.088	2.806
		μ_S	−3.368	−4.078	−3.377	−2.598
		σ^2_D	0.469	0.112	0.358	1.462
		σ^2_S	0.504	0.111	0.374	1.674
		$\rho_{D,S}$	0.179	−0.626	0.210	0.816
		v	2.636	0.856	2.373	5.943
		Sensitivity (pooled)	0.348	0.235	0.345	0.477
		Specificity (pooled)	0.937	0.904	0.939	0.962
		Sensitivity (predicted)	0.363	0.065	0.345	0.791
		Specificity (predicted)	0.923	0.755	0.939	0.987
		DIC	79.8			

4.1. Example: MRI

To analyze these data, we started by applying a bivariate Binomial–Normal model with two different link functions: logit and cloglog. The DIC of the model with logit link was 83.4, while the model with cloglog link was 82.4. We found that the last model slightly better fits the data, but not substantially. Given that the logit link is easily interpretable (e.g. as diagnostic odds rations, etc.) we choose this model for further analysis. A third model was fitted with a bivariate Binomial-t based on the scale mixture of normals, this model achieved a DIC with 79.8 indicating an interesting improvement in model fitting. In addition, the estimated degrees of freedom is $v = 2.634$, which shows that the Normal model is not adequate for these data.

Table IV summarizes numerical results of the Binomial–Normal and the Binomial-t models, respectively. The first important difference between these models is that the model based on the bivariate t-distribution gives narrower posterior distributions for the model parameters indicating an increase in efficiency in the use of the data. For example, the posterior distributions of the mean parameters (μ_D, μ_S) are 10 and 20 per cent larger under the normal distribution compared with the bivariate t-distribution. For the pooled sensitivity the posterior interval is 30 per cent narrow, whereas pooled specificity is similar under both models. Figure 2 compares the posterior distributions for the pooled sensitivity and specificity under these two models, we can clearly see the correction effect on the pooled specificity.

The effect on the predictive sensitivity and specificity is different to the pooled summaries. The model with bivariate t-distribution gives 10 and 70 per cent wider predictive posteriors for sensitivity and specificity respectively. Figure 3 presents the scatters of 200 predictive pairs of sensitivity and specificity under these two models. On the left panel the model with bivariate Normal shows less scatter of the predictive rates indicating inconsistence with the observed data. On the right panel the model with bivariate t-distribution shows substantially more scatter of the predictive rates given more consistency with the observed rates.

Posteriors of variance parameters in both models show that studies are not homogeneous and correlations show that D_i and S_i are uncorrelated. The posterior distribution of S_i is concentrated on negative values, which highlight that published studies overstated specificity at the expense of a low sensitivity.

The lower number of estimated degrees of freedom indicates the presence of studies with unusual study's effects and low mixture weights. Studies corresponding to Ho *et al.* 1992 and Kim *et al.* 1994 in Table III (lines 5 and 8) [35] had posterior mean mixture wights of 0.615 and 0.653, respectively. These two studies are located on the right panel of Figure 3 and indicted by larger circles. The study of Ho *et al.* 1992 reported 0 per cent sensitivity and 100 per cent

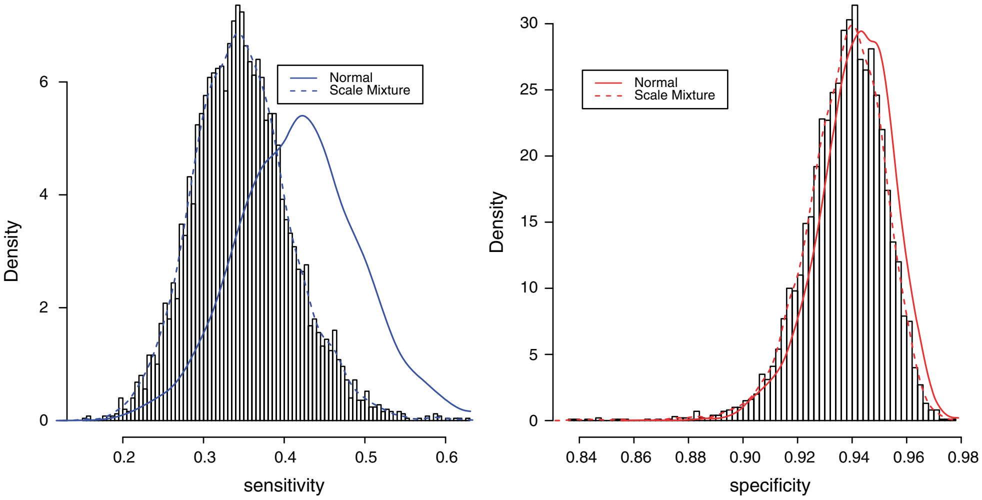


Figure 2. Results for MR data. Posterior distributions for the pooled sensitivity and specificity under Normal and Scale Mixture distribution.

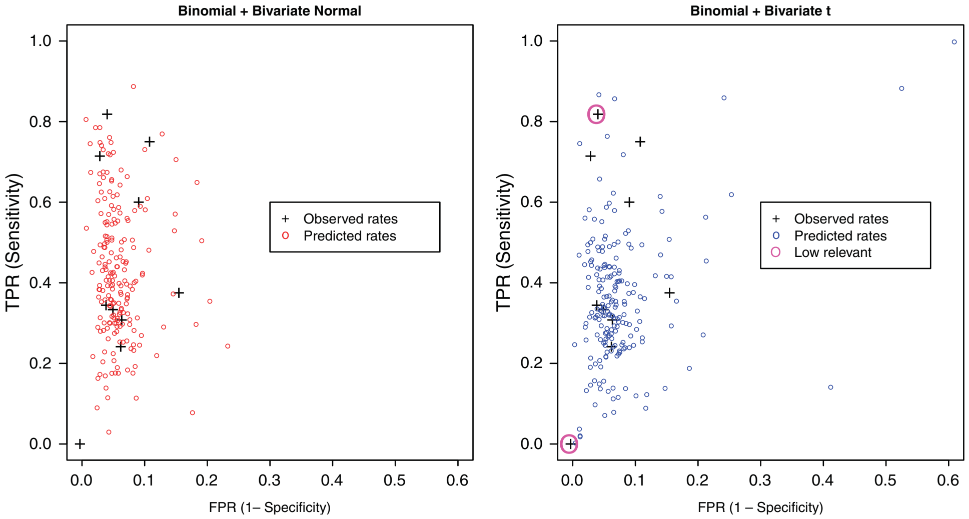


Figure 3. Posterior predicted sensitivity and specificity for MR data. Left panel: Model fitted with Normal random effects. Right panel: Model fitted with Scale mixture distribution.

specificity, a quite extreme result for these meta-analysis, while the study of Kim *et al.* 1994 is a big study with 272 patients with a high imbalance between disease and non-disease groups.

Comparing our analysis with the previous ones we can highlight the following points:

- For the pooled specificity results of the Bayesian Binomial-Normal model agree very much with results presented in Table I of Chu and Cole [20] and Table I of Martino and Rue [28]. But for the pooled sensitivity our approach gives slightly lower values. It is possible that the highly imbalanced and big study of Kim *et al.*(1994) influence the deterministic computations moving estimates upwards.
- The pooled summaries based on the Bayesian HSROC, Table III [11], showed the same pattern as Chu and Cole [20]. These authors also fit a t-distribution with fixed $df=2$ for the study effects under the HSROC model and they report in Section 4.6 that they obtain similar results as the Normal model, concluding that a Normal model gives a good fit for these data. Our analysis showed that there are clear outliers in these data and a Normal model can not accommodate these anomalies.
- In Table IV of Walter's [14] he presents estimations under the classical SROC and weighted SROC. The intercept parameter of the classical and weighted SROC is within the 95 per cent posterior interval of μ_D under the Binomial-Normal model, the slope parameters of the SROC are estimated close to zero which is in concordance with the posterior interval of $\rho_{D,S}(-0.614, 0.909)$. However, under the Binomial-t model the intercept parameters of

Walter's SROC analysis are outside the 95 per cent posterior interval of μ_D indicating that the SROC overestimates the diagnostic odds ratio in this meta-analysis.

- We fitted a Bayesian version of the three variate model with parametrization on (π_i, Se_i, Sp_i) and (P_i, PPV_i, NPV_i) of Chu *et al.* [25]. This analysis gave total DIC$=135.753$ for the (π_i, Se_i, Sp_i) and a partial DIC$=83.00$ for the components (Se_i, Sp_i). These results show that there is no improvement in model fitness compared with the bivariate Binomial-*t* model. We were not able to fit the (P_i, PPV_i, NPV_i) model because the study of Ho *et al.* 1992 in the Table III [35] (lines 5) corresponds to a Binomial sub-model $tp_5 \sim \text{Bin}(PPV_5, m_{5,1})$ with $m_{5,1}=0$.

4.2. *Example*: *CT appendicitis*

We start the analysis of these data with the same strategy of Example 4.1. Two models with a logit and a cloglog link functions where fitted using a Binomial–Normal model. The model with cloglog had a DIC$=414.7$ and the model with logit link had a DIC$=416.4$ showing no important differences in model fitness. As in Example 4.1, we chose the logit link for further analysis. A model with bivariate Binomial-*t* distribution was fitted, which resulted with a DIC$=405.7$, indicating a substantial improvement in model fitness.

In this example we found the same effect on the posterior summaries as the Example 4.1: the posterior distributions present narrower posterior credibility intervals in the Binomial-*t* model. The estimated degrees of freedom $v=4.7$ indicates that the Binomial–Normal model is not adequate for these data. Summary posteriors of variance parameters show that studies are not homogeneous and correlations show that D_i and S_i are negatively correlated. The posterior distribution of S_i is centered at zero indicating a balance between sensitivity and specificity for published studies.

The posterior of the pooled sensitivity is 0.955 (0.944, 0.964) and specificity is 0.952 (0.935, 0.966). These posterior summaries are very similar in both models. But posterior predictive summaries for sensitivity and specificity are 12.6 and 33.3 per cent wider under the Binomial-*t* distribution. The predictive posteriors for sensitivity and specificity under the Binomial-*t* model are 0.947 (0.868, 0.985) and 0.925 (0.669, 0.995), respectively.

In this example 7 studies were found with lower mixture weights, these are studies number (R)1, (R)3, (R)4, (R)7, (P)25, (P)29, (P)47 and they had posterior mean weights of 0.526, 0.563, 0.589, 0.664, 0.737, 0.685 and 0.387, respectively. Going back to study information in Table I, we found that these unusual diagnostic results have been produced by a remarkable imbalance between disease and non-disease groups, which is more accentuated in retrospective studies.

Given that study design gives the context in which these studies have been performed, we further explore its effect by including study design as covariate in the meta-regression model component and as a covariate in the study relevance model component, i.e. we fit the following model:

$$\mu_{i,D} = \alpha_0 + \alpha_1 x_{i,1} \tag{14}$$

$$\mu_{i,S} = \beta_0 + \beta_1 x_{i,1}, \tag{15}$$

$$w_i \sim \text{Gamma}(\rho, \mu_{w_i}), \tag{16}$$

$$\log(\mu_{w_i}) = \gamma_0 + \gamma_1 x_i, \tag{17}$$

where x_i is an indicator variable with

$$x_i = \begin{cases} 0 & \text{Prospective design,} \\ 1 & \text{Retrospective design.} \end{cases}$$

The DIC$=381.7$ indicates a great improvement in model fitness. However, the posterior mean and 95 per cent credibility interval for α_1 is -0.033 $(-0.814, 0.772)$ and for β_1 is 0.437 $(-0.352, 1.242)$, both results show no design effect. The improvement in DIC comes from explaining the increase of variability in the meta-analysis by including studies with retrospective design. This effect is measure by $\exp(\gamma_1)$ which has a posterior mean of 7.680 (1.665, 18.672) and posterior probability $\Pr(\exp(\gamma_1) > 1|\text{data}) = 0.997$.

The main reason to include studies with different design was to study the accuracy of CT in different populations and different CT setup. In order to study the influence of these characteristics in the meta-analysis, we extended the meta-regression model by including the covariates described in Table II. The DIC$=378.5$ shows slightly improvement in model fitness; however, some interesting patterns of these covariates are suggested in this analysis. Figure 4 summarizes these results. Each segment corresponds to the 2.5, 50 and 97.5 pe cent percentiles of the posterior distribution for each group of coefficients. The left panel corresponds to the posteriors for α_i's and the right panel for β_i's.

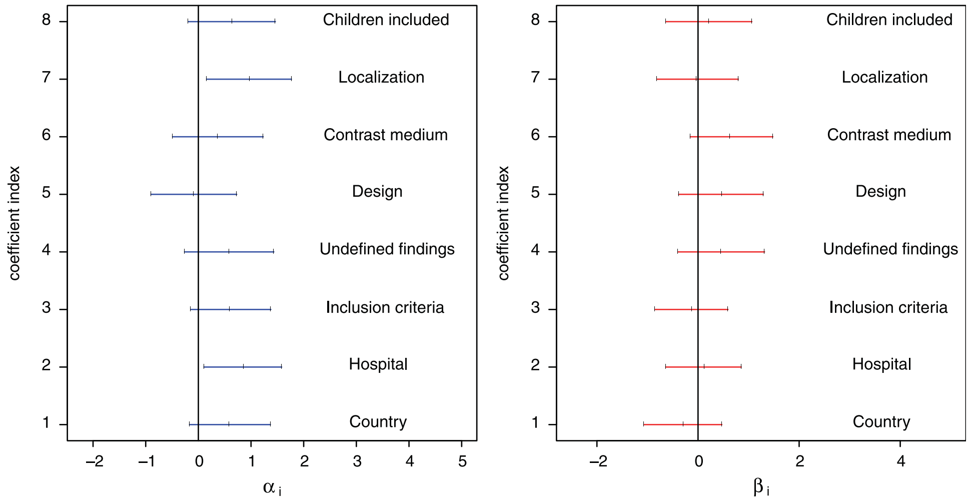


Figure 4. Summary plot for regression analysis. Left panel: regression coefficient α_i explaining influence of test discriminatory power. Right panel: regression coefficient β_i explaining influence of positive test results.

In Figure 4 we see that studies that have been performed outside university hospitals delivered higher diagnostic accuracy. Using more than one location in the CT setup have better diagnostic results. There is a tendency that studies which used contrast medium did not improve diagnostic results, but increased false positive rates.

5. Conclusions

In this article we have introduced a new Bayesian statistical model for meta-analysis of diagnostic test. The model is conceptually simple and combines familiar ideas of bivariate meta-analysis. In a way, this model may be interpreted as a Bayesian version of the classical bivariate meta-analysis approach, but versatile enough to deal with multiple sources of uncertainty which are usually present in this type of data.

A new data description based on study's diagnostic accuracy and study's dispersion has been presented. In particular the study's dispersion is interpreted as a measure of the rareness or incompatibility of a study included in the meta-analysis. Technically, this is achieved by using a scale mixture of bivariate Normals where the mixture weights for each study are directly interpreted as measuring excess of study's dispersion. This approach has the side effect of producing robust estimation of model parameters and delivering predictions compatible with the data at hand.

The data analysis of Section 4 shows that it is difficult to return to the naive idea that study effects follow a single bivariate normal distribution or equivalently that every study included in the review is worth the same amount of information. In this regard we agree with the recent work of Lee and Thompson [52], that inference regarding random effects should be based on distributions more flexible than the normal. As mentioned by the referee, other authors have already pointed to the fact that a random effects distribution will have heavier tails when there are unobserved confounding factors (Marshall and Spiegelhalter [53]).

Previous methodological work in this area summarized results by parameter estimates and uncertainties (typically in simple tabulation form) to conclude the analysis. They rarely address the problem of model checking, sometimes there is a model comparison using AIC, but rarely a graphical check showing the implications of the entire fitted model. In this work we recommend summarizing meta-analysis by predictive outcomes that vividly reflect the future use of meta-analytic results. These predictive quantities go hand in hand with model checking and can be used to spot out model deficiencies.

The model is extended to include covariates to explain systematic variability in the meta-analysis. As in the classical approach meta-regression can be used to explain diagnostic accuracy, but also, covariates can be used to explain structural dispersion. The use of study design as covariate to explain changes in variability give a parsimonious model to fit the data and it is an alternative to adding another hierarchical level as is commonly use in generalized evidence synthesis [54]. In this regard the model proposed in this work can be applied to other meta-analytic problems where studies with different designs have to be combine.

Given the complexity of the modeling procedures, we have found the MCMC calculations very stable and ready to use in routine meta-analytical work. This contrast with the HSROC approach, which in our experience can be hard to get convergence and with quadrature methods used in classical techniques, which can be unstable when we deal with small numbers of studies. In a recent simulation experiment Hamza *et al.* [55] reported unsatisfactory numerical results for quadrature methods when the number of studies included is 10 or less. The use of nested Laplace approximations is a powerful and accurate alternative to fit bivariate Bayesian random effects models, but the current implementation in the R package INLA [28] is not versatile enough to fit the models presented in this paper. Therefore, we recommend the use of MCMC techniques for bivariate meta-analysis. In particular our implementation is an straightforward application in WinBUGS and R.

Finally, we did not cover some aspects that may need future research, such as modeling several diagnostic tables per study, the inclusion of individual data, meta-analysis of diagnostic test with imperfect reference standard and publication bias in meta-analysis of diagnostic test.

Appendix A: R and BUGS code for the bivariate hierarchical model with scale normal mixtures

To perform the statistical analyzes described in this paper we run WinBUGS within R with the function `bugs()` from the package R2WinBUGS [48]. This approach combines the powerful MCMC calculations implemented in WinBUGS and gives flexibility for building plots and further summaries within R. It is the recommended form to make this type of Bayesian statistical analysis. In this appendix we describe the script to make the analysis of the MRI meta-analysis example. More details and the script for the CT example can be requested from the author.

The following script shows how to implement in BUGS language the bivariate binomial model with structural *t*-distribution based on scale mixtures. In order to fit this model we assume that the BUGS code is in the file `btlogit.tex` as follows:

```
#BUGS model: bivariate binomial t-distribution based on normal mixtures and with logit link.
model
{
for( i in 1 : n ) {
  tp[i] ~ dbin(tpr[i], n1[i]);  fp[i] ~ dbin(fpr[i], n2[i])
  m[i,1:2] ~ dmnorm(mu.0[1:2 ], sigma.inv[1:2, 1:2])
  w[i] ~ dgamma(nu.2, nu.2) I(0.001, 3)
  y[i, 1] <- mu[1] + m[i, 1] / sqrt(w[i])
  y[i, 2] <- mu[2] + m[i, 2] / sqrt(w[i])
  logit(tpr[i]) <- (y[i, 1] + y[i, 2])/2
  logit(fpr[i]) <- (y[i, 2] - y[i, 1])/2
 }

# Priors ...
mu[1] ~ dnorm(0, 0.25) ;  mu[2] ~ dnorm(0, 0.25)
mu.0[1] <- 0;  mu.0[2] <- 0; nu.2 <- nu/2
nu ~ dexp(1)
sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2], 3)

# Pooled summaries ...
x.pool <- (mu[1]+mu[2])/2;  y.pool <- (mu[2]-mu[1])/2
pool.se <- exp(x.pool) / ( 1 + exp(x.pool) )
pool.sp <- 1 - exp(y.pool) / ( 1 + exp(y.pool) )

# Predictive summaries...
mu.s[1:2] ~ dmt(mu[], sigma.inv[1:2, 1:2], nu)
x.s <- (mu.s[1] + mu.s[2])/2; y.s <- (mu.s[2] - mu.s[1])/2
new.se <- exp(x.s)/(1+exp(x.s)); new.sp <- 1-exp(y.s)/(1+exp(y.s))

# Variance covariance matrix for random-effects...
sigma[1:2, 1:2] <- inverse(sigma.inv[1:2, 1:2])
sigmaD <- sigma[1,1]; sigmaS <- sigma[2, 2]
rhoDS <- sigma[1,2]/(pow(sigmaD, 0.5) * pow(sigmaS, 0.5))
}
```

To fit this model, we need to specify two R objects, one with the names of *the data* and another with the names of *the parameters* of interest, so in R we have:

```
# Binomial + t + logit
# R is the scale matrix of  the Wishart
# tp true positives, etc..
data.t <- list ("R", "tp", "n1", "fp", "n2", "n")
parameters.t  <- c("nu", "w", "pool.se", "pool.sp", "new.se", "new.sp", "mu",
  "sigmaD", "sigmaS", "rhoDS")
```

The function `bugs()` has a series of arguments that are needed to run BUGS:

```
mt <- bugs(data.t, inits=NULL, parameters.t, "btlogit.txt", n.chains = 1,
  n.iter = 20000,  n.thin=1, bugs.directory = bugsdir, working.directory = getwd(),
  clearWD=TRUE, debug=TRUE)
```

The first argument refers to the data nodes, the second how initial values are generated (here `NULL` means that BUGS will generate these values randomly), `parameters.t` is the vector of parameters to monitor and `btlogit.txt` is the BUGS model. In this example, the argument `n.chains=1` indicates that we generate one chain, `n.inter = 20000` the length of the chain, by default the first `n.inter/2` iterations will be omitted for analysis. For more details see the help files of `bugs()`.

The resulting object `mt` is an R object from the class `mcmc.list`, which can be analyzed using the package `coda` or manually as we do here. For example the `print()` function gives a summary of the object:

```
> print(mt, digits=3)
...
            mean       sd    2.5%     25%     50%       75%    97.5%
nu          2.636    1.332  0.856   1.682   2.373     3.281    5.943
...
mu[1]       2.089    0.364   1.376   1.849   2.088     2.325    2.806
mu[2]      -3.368    0.375  -4.078  -3.622  -3.377    -3.128   -2.598
sigmaD      0.469    0.396  0.112   0.233   0.358    0.576    1.462
sigmaS      0.504    0.442  0.111   0.238   0.374    0.617    1.674
rhoDS        0.179    0.390  -0.626  -0.096   0.210    0.489    0.816
deviance   72.043    5.485  62.240  68.200   1.640    5.490   83.970
DIC info (using the rule, pD = Dbar-Dhat)
pD = 7.7 and DIC = 79.8
DIC is an estimate of expected predictive error (lower deviance is better).
>
```

The following lines show how to access *sensitivity* and *specificity* posterior distributions and plot them:

```
> sensitivity <- mt$sims.array[,1,"se"]
> specificity <- mt$sims.array[,1,"sp"]
> par(mfrow = c(1,2))
> hist(sensitivity, breaks=80, prob=T, main="", xlab="sensitivity")
> lines(density(sensitivity), lwd = 2, col ="blue")
> hist(specificity, breaks=80, prob=T, main="", xlab="specificity")
> lines(density(specificity), lwd = 2, col ="red")
> par(mfrow = c(1,1))
```

Acknowledgements

The author is very grateful to Christian Ohmann, Tibor Gilbers and Claus Franke for kindly allowing him to use the CT data and for innumerable discussions about meta-analysis of diagnostic test. An early version of this work was presented in the Department of Statistics at Stanford University, the author is grateful to Brad Efron and Ingram Olkin for additional discussions and several suggestions, which have improved this work. The author was partially financed by the Swiss National Science Foundation (51A240-104890). Finally, the author is grateful to the anonymous reviewer who gave valuable comments and suggestions.

References

1. Knottnerus JA (ed.). *The Evidence Base of Clicial Diagnostis*. BMJ Brooks: London, U.K., 2002.
2. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methos for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**:119–130.
3. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR* 2006; **187**:271–281.
4. Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine* 2002; **21**:1525–1537.
5. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic test. *The Journal of the American Medical Association* 1999; **282**:1061–1066.
6. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Medical Research Methodology* 2005; **5**:1471–2288.
7. Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003; **3**:25.
8. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; **12**:1293–1316.
9. Cochrane handbook for diagnostic test accuracy reviews (draft) in preparation by cochrane diagnostic test accuracy working group. Available from: http://srdta.cochrane.org/en/index.html. Accessed 21/08/2009.
10. Rutter C, Gatsonis C. Regression methods for meta-analysis of diagnostic test data. *Academic Radiology* 1995; **2**:S48–S56.

11. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**:2865–2884.
12. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003; **59**:936–946.
13. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**(9):925–932.
14. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine* 2002; **21**:1237–1256.
15. Walter SD. The partial area under the summary ROC curve. *Statistical in Medicine* 2005; **24**:2025–2040.
16. Walter SD, Irwig L, Glasziou PP. Meta-Analysis of diagnostic test with Imperfect reference standards. *Journal of Clinical Epidemiology* 1999; **52**:943–951.
17. Deeks JJ, Macaskill P, Irwig L. The performance of test of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005; **58**(9):865–866.
18. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**:982–990.
19. Arends LR, Hamza TH, Houwelingen JCv, Heijenbrok MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* 2008; **28**:621–638.
20. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparce data: a generalized linear mixed model approach. *Journal of Clincial Epidemiology* 2006; **59**:1331–1332.
21. Cong X, Cox DD, Cantor SB. Bayesian meta-analysis of Papanicolaou smear accuracy. *Gynecology and Oncology* 2007; **107**:133–137.
22. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**:239–251.
23. Chu H, Guo H. Letter to the editor. *Biostatistics* 2009; **10**(1):201–203.
24. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood rations in systematic reviews. *Statistics in Medicine* 2009; **27**:687–697.
25. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parametrization and model selection. *Statistics in Medicine*, 2009. Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3627.
26. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropiate for diagnostic meta-analyses? *Statistics in Medicine* 2009; **28**:2653–2668.
27. Paul M, Riebler A, Bachmann LM, Rue H, Held L. Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in Medicine* 2010; **29**:1325–1339.
28. Martino S, Rue H. Implementing approximate bayesian inference using integrated nested Laplace approximation: a manual for the Inla program. Department of Mathematical Sciences, NTNU, Norway 2009.
29. Riley RD, Dodd SR, Crain JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregated data. *Statistics in Medicine* 2008; **27**:6111–6136.
30. Ohmann C, Verde PE, Gilbers T, Franke C, Fuerst G, Sauerland S, Boehner H. Systematic review of CT investigation in suspected acute appendicitis. *Final Report*; Coordination Centre for Clinical Trials, Heinrich-Heine University. Moorenstr. 5, D-40225 Duesseldorf Germany, 2006.
31. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. *American Journal of Epidemiology* 1990; **5**:910–925.
32. Kraemer M, Ohmann C, Leppert R, Yang Q. Macroscopic assessment of the appendix at diagnostic laparoscopy is reliable. *Surgical Endoscopy* 2000; **7**:625–633.
33. Flum DR, Koepsell T. The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis. *Archives of Surgery* 2002; **137**(7):799–804.
34. The Cochrane methods group on systematic review of screening and diagnostic tests, recommended methods: screening and diagnostic tests, 2005. Available at: www.cochrane.org/cochrane/sadtdoc1.htm.
35. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *The Journal of the American Medical Association* 1997; **278**:1096–1101.
36. Carlin BP, Louis TA. *Bayesian Methods for Data Analysis* (3rd edn). CRC Press, Chapman & Hall: London, 2009.
37. Park T, Casella G. The Bayesian Lasso. *JASA* 2008; **103**:681–686.
38. Verde PE. Meta-analysis of diagnostic test data: modern statistical approaches. *Deutsche Nationalbibliothek* 2008. Available at: urn:nbn:de:hbz:061-20080715–114548-5.
39. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society A* 2009; **172**(Part 1):127–159.
40. Box GEP. Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* 1980 **143**:383–430.
41. Rubin DB. Estimation in parallel randomized experiments. *Journal of Statistics Education* 1981; **6**:377–400.
42. Rubin DB. Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 1984; **12**:1151–1172.
43. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (2nd edn). Chapman & Hall/CRC: London, Boca Raton, FL, 2004; 488–491.
44. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: Cambridge, 2007; 513–527.
45. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society* 2002; **B. 64**:583–640.
46. Spiegelhalter DJ, Thomas A, Best N. *WinBUGS, Version 1.4, Upgraded to 1.4.1, User Manual*. MRC Biostatistics Unit: Cambridge, 2004.
47. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2009. Available at: http://www.R-project.org.

48. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 2005; **12**(3):1–16.
49. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 1992; **7**:457–511.
50. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1997; **7**:434–455.
51. Kass RE, Carlin BP, Gelman A, Neal R. Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician* 1998; **52**:93–100.
52. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 2008; **27**:418–434.
53. Marshall EC, Spiegelhalter DJ. Institutional Performance. *Multilevel Modelling of Health Statistics*, Leyland AH, Goldstein H (eds). Wiley: New York, 2001; 127–142.
54. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society, A.* 2006; **169**:5–35.
55. Hamza TH. Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal–normal and binomical–normal bivariate summery ROC approaches. *Medical Decision Making* 2008; **28**:639–649.

Chapter 7

bamdit: an R Package for Meta-Analysis of Diagnostic Test Data

> ”That’s what a computer is to me: the computer is the most remarkable tool that we’ve ever come up with. It’s the equivalent of a bicycle for our minds.”
> Steve Jobs,

Submission history:

The paper presented in chapter has been submitted to the Journal of Statistical Software in June 2015. The R package **bamdit** version 2 was uploaded to The Comprehensive R Archive Network in June 2015.

E

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

bamdit: an R package for Bayesian meta-analysis of diagnostic test data

Pablo Emilio Verde

Abstract

In this paper we present the R package **bamdit**, its name stands for "**Ba**yesian **m**eta-analysis of **di**agnostic **t**est-data". **bamdit** was developed with the aim of simplifying the use of models in meta-analysis, that up to now have demanded great statistical expertise in Bayesian meta-analysis. The package implements a series of innovative statistical techniques including: the Bayesian Summary Receiver Operating Characteristic (BSROC) curve, the use of prior distributions that avoid boundary estimation problems of component of variance and correlation parameters, analysis of conflict of evidence and robust estimation of model parameters. In addition, the package comes with several published examples of meta-analysis that can be used for illustration or further research in this area.

Keywords: meta-analysis, diagnostic test data, hierarchical models, conflict of evidence, bias modeling, MCMC, JAGS, R.

1. Introduction

One of the most important decisions in the presence of illness is the correct medical diagnosis. Ideally, for a particular diagnostic problem we should have a collection of studies which indicate the best way to proceed. However, this is not the case in clinical and other areas of empirical research. Instead, researchers have to face a heterogeneous and fragmented evidence that has to be analyzed.

Meta-analysis is a branch of statistical techniques that helps researchers to combine evidence from a multiplicity of sources. In particular, meta-analysis of diagnostic test data differs from other types of meta-analysis in several aspects: First, the diagnostic summaries that we aim to combine (e.g. sensitivity and specificity) could be interdependent and a marginal combination by pooling these quantities might be misleading (Irwig, Macaskill, Glasziou, and Fahey 1995). Second, diagnostic studies are usually performed under slightly different diagnostic setups and they can be applied to different patients' populations. Hence, we can expect high heterogeneity between studies' results. In addition, the number of studies included might be small and with different qualities (e.g. they might have different study designs) (Lijmer, Mol, Heisterkamp, Bonsel, Prins, van der Meule, and Bossuyt 1999; Lijmer, Bossuyt, and Heisterkamp 2002; Westwood, Whiting, and Kleijnen 2005). Hence, conducting meta-analysis and combining results from diagnostic studies may become a challenge.

In this paper we present the R package **bamdit**. The name of the package stands for "**Ba**yesian **m**eta-analysis of **di**agnostic **t**est-data". The development of the package started with the following question: "How can we make complex meta-analysis in an automatic fashion?"

The initial release of **bandit** was the version 1.0 of the Summer 2011. This version was an experimental package where the aim was to investigate different statistical software architectures to fit complex meta-analysis models. During the last years we have rewritten and updated the package several times with the intention of making the package more user friendly. The current release corresponds to version 2.0 which is presented in this paper.

The package may be helpful to practitioners who are not familiar with complex Bayesian modeling and who do not have the skills to implement these models in general proposed Bayesian software such us WinBUGS/OpenBUGS Lunn, Spiegelhalter, Thomas, and Best (2009) or JAGS Plummer (2003).

For more than a decade meta-analysis of diagnostic tests has been an active area of research. Statistical methods have fallen into two main approaches: On the one hand we have techniques that have a focus on making a meta-analysis summary by recovering an underlined Receiver Operating Characteristic (ROC) curve. This in the case of the summary ROC (SROC) curve introduced by Moses, Shapiro, and Littenberg (1993) and the hierarchical ROC (HROC) curve presented in Rutter and Gatsonis (1995, 2001); Macaskill (2004).

On the other hand we have approaches that directly model the diagnostic outcomes as a bivariate meta-analysis (Reitsma, Glas, Rutjes, Scholten, Bossuyt, and Zwinderman 2005; Chu and Guo 2009). The relationships between these two approaches have been investigated by Harbord, Deeks, Egger, Whiting, and Sterne (2007) and Arends, Hamza, Van Houwelingen, Heijenbrik, Hunink, and Stijnen (2008) from the classical perspective and by Novielli, Cooper, Sutton, and Abrams (2010) from the Bayesian perspective.

Recent research in meta-analysis of diagnostic test data has focused on the problem of modeling heterogeneity (Verde 2010b), measuring heterogeneity (Zhou and Dendukuri 2014), assessing publication bias (Buerkner and Doebler 2014) and modeling results in the presence of imperfect reference standard (Menten, Boelaert, and Lesaffre 2013).

Software for meta-analysis has been available for many years, in particular in R (Team 2013) several packages have been developed for different meta-analytic problems. An extensive list with a comprehensive description of these packages is presented in the CRAN task view ”Meta-Analysis” (Dewey 2014). In particular the following R packages have been developed for meta-analysis of diagnostic test data: **mada** implements the bivariate method of Reitsma *et al.* (2005). **HSROC** provides the implementation of the hierarchical summary receiver operating characteristic (HSROC) method of Rutter and Gatsonis (2001). **Meatron** includes the implementation of the Reitsma *et al.* (2005) model including the case of diagnostic test with an imperfect reference standard. **metamisc** implements the method of Riley, Lambert, Staessen, Wang, Gueyffier, Thijs, and Boutitie (2008) which estimates a common within and between correlation when the within-study correlations are unknown.

Implementation of different Bayesian meta-analysis models for diagnostic test data in WinBUGS software is discussed in Rutter and Gatsonis (2001), Verde (2008, 2010b) and Novielli *et al.* (2010). Approximate Bayesian methods using INLA (Integrated Nested Laplace Approximation) can be found in Paul, Riebler, Bachmann, Rue, and Held (2010).

The rest of the paper is organized as follows: In Section 2 we describe the software implementation of **bamdit**. In Section 3 we present methodological details of the Bayesian statistical model. In Section 4 we show how to use **bamdit** in practice. Finally, in Section 5 we give a brief summary of the work and we discuss future developments of package.

2. Software implementation

In the implementation of **bamdit** we have considered that the package should be easy to use for practitioners familiar with R, but without Bayesian statistical background. We also considered that the package has to be portable between different operative systems. **bamdit** uses JAGS for MCMC computations, therefore the main system requirement is that JAGS ($\geq$ 3.4.0) is installed in your computer (see http://mcmc-jags.sourceforge.net).

From the statistical point of view, the software reduces the risk of having boundary problems in the estimation of the variance components and correlation between random effects of the meta-analysis model. In this regard it can be applied to problems where classical approaches fail (see Section 4). In addition, **bamdit** is equipped with an automatic analysis of conflict of evidence (Verde 2014) which allows to spot out studies with unusual results that have been included in the meta-analysis.

A single function called `metadiag()` performs the meta-analysis. This function allows to fit bivariate Normal random effects or bivariate scale mixture of Normals. The default link function is the logistic link, but the user can choose between the three classical link functions of binomial data: logistic, complementary log-log or probit. The output of this function can be analyzed with **R2jags** or with **rjags** packages. Internally, this function writes the BUGS script and send the script to JAGS where MCMC (Markov Chain Monte Carlo) computations are performed and returned to R. Further statistical details of the model behind **bamdit** is presented in Section 3.

Convergence of the MCMC computations can be analyzed using the R package **coda** (Plummer, Best, Cowles, and Vines 2006). In addition, we have implemented a series of graphical functions that can be used to summarize results and to compare results between models. We demonstrate this software's functionality in Section 4.

3. Bayesian meta-analysis of diagnostic test data

3.1. Data model for diagnostic test results

We assume that the pieces of evidence that we aim to combine are the results of N diagnostic studies, where results of the ith study ($i = 1, \dots, N$) are summarized in a 2×2 table as follows:

		Patient status	
		With disease	Without disease
Test	+	tp_i	fp_i
outcome	-	fn_i	tn_i
Sum:		$n_{i,1}$	$n_{i,2}$

where tp_i and fn_i are the number of patients with positive and negative diagnostic results from $n_{i,1}$ patients with disease and fp_i and tn_i are the positive and negative diagnostic results from $n_{i,2}$ patients without disease.

Assuming that $n_{i,1}$ and $n_{i,2}$ have been fixed by design, we model the tp_i and fp_i outcomes with two independent Binomial distributions:

$$tp_i \sim \texttt{Binomial}(\mathrm{TPR}_i, n_{i,1}) \quad \texttt{and} \quad fp_i \sim \texttt{Binomial}(\mathrm{FPR}_i, n_{i,2}), \tag{1}$$

where TPR_i is the true positive rate or sensitivity of study i and FPR_i its the false positive rate or complementary specificity (1-specificity).

At face value, diagnostic performance of each study is summarized by the empirical true positive rate and true negative rate or specificity,

$$\widehat{\mathrm{TPR}}_i = \frac{tp_i}{n_{i,1}} \quad \text{and} \quad \widehat{\mathrm{TNR}}_i = \frac{tn_i}{n_{i,2}} \tag{2}$$

and the complementary empirical rates of false positive rate and false negative diagnostic results,

$$\widehat{\mathrm{FPR}}_i = \frac{fp_i}{n_{i,2}} \quad \texttt{and} \quad \widehat{\mathrm{FNR}}_i = \frac{fn_i}{n_{i,1}}. \tag{3}$$

The main question in meta-analysis of diagnostic test data is: How can we combine the multiplicity of diagnostic accuracy rates in a single coherent model? In this work we recognize that in order to combine results of different studies we have to explicitly model the variability between studies, which is the topic of the next section.

3.2. Random effects model

We model between studies variability with the following random components:

$$D_i = g(\mathrm{TPR}_i) - g(\mathrm{FPR}_i) \quad \text{and} \quad S_i = g(\mathrm{TPR}_i) + g(\mathrm{FPR}_i), \tag{4}$$

where $g(\cdot)$ corresponds to a link function which maps the diagnostic rates to the real scale $(-\infty, \infty)$. The canonical link function used in this work is the logistic link $g(p) = \log(p/(1-p))$, but other links are also possible (e.g. the complementary log-log link function $g(p) = \log(-\log(1-p))$.

The random component D_i represents the study effect associated with the diagnostic discriminatory power. For example, the logistic link function of D_i corresponds to the diagnostic odds ratio in the logarithmic scale:

$$D_i = \log\left(\frac{\mathrm{TPR}_i}{1-\mathrm{TPR}_i}\right) - \log\left(\frac{\mathrm{FPR}_i}{1-\mathrm{FPR}_i}\right). \tag{5}$$

Meta-analysis based on odds ratios is a common practice for therapeutic outcomes and for diagnostic studies one could also follows this approach. However, diagnostic results are sensitive to the diagnostic settings (e.g. the use of different thresholds) and to the populations where the diagnostic procedure under investigation is applied. These issues are associated with the *external validity* of diagnostic results.

Following the footsteps of Moses *et al.* (1993), Verde (2010a) introduced the random effect S_i. This random effect quantifies variability produced by patients' characteristics, study design and diagnostic setup, that may produced a correlation between the observed $\widehat{\mathrm{TPR}}$s and $\widehat{\mathrm{FPR}}$s. In short, we called S_i **the threshold effect** of study i and represents and adjustment of external validity in the meta-analysis.

Conditionally to a study weight w_i, the study effects D_i and S_i are modeled as exchangeable between studies and they follow a *scale-mixture of bivariate Normal* distributions with mean and variance:

$$E\left[\left(\begin{array}{c} D_i \\ S_i \end{array}\right) \middle| w_i\right] = \left(\begin{array}{c} \mu_D \\ \mu_S \end{array}\right), \quad \texttt{and} \quad var\left[\left(\begin{array}{c} D_i \\ S_i \end{array}\right) \middle| w_i\right] = \frac{1}{w_i}\left(\begin{array}{cc} \sigma_D^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_S^2 \end{array}\right) = \Sigma_i, \tag{6}$$

and scale mixing density

$$w_i \sim p(w_i). \tag{7}$$

The inclusion of the random weights w_i into the model was proposed by Verde (2010a), where $p(w_i)$ allows for a great flexibility to model the marginal distribution of D_i and S_i. Two important cases are: $w_i \sim \chi^2(\nu)$, which corresponds to a marginal bivariate t-distribution with known degrees of freedom ν, and $p(w_i = 1) = 1$ which corresponds to a bivariate Normal distribution. In the case of the bivariate t-distribution by integrating w_i from the conditional distribution of $(D_i, S_i|w_i)$ we have a marginal variance of

$$var\left[\left(\begin{array}{c} D_i \\ S_i \end{array}\right)\right] = \frac{\nu}{\nu-2}\left(\begin{array}{cc} \sigma_D^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_S^2 \end{array}\right), \tag{8}$$

hence we have to restrict $\nu > 2$ in order to have non infinite marginal variance in the random effects.

Another important aspect of w_i is its interpretation as **estimated bias correction**. *A priori* all studies included in the review have a mean of $E(w_i) = 1$, we can expect that studies which are unusually heterogeneous will have posteriors substantially greater than 1. If the model is not corrected by the influence of unusual study results, then the meta-analysis may produce biased results.

Unusual studies' results could be produced by factors that may affect the quality of the study, such as errors in recording diagnostic results, confounding factors, loss to follow-up, etc. For that reason, the studies' weights w_i can be interpreted as an adjustment of studies' **internal validity bias**.

Figure 1 displays the Directed Acyclic Graph (DAG) of the model presented in this section. In the usual DAG notation, elliptical nodes represent random variables (parameters and data), rectangular nodes represent fixed parameters, single arrows correspond to stochastic dependencies between nodes and double arrows correspond to deterministic relationships. Model parameters with priors are depicted with ellipses with dashed lines. Repeated structures of the graph are represented by the central plate, where each 2×2 table is modeled as the result of diagnostic parameters (TPR_i and FPR_i) which are the result of random study effects (D_i and S_i). The model of interest is framed with a rectangle containing the hyper-parameters of the model ($\mu_D, \mu_S, \sigma_D, \sigma_S, \rho$).

The DAG of Figure 1 links the statistical model to the MCMC computations implemented in JAGS. Using an automatic theorem proof algorithm JAGS factorized the joint posterior distribution in a set of conditional distributions which are used for Gibbs sampling. In addition the DAG representation helps to understand how to extend the model of interest. For example, the *pooled Sensitivity* and the *pooled Specificity* are the result as functional parameters of the hyper-parameters (see Section 3.5).

3.3. Splitting the studies' weights and conflict of evidence analysis

In Verde (2014) I conjectured that a way to perform conflict of evidence in a multi-parameter meta-analysis model was to extend the the random effects distribution by using a scale mixture of normal distributions per random effect. I have called this technique *"splitting the studies' weights"* and it is implemented in the **bamdit**'s function `metadiag()` by using the argument `split.w = TRUE`.

The study's weight w_i is now "split" in two components weights $w_{i,1}$ and $w_{i,2}$, these weights measure individual conflict for the components D_i and S_i respectively. For example, if the sources of conflict are studies with unusual specificity the posteriors of $w_{i,2}$ will be away from a prior mean $E(w_{i,2}) = 1$, while the corresponding posteriors of $w_{i,1}$ will be concentrated around the prior mean. We illustrate how to use this technique in the examples of Section 4.

Conditionally to a study weights $w_{i,1}$ and $w_{i,2}$, the study effects D_i and S_i are modeled as exchangeable between studies. We use as a common scale mixing density a χ^2 distribution:

$$w_{1,1}, \ldots, w_{n,1}, w_{1,2} \ldots, w_{n,2} \sim \chi^2(\nu), \tag{9}$$

with known degrees of freedom ν.

3.4. Priors for Hyperparameters

The formulation of the model for aggregate data is completed by specifying the priors for the

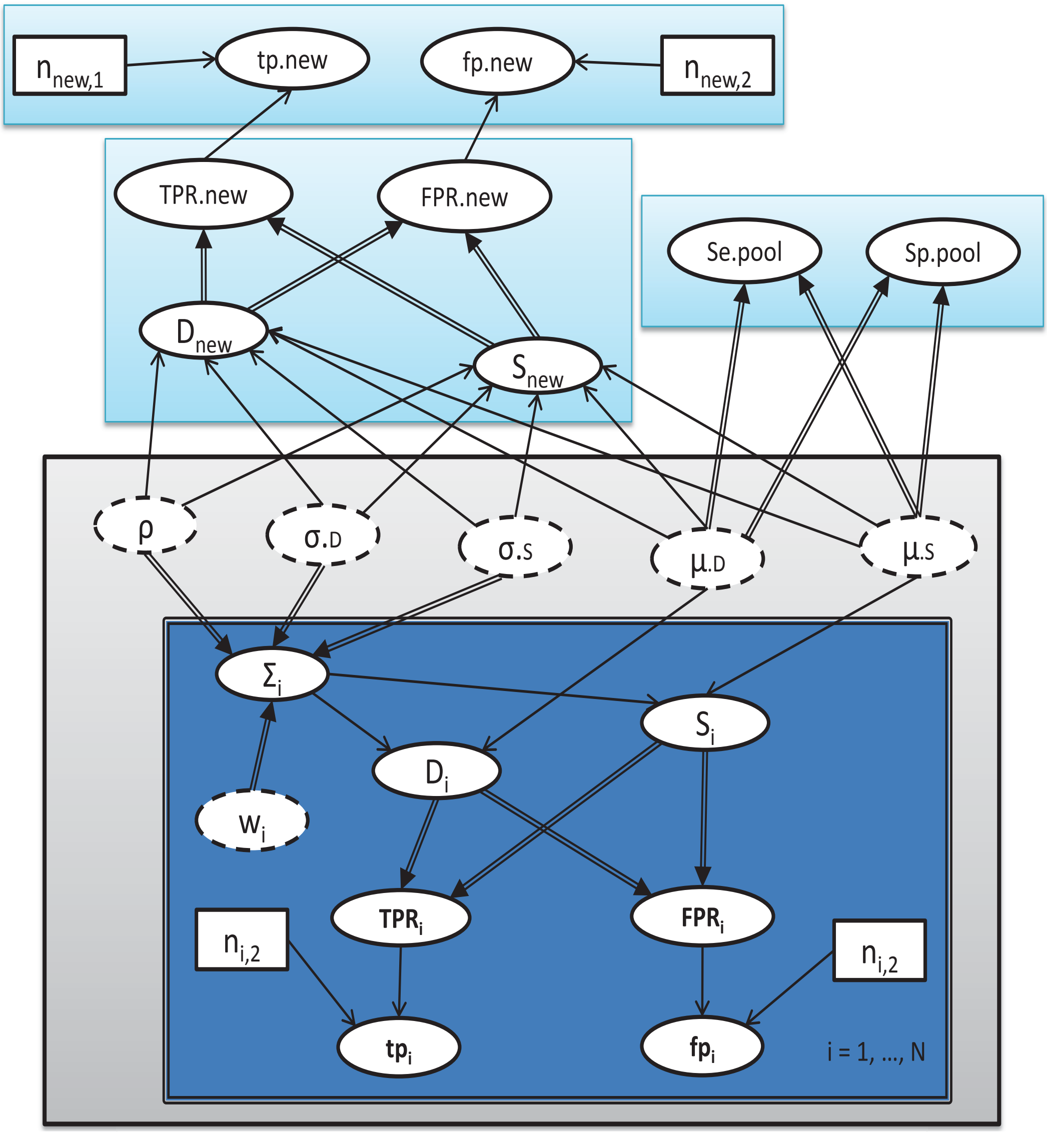


Figure 1: DAG for the model which combines diagnostic accuracy results. Elliptical nodes represent random variables (parameters and data), rectangular nodes represent fixed parameters, single arrows correspond to stochastic dependencies between nodes and double arrows correspond to deterministic relationships. Model parameters with priors are depicted with ellipses with dashed lines. Repeated structures of the graph are represented by the central plate. The model of interest is framed with a rectangle containing the hyper-parameters of the model $(\mu_D, \mu_S, \sigma_D, \sigma_S, \rho)$.

hyperparameters $\mu_D, \mu_S, \sigma_D, \sigma_S$ and ρ. We assume that parameters are independent and we use the following set of priors:

$$\mu_D \sim \texttt{Logistic}(m_1, v_1), \quad \mu_S \sim \texttt{Logistic}(m_2, v_2) \tag{10}$$

and

$$\sigma_D \sim \texttt{Uniform}(0, u_1), \quad \sigma_S \sim \texttt{Uniform}(0, u_2). \tag{11}$$

The correlation parameter ρ, is transformed by using the Fisher transformation,

$$z = \texttt{logit}\left(\frac{\rho + 1}{2}\right)$$

and a Normal prior is used for z:

$$z \sim \texttt{Normal}(m_r, v_r). \tag{12}$$

Modeling priors in this way guarantees that in each MCMC iteration the variance-covariance matrix of the random effects θ_1 and θ_2 is positive definite. The values of the constants $m_1, v_1, m_2, v_2, u_1, u_2, m_r$ and v_r have to be given. They can be used to include valid prior information, which might be empirically available or they could be the result of expert elicitation. If such information is not available, we recommend setting these parameters to values that represent weakly informative priors. In this work, we use $m_1 = m_2 = m_r = 0$, $v_1 = v_2 = 1$ and $v_r = 1$ as weakly informative prior setup.

These values are fairly conservative, in the sense that they induce prior uniform distributions for TPR_i and FPR_i. They give locally uniform distributions for μ_1 and μ_2; uniforms for σ_1 and σ_2; and a symmetric distribution for ρ centered at 0. In our experience, the most difficult parameter to estimate in this model is ρ. Therefore, we recommend to make *a prior to posterior sensitivity analysis* by giving different values for m_r and v_r to understand their influence in the analysis.

Finally, in the current implementation of **bamdit** we give a fixed value of the degrees of freedom ν with a default value of $\nu = 4$.

3.5. Pooled and predictive summaries

In meta-analysis of diagnostic data we are interested in summarizing the overall accuracy of the test in term of the *pooled Sensitivity* and the *pooled Specificity*.

These quantities are calculated as functions of μ_D and μ_S as following:

$$\text{Sensitivity}^{pooled} = g^{-1}[(\mu_D + \mu_S)/2], \quad \text{Specificity}^{pooled} = 1 - g^{-1}[(\mu_D - \mu_S)/2]. \tag{13}$$

In Figure 1 these quantities are represented as functions of logical nodes, statistical inference is based on sampling from their marginal posterior distributions:

$$p(\text{Sensitivity}^{pooled}|\texttt{Data}) \quad p(\text{Specificity}^{pooled}|\texttt{Data}). \tag{14}$$

Another important summary is the predicted pairs of rates (FPR, TPR) for a study that has not been included in the meta-analysis. Statistical inference of these quantities is based on sampling from the bivariate predictive posterior

$$p(\text{TPR}^{new}, \text{FPR}^{new}|\texttt{Data}). \tag{15}$$

In Figure 1 we display how this posterior is built by defining a stochastic node (D^{new}, S^{new}) which is used to calculate $\text{TPR}^{new}, \text{FPR}^{new}$ in each MCMC iteration.

The predictive posterior(15) can be used graphically in order to report the predictive surface at a given credibility level (e.g. 95%). We call this summary the Bayesian Predictive Surface (BPS). Clearly, in this model framework we can calculate the marginal predictive posteriors $p(\text{TPR}^{new}|\texttt{Data})$ and $p(\text{FPR}^{new}|\texttt{Data})$.

The predictive posterior (15) can be used to generate predictive data. This process is described at the top of Figure 1. A total number of patients is fixed in each group n_1^{new} and n_2^{new} and the predictive number of true positive and false positive results is generated by using two independent Binomial distributions with predictive rates $\text{TPR}^{new}, \text{FPR}^{new}$. These predictive data can be used to assess what is expected in a new diagnostic study with n_1^{new} and n_2^{new} patients per group.

Data prediction can be extended to generate N studies with the same number of $n_{i,1}$ and $n_{i,2}$ as the original ones. The resulting predictive data can be compared with the observed data to assess model misfit.

3.6. Conditional summaries and the Bayesian SROC curve (BSROC) and the area under the curve (BAUC)

The most commonly statistical technique use for practitioners to summarize meta-analysis of diagnostic data is the Summary Receiving Operating Characteristic (SROC) curve introduced by Moses *et al.* (1993). The model presented in Section 3 allows to build the Bayesian version of the SROC curve introduced by Verde (2008).

An alternative representation of the marginal model presented in Section 3.2 is the model based on the conditional distribution of $(D_i|S_i = x)$ and the marginal distribution of S_i. The conditional mean of $(D_i|S_i = x)$ is given by:

$$E(D_i|S_i = x) = \text{A} + \text{B}\, x \tag{16}$$

where the functional parameters A and B are

$$\text{A} = \mu_D, \quad \text{and} \quad \text{B} = \rho\frac{\sigma_D}{\sigma_S}. \tag{17}$$

We define the Bayesian SROC curve (BSROC) by transforming back results from (S, D) to (FPR, TPR) with

$$\text{BSROC}(\text{FPR}) = \text{g}^{-1}\left[\frac{A}{(1-B)} + \frac{B+1}{(1-B)}\,\text{g}(\text{FPR})\right]. \tag{18}$$

The BSROC curve is obtained by calculating TPR in a grid of values of FPR which gives a posterior conditionaly on each value of FPR. Therefore, it is straightforward to give credibility intervals for the BSROC for each value of FPR.

One important aspect of the BSROC is that incorporates the variability of the model's parameters, which influence the wide of its credibility intervals. In addition, give that FPR is modeled as random variable, the curve is corrected by measurement error bias in FPR.

Finally, we can define a Bayesian area under the SROC curve (BAUC) by numerically integrating the BSROC for a range of values of the FPR:

$$\text{BAUC} = \int_{fpr_0}^{fpr_1} \text{BSROC}(x)\, dx. \tag{19}$$

We recommend to use the limits fpr_0 and fpr_1 within the observed values of $\widehat{\mathrm{FPR}}$s. The BAUC has the appealing interpretation to be the probability that in a pair of disease and non-disease subjects, the disease subject will be classified as more likely to have the disease.

We have implemented these conditional summaries in the function `bsroc()`, the function plots the study results with the fitted SROC curve, its credibility intervals and the posterior distribution of the BAUC. We illustrate this functionality in Section 4.

4. Application of bamdit in practice

4.1. Example: Diagnostic of bladder cancer

Glas, Lijmer, Prins, Bonsel, and Bossuyt (2003) performed a systematic review to investigate diagnostic procedures for tumor markers used for diagnostic of bladder cancer. One of this markers was telemerase, a ribonucleoprotein enzyme,which was evaluated in 10 studies. Riley, Abrams, Sutton, and Thompson (2007) used this example to present issues regarding boundary problems in the estimation of the correlation between random effects. Paul *et al.* (2010) illustrate the use of INLA computations in this example as well.

Looking at the data

The data of this meta-analysis can be found in the `glas` data frame in **bamdit**. We can have a quick view of the different subgroups of markers by using the function `plotdata()`, here we present some of its functionality:

```
R> library(bamdit)
R> data(glas)
R> head(glas)
```

```
  tp n1 fp  n2   Author cutoff(U/ml) marker
1  1  2 15  52 Kirollos         <NA>    BTA
2 17 60  9  70 Johnston         <NA>    BTA
3  8 28  7  34   Murphy         <NA>    BTA
4 19 47  8  30  Landman         <NA>    BTA
5 33 41 27 304     Leyh         <NA>    BTA
6  8 12 12  35    Chong         <NA>    BTA
```

```
R> plotdata(glas,                  # Data frame
+            group = glas$marker, # groupping variable
+            max.size = 5)        # scale of circles
```

We extract the subset of studies which have been reported results by using the telemerase marker:

```
R> glas.t <- glas[glas$marker == "Telomerase", 1:4]
```

and we plot this subgroup by

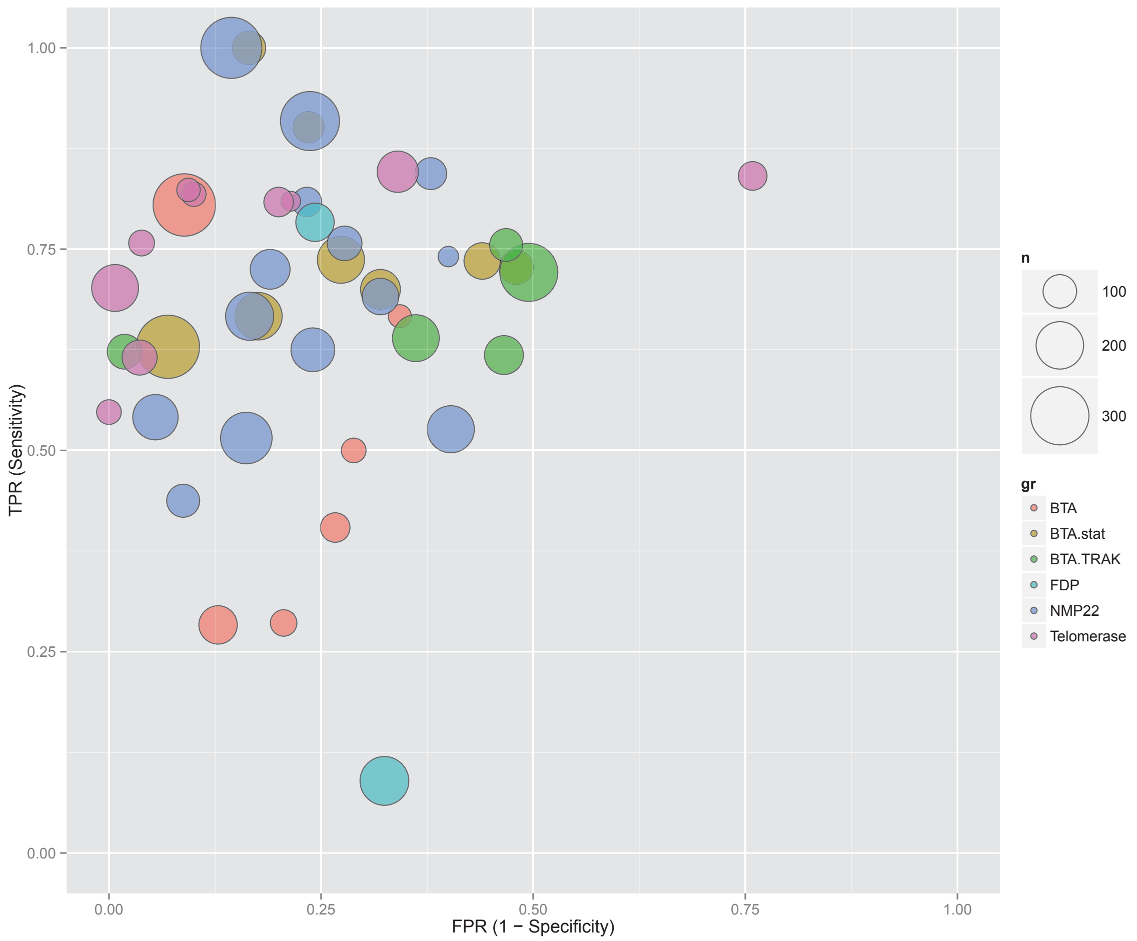


Figure 2: Display of the meta-analysis results of the data frame glas: each circle identifies the true positive rate vs. the false positive rate of each study. Different colours are used for different markers and different sizes for sample sizes.

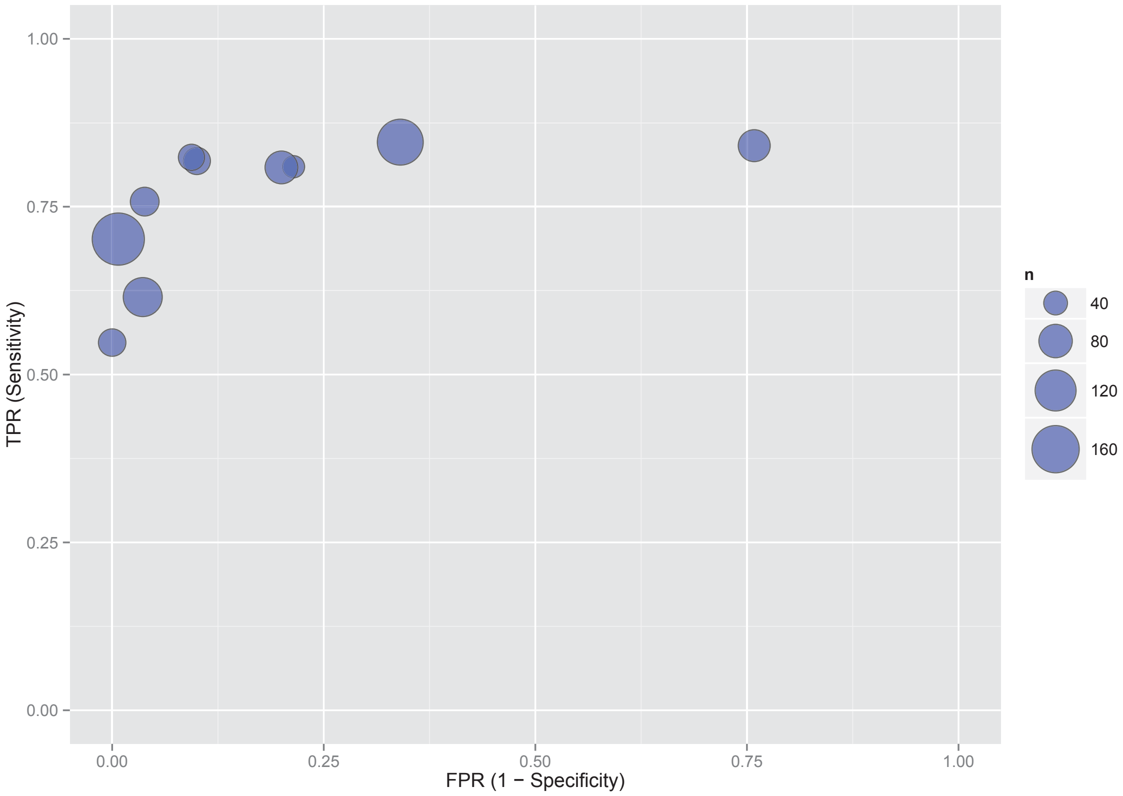


Figure 3: Display of the meta-analysis results of studies with telemerase marker in the data frame glas.

```
R> plotdata(glas.t)
```

Fitting Bayesian meta-analysis models

A single function called `metadiag()` is used to fit different type of Bayesian meta-analysis models. Below we illustrate some of the arguments of this function. For example, to fits a model, with bivariate Normal distribution with logistic link function, type:

```
R> glas.m1 <- metadiag(glas.t,                # Data frame
+                      re = "normal",         # Random effects distribution
+                      link = "logit",        # Link function
+                      nr.burnin = 1000,      # Iterations for burnin
+                      nr.iterations = 10000, # Total iterations
+                      nr.chains = 4,         # Number of chains
+                      r2jags = TRUE)         # Use r2jags as interface to jags

module glm loaded

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 208

Initializing model
```

To see the results of this computations just print the object by:

```
R> glas.m1

Inference for Bugs model at "5", fit using jags,
 4 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 4000 iterations saved
         mu.vect sd.vect   2.5%    25%    50%    75%  97.5% Rhat n.eff
fp.new     9.152  10.567  0.000  2.000  5.000 13.000 40.000 1.00  1600
mu.D       3.375   0.527  2.378  3.032  3.355  3.696  4.491 1.00  1700
mu.S      -0.926   0.761 -2.467 -1.395 -0.930 -0.446  0.544 1.00  4000
rho       -0.483   0.224 -0.822 -0.651 -0.519 -0.351  0.042 1.00  4000
se.new     0.734   0.179  0.276  0.644  0.771  0.866  0.971 1.00  2600
se.pool    0.767   0.062  0.628  0.732  0.771  0.807  0.877 1.00  1500
sigma.D    1.355   0.474  0.690  1.031  1.277  1.574  2.478 1.00  2100
sigma.S    2.191   0.643  1.280  1.748  2.079  2.521  3.737 1.00  4000
sp.new     0.817   0.204  0.226  0.745  0.898  0.961  0.997 1.01  2100
sp.pool    0.885   0.055  0.749  0.856  0.895  0.924  0.965 1.00  4000
tp.new    36.697   9.429 13.000 32.000 39.000 44.000 49.000 1.00  2400
deviance  82.065   5.998 72.456 77.705 81.317 85.709 95.620 1.00  2700

For each parameter, n.eff is a crude measure of effective sample size,
```

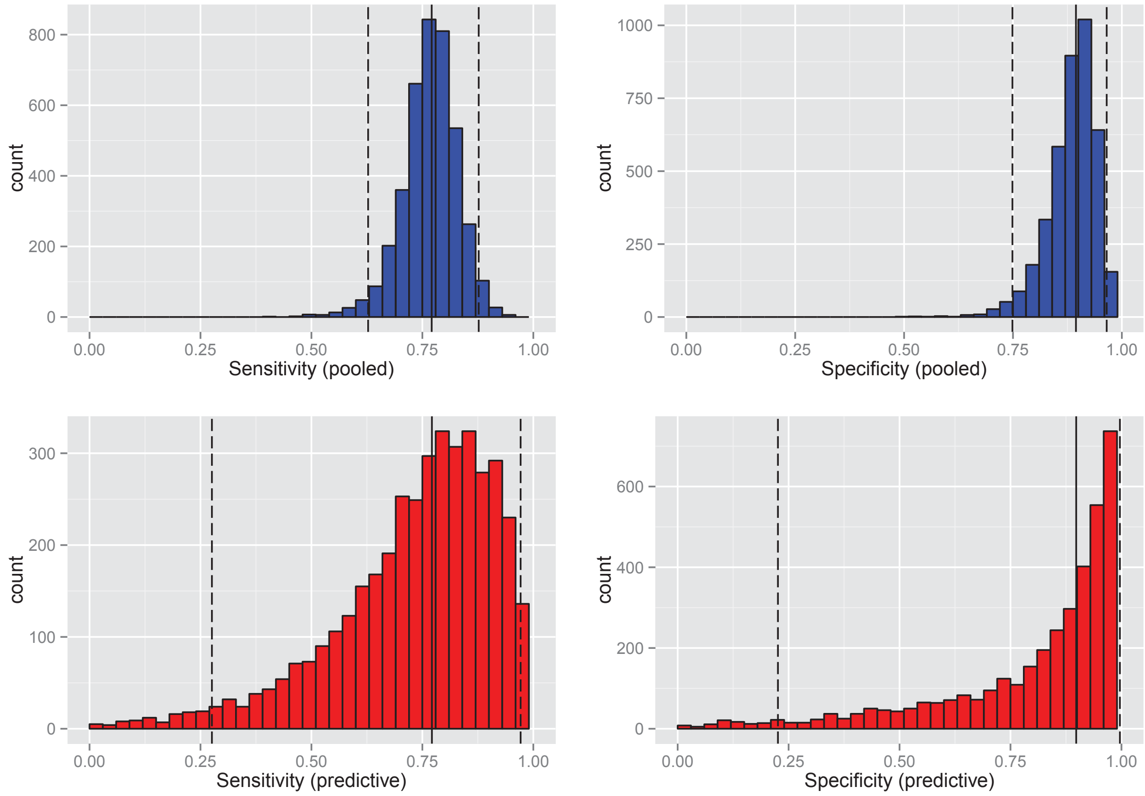


Figure 4: Results of the meta-analysis: Posterior distributions for the pooled sensitivity and specificity and their predictive posteriors.

```
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 18.0 and DIC = 100.0
DIC is an estimate of expected predictive error (lower deviance is better).
```

We can see that hyper-parameters, like the component of variances (σ_D and σ_S) and the correlation between random effects (ρ) are estimated without boundary problems.

Displaying meta-analysis summaries

The function `plotsesp()` is a user friendly function in **bamdit** which displays the posterior distribution of the pooled sensitivity and specificity and their predictive posteriors. We can display these posteriors as follows:

```
R> plotsesp(glas.m1)
```

Figure 4 shows the output, clearly the low number of studies influence the ability to predict the result of a future study.

It is very useful to display the the Bayesian Predictive Surface by contourns at different

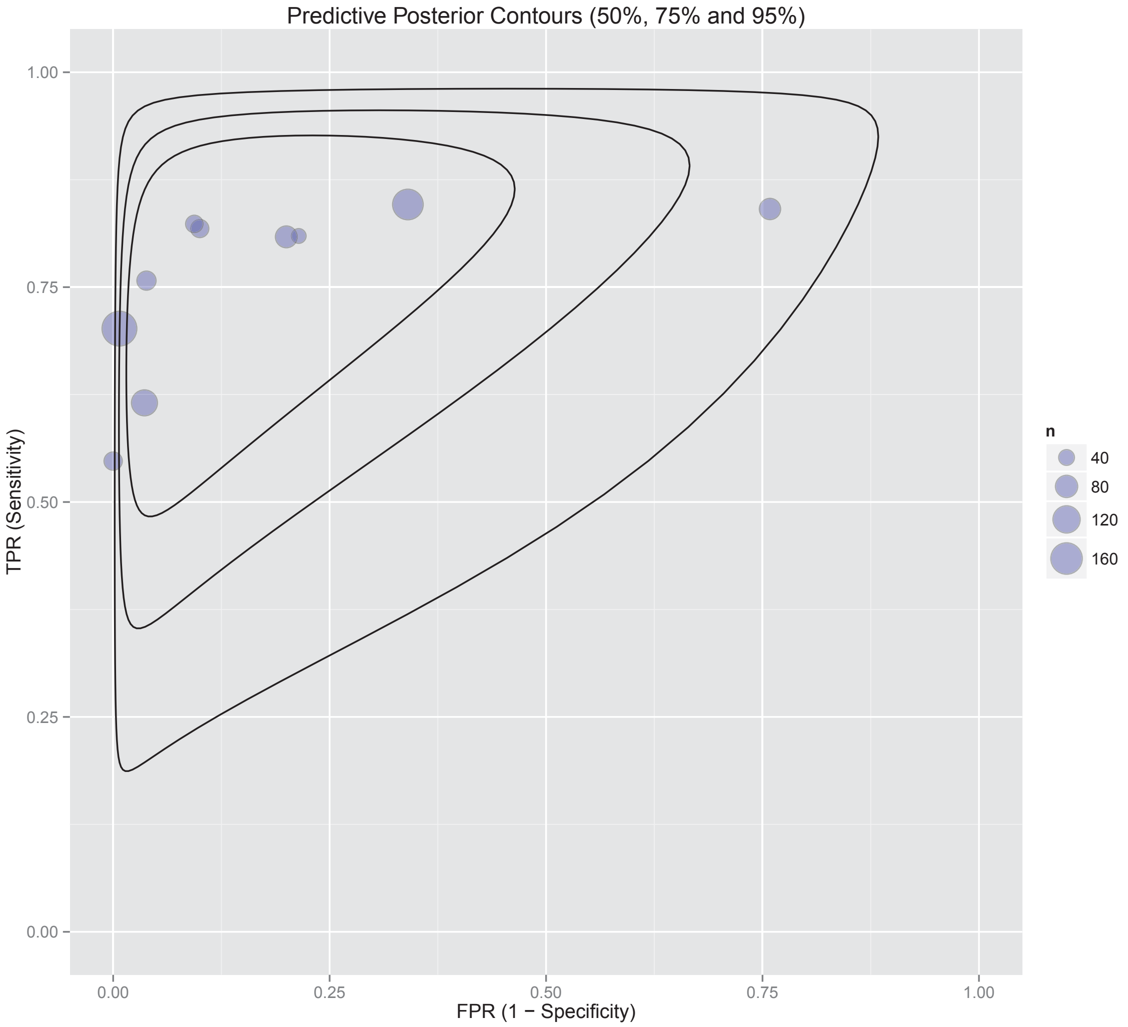


Figure 5: Results of the meta-analysis: Bayesian Predictive Surface by contours at different credibility levels.

credibility levels and compare these curves with the observed data. The function `plotcont` displays parametric and non-parametric predictive contours:

```
R> plotcont(m = glas.m1,                # Fitted model
+           data = glas.t,              # Data frame with studies' results
+           level = c(0.5, 0.75, 0.95), # Credibility levels
+           parametric.smooth = TRUE)   # Parametric curve
```

The BSROC curve and its area under the curve are useful summaries of a meta-analysis, we can easily display these summaries by using the function `bsroc()` as follows:

```
R>  bsroc(glas.m1,                      # Fitted model
+        data = glas.t,                 # Data frame with studies' results
+        level = c(0.025, 0.5, 0.975),  # Credibility levels
```

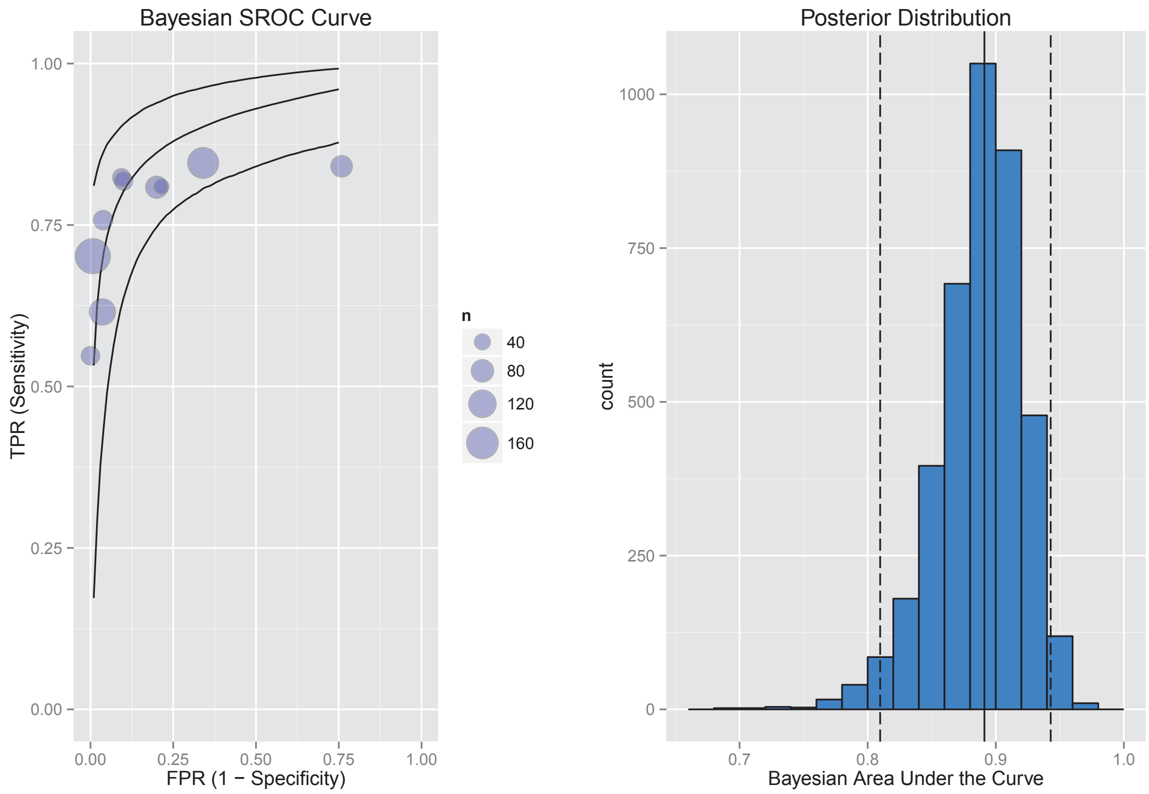


Figure 6: Conditional summaries: Left panel shows the BSROC curve, the central line corresponds to the posterior median and the upper and lower curves to the quantiles of the 2.5 and 97.5 precent respectively. The right panel displays the posterior distribution of the area under the BSROC curve.

```
+         plot.post.bauc = TRUE,     # include the posterior of the AUC
+         binwidth.p = 1/50, # histogram class length is range*binwidth.p
+         fpr.x = seq(0.01, 0.75, 0.01), # grid of values for FPR
+         lower.auc = 0, # lower limit for the BAUC
+         upper.auc = 0.99) # upper limit for the BAUC
```

```
Summary results for the Bayesian Area Under the Curve (BAUC)
------------------------------------------------------------
 2.5%   25%   50%   75% 97.5%
0.810 0.869 0.891 0.911 0.943
------------------------------------------------------------

NULL
```

Interesting, the BAUC results and the BSROC , which is display in Figure 6, show promising diagnostic ability of this marker.

Hyper-parameters posterios and checking convergence of MCMC computations

If we are interested in visualizing the posterior distributions of all hyper-parameters simultaneusly, we can use one of the alternative matrix plot function in R. For example, we can use the `ggpairs()` function from the package **GGally** as follows:

```
R> library(ggplot2)
R> library(GGally)
R> library(R2jags)

Loading required package: rjags
Loading required package: coda
Linked to JAGS 3.4.0
Loaded modules: basemod,bugs,dic,glm

Attaching package: 'R2jags'

The following object is masked from 'package:coda':

    traceplot

R> attach.jags(glas.m1)
R> hyper.post <- data.frame(mu.D, mu.S, sigma.D, sigma.S, rho)
R>
R> ggpairs(hyper.post,                    # Data frame
+ title = "Hyper-Posteriors",           # title of the graph
+ upper = list(params = c(size = 5)),  # print correlations
+ lower = list(continuous = "density") # contour plots
+                                    )
```

In Figure 7 we can also see in the lower diagonal panels the correlation structure of this multivariate posterior. Clearly hyper-parameters are uncorrelated with the exception of μ_D and μ_S.

Convergence of the MCMC computations can be investigated with the package **ggmcmc**, this package also offers an alternative to display results:

```
R> library(ggmcmc)

Loading required package: dplyr

Attaching package: 'dplyr'

The following object is masked from 'package:GGally':

    nasa

The following object is masked from 'package:stats':
```

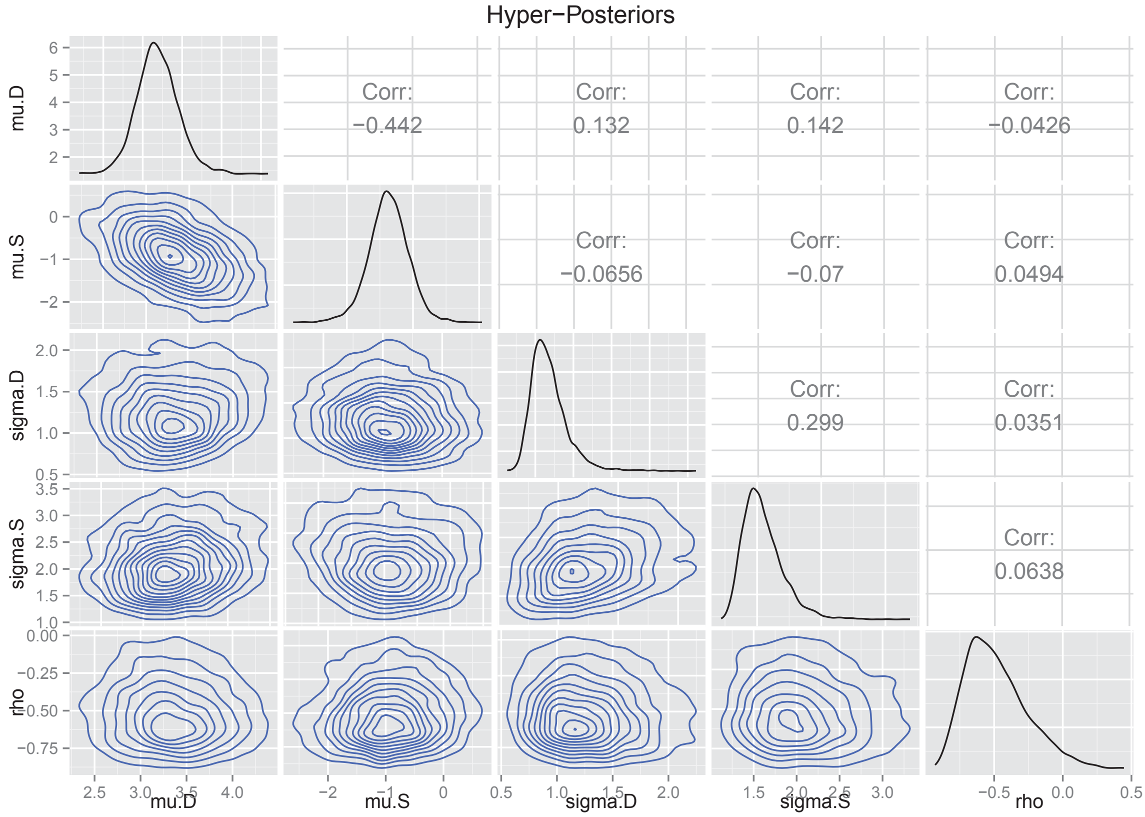


Figure 7: Posterior distributions for the hyperparameters of the model.

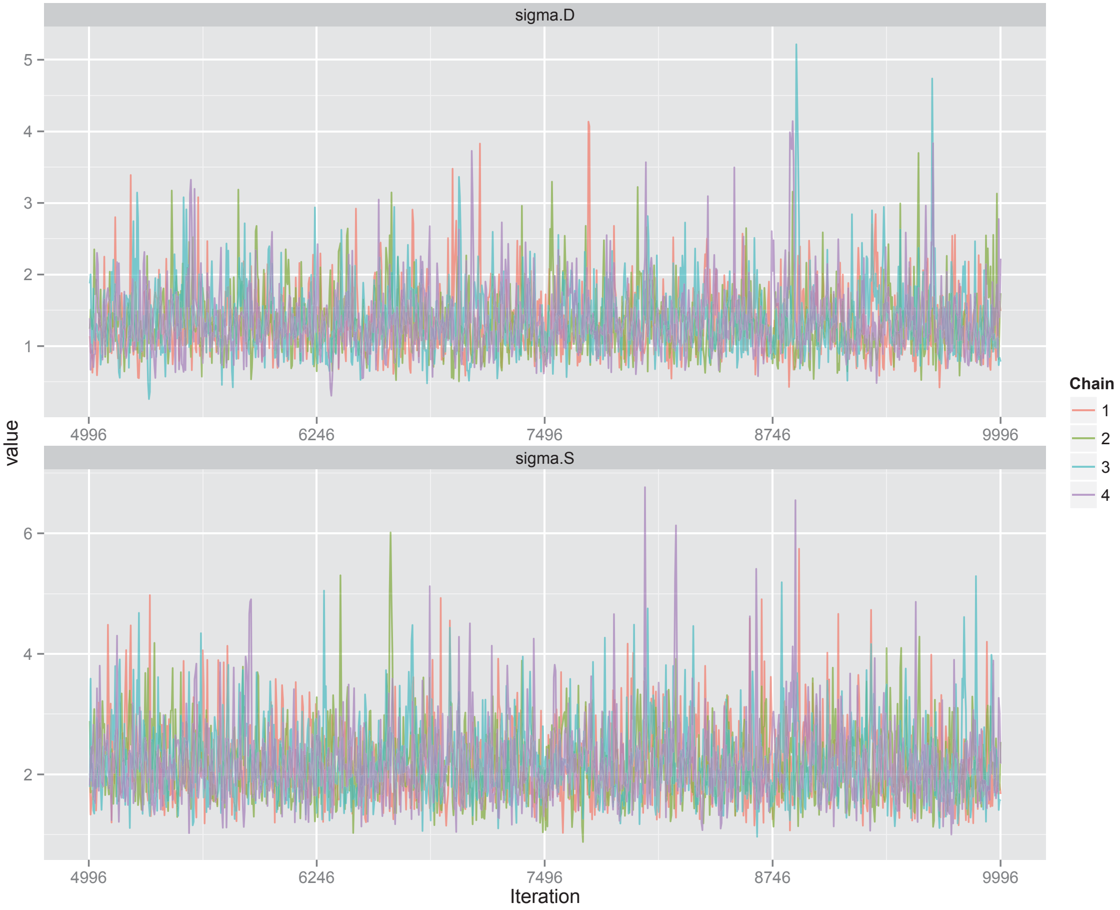


Figure 8: Trace plots for MCMC computations: Posteriors for components of variances.

```
    filter

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Loading required package: tidyr
```

```
R> out.m1 <- as.mcmc(glas.m1)
R> out.m1 <- ggs(out.m1)
R> ggs_traceplot(out.m1, family = c("sigma"))
```

Conflict of evidence analysis by using scale mixtures random-effects

We can fit a model with scale mixtures as random effects to investigate if there are conflict of evidence between the studies included in the systematic review. The following code gives an example:

```
R> glas.m2 <- metadiag(glas.t, # Data frame
+               re = "sm",     # Scale mixture of normals
+             link = "logit",  # Link function
+               df = 4,        # Degrees of freedom
+          split.w = TRUE,     # Different weights for each component
+        nr.burnin = 1000,     # Iterations for burnin
+    nr.iterations = 10000,    # Total iterations
+        nr.chains = 4,        # Number of chains
+           r2jags = TRUE)     # Use r2jags as interface to jags

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 278

Initializing model
```

The results are printed as usual:

```
R> glas.m2

Inference for Bugs model at "6", fit using jags,
 4 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 4000 iterations saved
         mu.vect sd.vect   2.5%    25%    50%    75%  97.5% Rhat n.eff
fp.new     9.008  10.659  0.000  2.000  5.000 12.000 42.000 1.00  1300
mu.D       3.405   0.497  2.476  3.085  3.393  3.722  4.433 1.00  4000
mu.S      -0.980   0.758 -2.493 -1.466 -0.979 -0.508  0.534 1.00  4000
rho       -0.462   0.231 -0.816 -0.633 -0.496 -0.330  0.078 1.00  3600
se.new     0.727   0.190  0.203  0.646  0.770  0.861  0.978 1.01  4000
se.pool    0.765   0.061  0.628  0.731  0.769  0.806  0.873 1.00  4000
sigma.D    1.138   0.426  0.521  0.843  1.064  1.350  2.185 1.00  3100
sigma.S    1.824   0.637  0.901  1.390  1.710  2.144  3.387 1.00  2100
sp.new     0.820   0.208  0.166  0.765  0.898  0.960  0.997 1.02  1500
sp.pool    0.889   0.052  0.762  0.863  0.899  0.926  0.964 1.00  4000
tp.new    36.417   9.913 10.000 32.000 39.000 43.000 49.000 1.00  4000
w1[1]      1.579   2.080  0.351  0.685  1.063  1.729  5.826 1.00  4000
w1[2]      1.556   2.322  0.341  0.670  1.045  1.714  5.418 1.00  4000
w1[3]      1.565   2.118  0.342  0.684  1.048  1.713  5.886 1.00  4000
w1[4]      1.558   1.918  0.336  0.684  1.043  1.745  5.852 1.00  2000
w1[5]      1.876   2.372  0.366  0.773  1.235  2.042  7.418 1.00  2000
w1[6]      1.493   1.705  0.339  0.660  1.003  1.659  5.726 1.00  4000
w1[7]      1.685   2.255  0.354  0.697  1.127  1.846  6.556 1.00  1900
w1[8]      1.514   1.685  0.338  0.661  1.037  1.741  5.730 1.00  4000
w1[9]      1.601   2.142  0.348  0.694  1.048  1.733  6.013 1.00  4000
w1[10]     2.481   3.632  0.401  0.910  1.517  2.755 10.386 1.00  2400
w2[1]      1.530   1.815  0.343  0.677  1.053  1.732  5.539 1.00  1500
```

```
w2[2]      1.525   1.798  0.325  0.670  1.056  1.748  5.356 1.00  3800
w2[3]      1.904   2.091  0.387  0.800  1.264  2.161  7.376 1.00  3900
w2[4]      1.800   2.222  0.370  0.765  1.194  2.047  6.622 1.00  4000
w2[5]      2.342   3.962  0.438  0.916  1.445  2.499  9.303 1.00  4000
w2[6]      1.521   2.246  0.354  0.681  1.037  1.693  5.544 1.00  4000
w2[7]      2.282   3.267  0.423  0.883  1.423  2.499  8.980 1.00   760
w2[8]      1.493   2.173  0.336  0.661  0.996  1.638  5.342 1.00  4000
w2[9]      1.433   1.677  0.337  0.656  1.006  1.639  4.933 1.00  2900
w2[10]     3.434   5.605  0.575  1.264  2.124  3.754 13.959 1.00  4000
deviance  82.030   6.109 72.127 77.626 81.454 85.471 96.029 1.00  2400

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 18.6 and DIC = 100.7
DIC is an estimate of expected predictive error (lower deviance is better).
```

Although, this model shows similar results as the model with bivariate normal random effects, there is about 5% of reduction of the standard deviations of the pool summaries and we have the additional information coming from the posterior weights. The function `plotw` plots a the posteriors of the weights:

```
R>   plotw(m = glas.m2)
```

Figure 9 summarize the results of the component weights w_1 and w_2. If the bivariate normal random effects is correct, then we expect that the posteriors are centered at 1. Studies 5 and 7 showed a moderate deviation and Study 10 a clear deviation. We can print the original data to explain these results

```
R> glas.t[c(5, 7, 10), ]

   tp n1 fp  n2
38 40 57  1 138
40 23 42  0  12
43 37 44 22  29
```

and calculate the empirical rates

```
R> dat.hat <- data.frame(tpr = glas.t[,1]/glas.t[,2],
+                        fpr = glas.t[,3]/glas.t[,4],
+                        n = glas.t[,2] + glas.t[,4])
R> dat.hat[c(5, 7, 10), ]

     tpr     fpr   n
5  0.702 0.00725 195
7  0.548 0.00000  54
10 0.841 0.75862  73
```

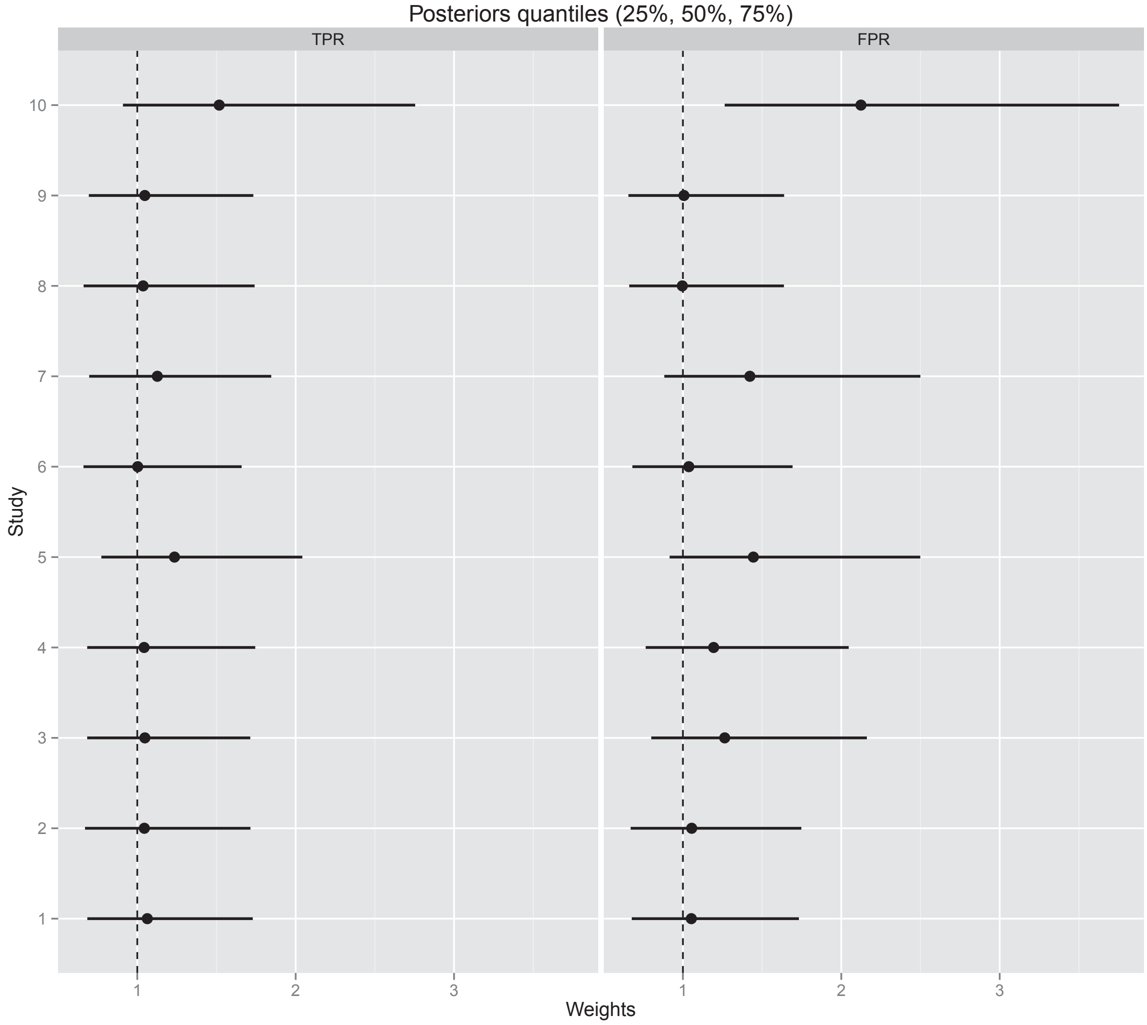


Figure 9: Posterior distributions of the component weights: It is expect that the posterior is centered at 1. Studies 5 and 7 showed a moderate deviation and Study 10 a clear deviation.

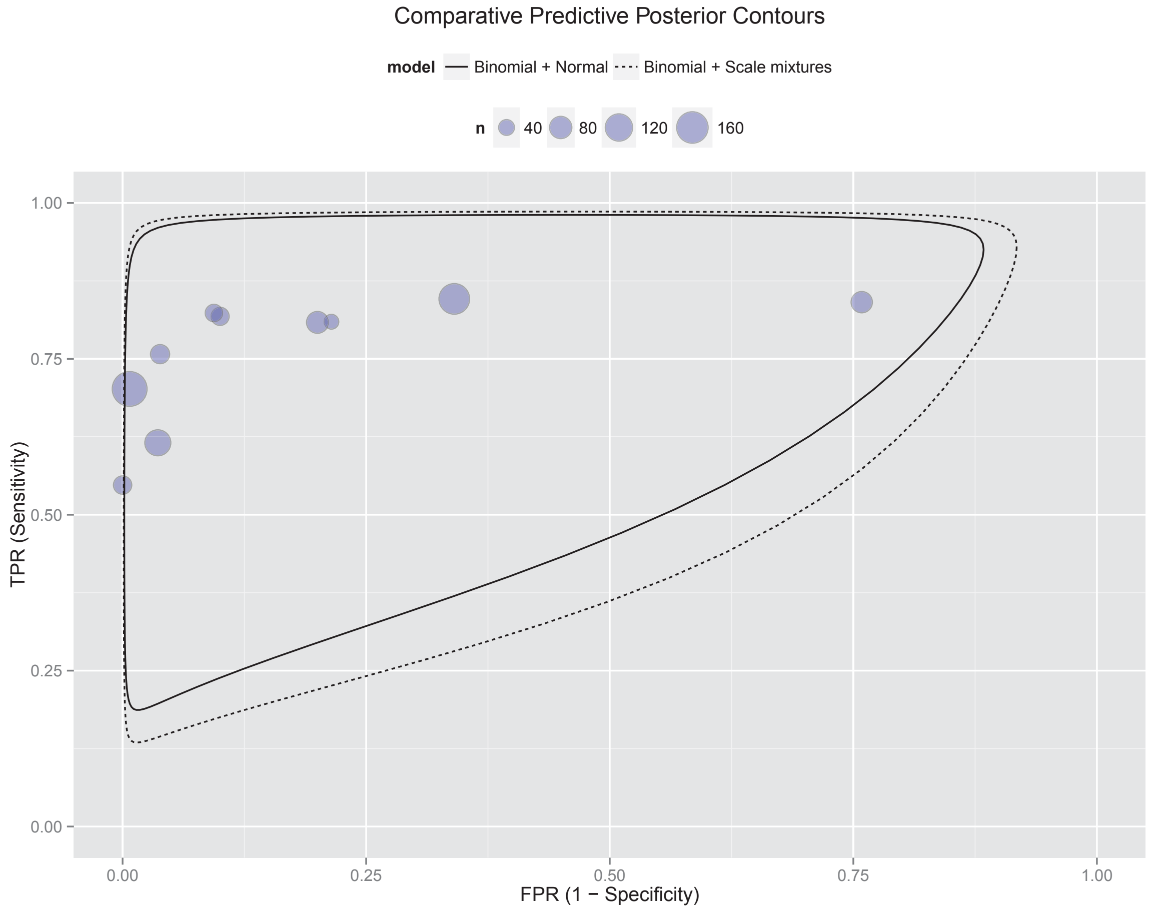


Figure 10: Comparative results of the Bayesian Predictive Surface at the 95 percent credibility level.The Normal random effects model corresponds to the solid line and the scale mixtures of random effects to the dotted line.

Studies 5 and 7 have a very low false positive rate, may be too low to be true! Study 10 has over 75% false positive rate, which is extreme for these data. We can use the function `plotcompare()` to display the differences between two models with respect to the predictive posterior contours:

```
R> plotcompare(m1 = glas.m1,                            # Model 1 object
+              m2 = glas.m2,                            # Model 2 object
+              data = glas.t,
+              m1.name = "Binomial + Normal",           # Label for Model 1
+              m2.name = "Binomial + Scale mixtures", # Label for Model 2
+              level = 0.95)
```

Figure 10 shows that the model with the scale mixture random effects extends the predictive contours in the lower direction of sensitivity and in the upper direction of the false positive rate.

Computer tomography (CT) scans in the diagnosis of appendicitis

This example refers to a meta-analysis 51 studies investigating the accuracy performance of Computer Tomography (CT) scans in the diagnosis of appendicitis Verde (2008).

One characteristic of this meta-analysis is the combination of disparate data. From the 51 studies 22 were retrospective and 29 were prospective. Verde (2008) analyzed this characteristic and found that retrospective studies had substantial more heterogeneity than prospective ones, which led to the structural dispersion model of Verde (2010a). Recently, Zhou and Dendukuri (2014) used this data to illustrate measurement heterogeneity in a bivariate random effects meta-analysis.

Looking at the data

The data of this meta-analysis can be found in the `ct` data frame in **bamdit**. In addition to the test performance results, this data frame contains information about study characteristics, patient characteristics, study design, and diagnostic setup.

```
R> data(ct)
R> gr <- with(ct, factor(design,
+                        labels = c("Retrospective study", "Prospective study")))
R>
R> plotdata(ct,               # Data frame
+          group = gr,        # Groupping variable
+          y.lo = 0.75,       # Lower limit of y-axis
+          x.up = 0.75,       # Upper limit of x-axis
+          alpha.p = 0.5,     # Transparency of the balls
+          max.size = 5)      # Scale the circles
```

Analyzing conflict of evidence of studies with different design

We analyze these data to show how to compare the posterior weights for different groups of studies. In the following example we compare these posteriors by using the function `plotw` and given to the argument `group` the factor variable which indicates if a study has prospective or retrospective design.

```
R> ct.m <- metadiag(ct,
+                re = "sm",      # Scale mixture of normals
+              link = "logit",   # Link function
+                df = 4,         # Degrees of freedom
+           split.w = TRUE,      # Different weights for each component
+         nr.burnin = 1000,      # Iterations for burnin
+     nr.iterations = 10000,     # Total iterations
+         nr.chains = 4,         # Number of chains
+            r2jags = TRUE)      # Use r2jags as interface to jags

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
```

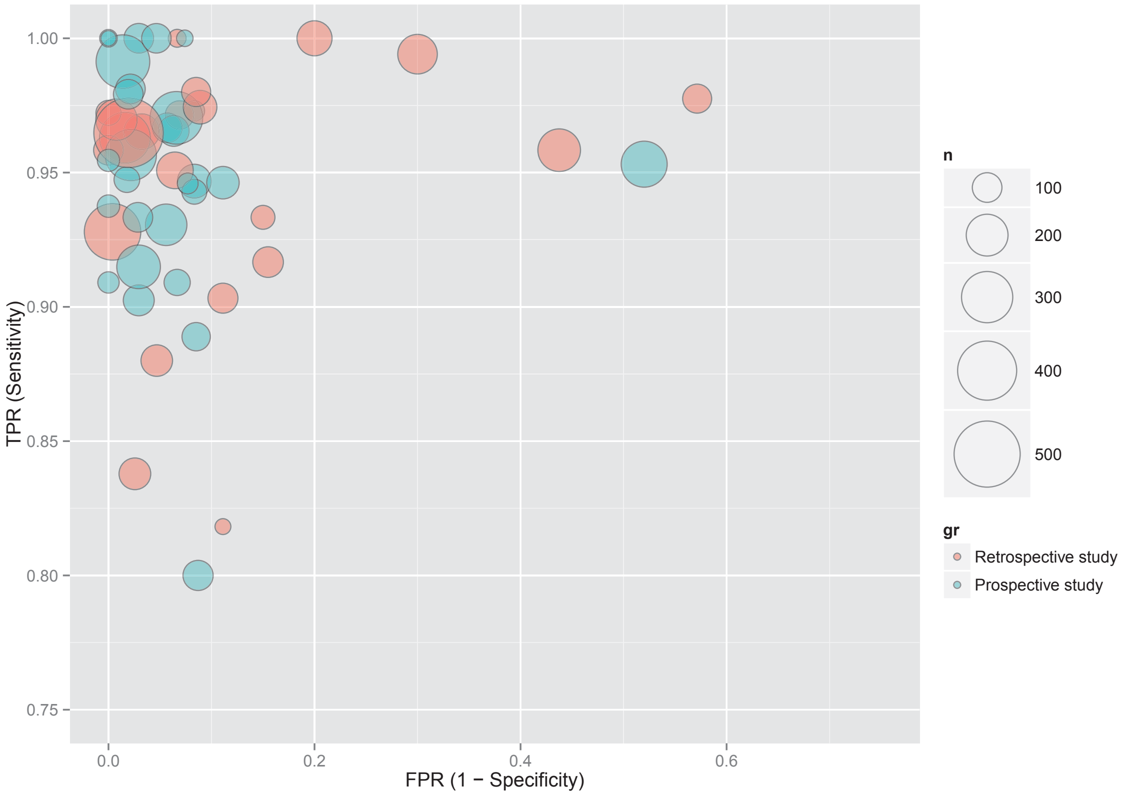


Figure 11: Display of the meta-analysis results of the data frame ct: each circle identifies the true positive rate vs. the false positive rate of each study. Different colours are used for different study designs and different sizes for sample sizes.

```
   Graph Size: 1180

Initializing model

R> plotw(m = ct.m,                # The fitted model
+       group = gr                # The groupping factor
+       )
```

Figure 12 displays the posteriors of each components' weights. The right panel shows that prospective studies number 25 and 33 deviate with respect to the prior mean of 1, while on the left panel we see that a prospective study (number 47) and five retrospective studies have substantial variability.

The function `plotcompare()` can be used to compare the predictive differences between retrospective and prospective studies:

```
R> m1.ct <- metadiag(ct[ct$design==1, 1:4]) # Restrospective studies

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 388

Initializing model

R> m2.ct <- metadiag(ct[ct$design==2, 1:4]) # Prospective studies

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 493

Initializing model

R> plotcompare(m1.ct, m2.ct, data = ct,
+             m1.name = "Retrospective design",
+             m2.name = "Prospective design",
+               group = gr,
+            limits.x = c(0, 0.75), limits.y = c(0.65, 1))
```

Finally, Figure 13 presents the 95% predictive posterior contours for studies with retrospective and prospective design, we can clearly see the effects of study design in the meta-analysis. In synthesis, retrospective studies are less specific and more uncertain than prospective ones.

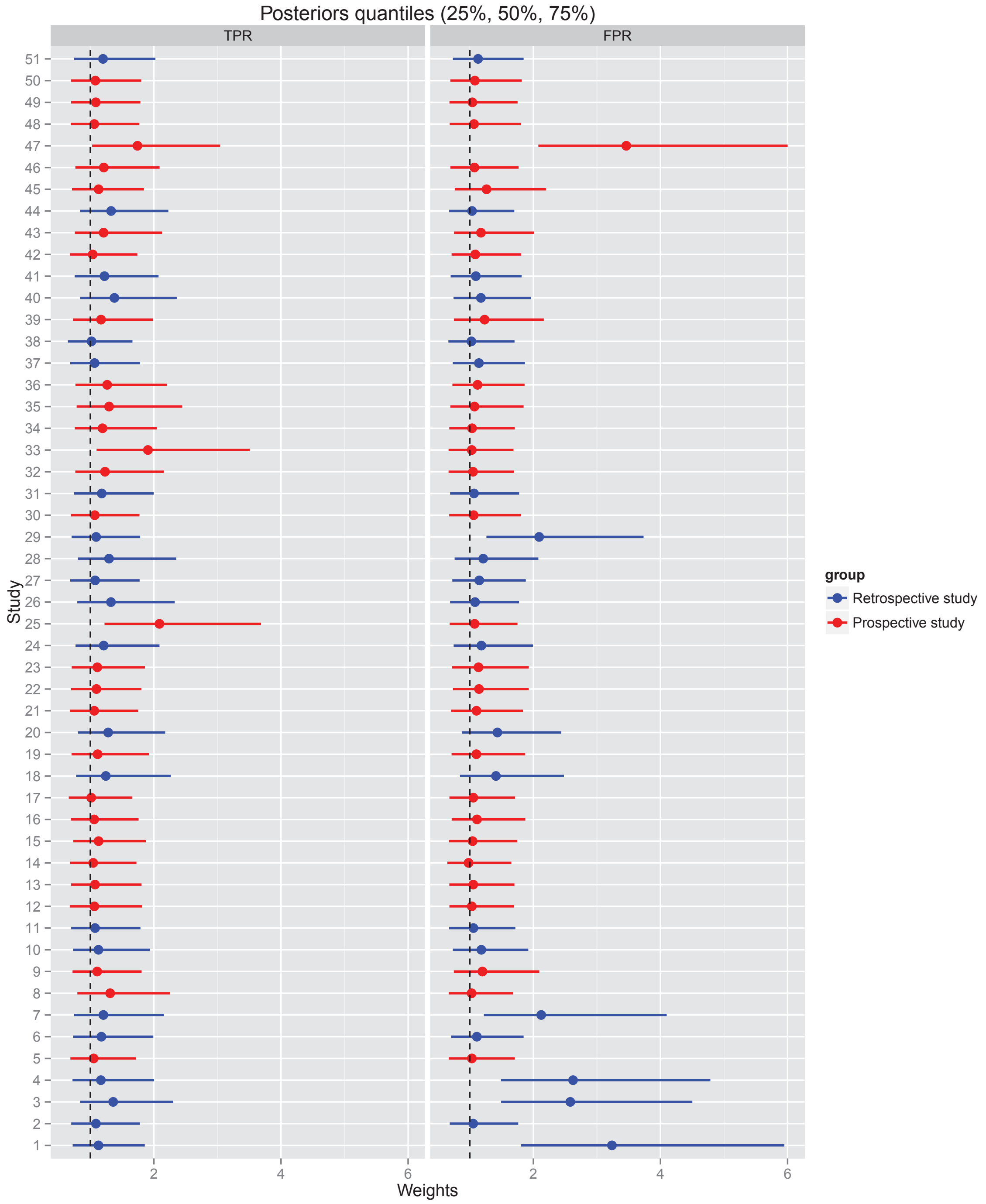


Figure 12: Posterior distributions of the component weights: It is expect that the posterior is centered at 1. Studies with retrospective design tend to present deviations in FPR.

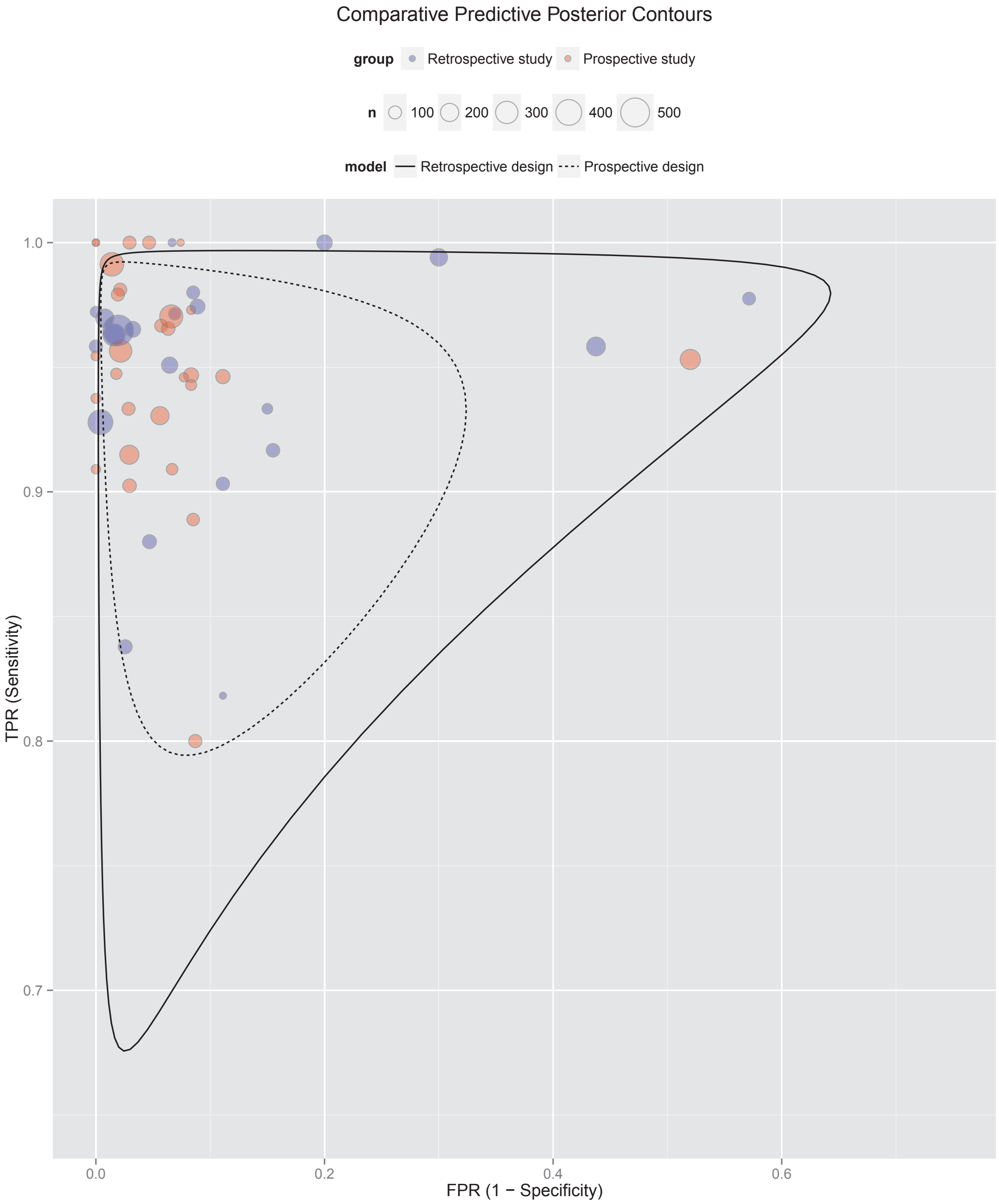


Figure 13: Predictive posteriors contours at 95 credibility level: Two models with Normal random effects are fitted to studies with retrospective (blue points) and prospective (red points) design.

5. Conclusions

When developing **bamdit**, our aim was to simplify the application of a meta-analysis model which was accessible to practitioners but which up to now had required a large amount of statistical expertise. The package implements a series of innovative statistical techniques to avoid boundary estimation of parameters, conflict of evidence and robust estimation of model parameters.

The first example in Section 4 shows that the MCMC algorithm implemented in **bamdit** outperforms a classical bivariate random effects approach based on REML estimation, which can be unreliable when the meta-analysis contains a small number of studies with a large heterogeneity (Riley *et al.* 2007). Moreover the flexible random effects distribution used in **bamdit** helps to better understand the studies' results by pointing out unusual results.

The conflict of evidence assessment is the deconstructionist side of meta-analysis, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. One possibility for this type of analysis is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. Both examples in Section 4 demonstrated that we could apply a double scale mixture of bivariate normal distributions and we made conflict diagnostics by direct interpretation of the scale weights.

One important topic currently not implemented in **bamdit** is the meta-regression and the indirect comparison of several diagnostic procedures. These topics are linked to the problematic of ecological bias and are topics of current research. However, we plan to update **bamdit** to include this functionality soon.

Acknowledgments

The author is greatful to the following talented and motivated students from the Institute of Bioinformatics at the University of Düsseldorf: Arnold Sykosch, who designed and developed the first version of the **bamdit** packge, and to Marc Daxer, who implemented new plot functions. This work was supported by the German Research Foundation project DFG VE 986/1-1.

References

Arends L, Hamza T, Van Houwelingen J, Heijenbrik K, Hunink M, Stijnen T (2008). "Bivariate random effects meta-analysis of ROC curves." *Medical Decision Making*, **28**(5), 621–638.

Buerkner P, Doebler P (2014). "Testing for publication bias in diagnostic meta-analysis: a simulation study." *Statistics in Medicine*, **33**, 3061–3077.

Chu H, Guo H (2009). "Letter to the editor." *Biostatistics*, **10**(1), 201–203.

Dewey M (2014). "CRAN Task View: Meta-Analysis." Version 2014-07-25, URL `http://CRAN.R-project.org/view=MetaAnalysis`.

Glas A, Lijmer J, Prins M, Bonsel G, Bossuyt P (2003). "The diagnostic odds ratio: a single indicator of test performance." *Journal of Clinical Epidemiology*, **56**(11), 1129–1135.

Harbord R, Deeks J, Egger M, Whiting P, Sterne J (2007). "A unification of models for meta-analysis of diagnostic accuracy studies." *Biostatistics*, **1**, 1–21.

Irwig L, Macaskill P, Glasziou P, Fahey M (1995). "Meta-analytic methods for diagnostic test accuracy." *Journal of Clinical Epidemiology*, **48**, 119–130.

Lijmer J, Bossuyt P, Heisterkamp S (2002). "Exploring sources of heterogeneity in systematic reviews of diagnostic tests." *Statistics in Medicine*, **21**, 1525–1537.

Lijmer J, Mol B, Heisterkamp S, Bonsel G, Prins M, van der Meule J, Bossuyt P (1999). "Empirical evidence of design-related bias in studies of diagnostic test." *The Journal of the American Medical Association*, **282**, 1061–1066.

Lunn D, Spiegelhalter D, Thomas A, Best N (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, **28 (25)**, 3049–3067.

Macaskill P (2004). "Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis." *Journal of Clinical Epidemiology*, **57**(9), 925–932.

Menten J, Boelaert M, Lesaffre E (2013). "Bayesian meta-analysis of diagnostic test allowing for Imperfect reference standards." *Statistics in Medicine*, **32**, 5398–5413.

Moses L, Shapiro D, Littenberg B (1993). "Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations." *Statistics in Medicine*, **12**, 1293–1316.

Novielli N, Cooper NJ, Sutton AJ, Abrams K (2010). "Bayesian model selection for meta-analysis of diagnostic test accuracy data: application to Ddimer for deep vein thrombosis." *Res. Syn. Meth.*, **1**, 226–238.

Paul M, Riebler A, Bachmann L, Rue H, Held L (2010). "Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations." *Statistics in Medicine*, **29**(12), 1325–1339.

Plummer M (2003). "JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling JAGS : Just Another Gibbs Sampler." *Proceedings of DSC*, (Dsc). URL `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf`.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL `http://CRAN.R-project.org/doc/Rnews/`.

Reitsma J, Glas A, Rutjes A, Scholten R, Bossuyt P, Zwinderman A (2005). "Bivariate analysis of sensitivity and specifity produces informative summary measures in diagnostic reviews." *Journal of Clinical Epidemiology*, **58**, 982–990.

Riley R, Abrams K, Sutton Lambert P, Thompson J (2007). "Bivariate random-effects meta-analysis and the estimation of between-study correlation." *BMC Medical Research Methodology*, **7**, 3.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F (2008). "Meta-analysis of continuous outcomes combining individual patient data and aggregate data." *Statistics in Medicine*, **27**(11), 1870–1893. ISSN 1097-0258. doi:10.1002/sim.3165. URL http://dx.doi.org/10.1002/sim.3165.

Rutter C, Gatsonis C (1995). "Regression methods for meta-analysis of diagnostic test data." *Academic Radiology*, **2**, 48–56.

Rutter C, Gatsonis C (2001). "A hierachical regression approach to meta-analysis of diagnostic test accuracy evaluations." *Statistics in Medicine*, **20**, 2865–2884.

Team RC (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Verde P (2008). "Meta-analysis of diagnostic test data: modern statistical approaches." *Deutsche Nationalbibliothek*.

Verde PE (2010a). "An introduction of Bayesian data analysis with R and BUGS: a simple worked example." *Estadistica*, **62**, 21–44.

Verde PE (2010b). "Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine*, **29**(30), 3088–3102. doi:10.1002/sim.4055. URL http://doi.wiley.com/10.1002/sim.4055.

Verde PE (2014). "A comment mentioning possible application in meta-analysis of Dirichlet t-distributions." *Bayesian Analysis*, **9**(3), 589–590.

Westwood M, Whiting P, Kleijnen J (2005). "How does study quality affect the results of a diagnostic meta-analysis?" *BMC Medical Research Methodology*, **5**, 1471–2288.

Zhou Y, Dendukuri N (2014). "Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of bibinar data: the case of meta-analyses of diagnostic test." *Statistics in Medicine*, **33**, 2701–2717.

Affiliation:

Pablo Emilio Verde
Coordination Center for Clinical Trials
University of Duesseldorf
Mooren str. 5
40225, Duesseldorf
Germany
E-mail: pabloemilio.verde@hhu.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd