University of Groningen

Making better use of clinical trials

Valkenhoef, Gerardus Hendrikus Margondus van

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2012

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Valkenhoef, G. H. M. V. (2012). *Making better use of clinical trials: computational decision support methods for evidence-based drug benefit-risk assessment*. s.n.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

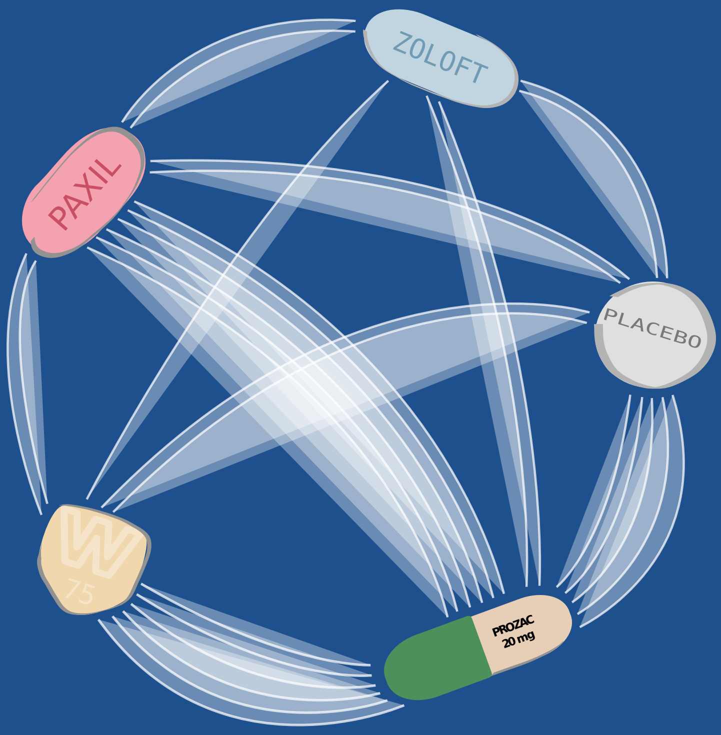
Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 04-06-2022

Making better use of clinical trials

Computational decision support methods for evidence-based drug benefit-risk assessment


ZOLOFT
PAXIL
PLACEBO
W
75
PROZAC
20 mg

Gert van Valkenhoef

rijksuniversiteit
groningen

Making better use of clinical trials

Computational decision support methods for evidence-based drug benefit-risk assessment

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
woensdag 19 december 2012
om 14:30 uur

door

Gerardus Hendrikus Margondus van Valkenhoef

geboren op 25 juli 1985
te Amersfoort

Promotores: Prof. dr. J.L. Hillege
Prof. dr. E.O. de Brock

Copromotor: Dr. T.P. Tervonen

Beoordelingscommissie: Prof. dr. A.E. Ades
Prof. dr. E.R. van den Heuvel
Prof. dr. M.J. Postma

ISBN 978-90-367-5884-0 (PDF e-book)

This thesis was produced in the context of the Escher Project (T6-202), a project of the Dutch Top Institute Pharma. The Escher Project brings together university and pharmaceutical partners with the aim of energizing pharmaceutical R & D by identifying, evaluating, and removing regulatory and methodological barriers in order to bring efficacious and safe medicines to patients in an efficient and timely fashion. The project focuses on delivering evidence and credibility for regulatory reform and policy recommendations.

The work was performed at the Faculty of Economics and Business, University of Groningen (2009 – 2010), at the Department of Epidemiology, University Medical Center Groningen (2010-2012), and during a series of research visits to the Department of Community Based Medicine, University of Bristol. The work was supported by the GUIDE institute of the Graduate School of Medical Sciences, University Medical Center Groningen.

Abstract

Drug regulators and other strategic health care decision makers assess the health impact of alternative treatment options (e.g. drug A, drug B, or no treatment at all), in addition to other potential factors such as costs. The health impact of a drug is assessed in terms of its favorable effects, or *benefits* and its unfavorable effects, or *risks*. Whether the benefits of a treatment outweigh its risks (relative to the other options), i.e. the *benefit-risk balance*, is essential to this assessment. Thus, we refer to it as *benefit-risk assessment*. In drug regulation, the benefit-risk assessment is generally based on pivotal evidence provided by clinical trials. Although various information systems store and process such evidence, there are still large gaps in the transfer of this information from the individual studies to the relevant decision makers.

First, clinical trials data are primarily disseminated in text-based reports with no formal structure. This makes the information impossible to process automatically and thus leads to difficulty in finding and making use of the relevant data (*data acquisition*). Second, current decision making in drug regulation relies entirely on the expert judgment of the assessors. Reliance on subjective assessment hides the reasoning that supports the decision and causes the regulatory process to be insufficiently transparent and traceable. Furthermore, the benefit-risk balance is seldomly made explicit, least quantified. Formal methods and computer models can facilitate the decision making process (*decision aiding*) as well as make it more explicit and transparent. Finally, while current regulatory decision making is based on the clinical trials data delivered by the company, general benefit-risk decision making requires the assessment of all viable alternatives. Such *relative* benefit-risk assessments are expected to gain importance for regulators as well. Thus benefit-risk assessment is a problem involving multiple alternatives as well as multiple criteria, and the evidence supporting it is provided by complex networks of clinical trials. Appropriate statistical methods are required to coherently integrate the available data (*evidence synthesis*). Moreover, the evidence synthesis method(s) must be compatible with the decision aiding model(s) in order to be useful.

In order to increase the transparency and quality of decision making in benefit-risk assessment all three problems must be addressed by a single integrated decision support system. The evidence synthesis and decision aiding models must fit together in an integrated work flow, and constructing them should be straightforward so that models can be created and revised on the fly for the decision at hand. This requires straightforward data acquisition, which is only possible if the evidence is available in a sufficiently structured format to enable computer support.

In this thesis we address these problems by the development of the Aggregate Data Drug Information System (ADDIS), a decision support system for evidence based strategic health care decision making. The development of the ADDIS software was carried out simultaneously with research into the appropriate data model, evidence synthesis methods and decision support models. ADDIS was developed iteratively according to agile software development principles, meaning that a working system was available throughout the project and frequent (quarterly) releases were made to the project stakeholders and the general public to ensure rapid feedback. ADDIS is working, freely available, and open source software.

The evidence synthesis problem was solved by developing algorithms to automatically generate the complex statistical models that underly network meta-analysis, a method to evaluate any number of treatments using a network of clinical trials. Previously, network meta-analysis required the manual specification of these models. We developed Multiple Criteria Decision Analysis (MCDA) methods to support drug regulatory decisions, and showed how they can be used in combination with network meta-analysis. Specifically, we used the Stochastic Multicriteria Acceptability Analysis (SMAA) method to take into account the uncertainty inherent in clinical trials information. In addition, SMAA enables decision aiding when the decision maker is not able or willing to commit to precise trade-off valuations (preferences) by allowing them to be completely or partially unspecified. Finally, we developed the ADDIS decision support system and data model, which are the result of iterative development focused on delivering evidence synthesis and decision aiding functionality rather than top-down design. While other data models for summary-level clinical trials data are in development, they have thus far not led to useful applications. On the other hand, statistical software exists for evidence synthesis of clinical trials, but their data formats are analysis-specific so extracted data can not be reused. In contrast, our data model enables data analysis and decision aiding, and that data have to be extracted only once in order to support many different analyses and decision aiding scenarios.

Samenvatting

Beleidsmakers in de gezondheidszorg, zoals registratieautoriteiten die bepalen of een geneesmiddel op de markt komt, beoordelen geregeld de effecten van verschillende behandelingsstrategieën. Hiervoor vergelijken ze een nieuw middel met een of meer bestaande middelen en/of een placebo. Bij dergelijke beslissingen staan de gezondheidseffecten van het middel centraal. Andere beleidsmakers, zoals vergoedingsautoriteiten, nemen ook andere factoren zoals de kosten en de maatschappelijke effecten van behandeling met de verschillende middelen mee in hun beslissing. Over het algemeen heeft een geneesmiddel zowel wenselijke als schadelijke gezondheidseffecten. De vraag die beleidsmakers beantwoorden is of de wenselijke effecten zwaarder wegen dan de schadelijke effecten: de *balans werkzaamheid-schadelijkheid*. De beoordeling van werkzaamheid-schadelijkheid is altijd relatief tot een of meer andere opties. Zo zal de balans voor een nieuw geneesmiddel ten minste beter moeten uitvallen dan wanneer er niet behandeld wordt. Bij markttoegang is deze beoordeling primair gebaseerd op bewijs uit gerandomiseerde klinische studies met een controlegroep (randomized controlled trials, RCTs). Hoewel verschillende informatiesystemen deze informatie opslaan en verwerken is de overdracht van informatie uit individuele RCTs naar de beleidsmakers niet efficiënt, gemakkelijk, of transparant. De beslissingen zelf zijn ook niet altijd op transparante wijze genomen of duidelijk onderbouwd.

Een aantal problemen liggen ten grondslag aan deze situatie (zie hoofdstuk 2 en appendix A). Ten eerste vindt de informatieoverdracht plaats door middel van tekstuele documenten zonder formele structuur, waardoor de informatie niet goed door computers te verwerken is. Het is dus moeilijk om de juiste informatie terug te vinden en te gebruiken (*informatieverwerving*). Ten tweede is de besluitvorming bij markttoegang grotendeels informeel: er wordt vertrouwd op het (subjectieve) oordeel van de groep experts die de beslissing neemt. Hierdoor is de onderbouwing van beslissingen soms onduidelijk en is niet transparant welke rol het bewijs uit RCTs gespeeld heeft bij de beslissing. Doordat niet expliciet gemaakt wordt hoe de wenselijke en schadelijke effecten gewogen worden in de balans werkzaamheid-schadelijkheid zijn de beslissingen onvoldoende reproduceerbaar en voorspelbaar. Formele methoden en computermodellen kunnen de besluitvorming ondersteunen en deze tegelijkertijd explicieter en transparanter maken (*beslissingsondersteuning*). Ten derde worden beslissingen nu gebaseerd op een klein aantal RCTs die door het bedrijf wat markttoegang aanvraagt uitgevoerd zijn. Meestal is er echter een uitgebreide wetenschappelijke literatuur beschikbaar waarin RCTs worden beschreven die andere behandelingsopties met elkaar vergelijken. Het is van belang dat ook deze achtergrondinformatie gebruikt wordt bij de beoordeling van een nieuw middel. In het algemeen moet de balans werkzaamheid-schadelijkheid dus bepaald worden op basis van complexe netwerken van RCTs waarin mogelijk een groot aantal middelen met elkaar vergeleken zijn. Statistische methoden zijn nodig om de informatie uit deze individuele studies op een consistente manier te combineren (*meta-analyse*). Om van praktisch nut te zijn moeten de methoden voor meta-analyse en beslissingsondersteuning goed op elkaar aansluiten. Om de transparantie en kwaliteit van beoordelingen van de balans werkzaamheid-schadelijkheid te verbeteren moeten deze problemen alle drie opgelost worden. De oplossing(en) moeten beschikbaar zijn in één enkel informatiesysteem waarin de relevante RCTs op efficiënte wijze gecombineerd kunnen worden en waarin de balans werkzaamheid-schadelijkheid op basis hiervan in kaart gebracht kan worden.

Om dit te bewerkstelligen wordt in dit proefschrift het informatiesysteem ADDIS (Aggregate Data Drug Information System) gepresenteerd (hoofdstuk 9). De ontwikkeling van ADDIS verliep parallel met onderzoek naar de juiste methoden voor meta-analyse (hoofdstuk 3-4), beslissingsondersteuning (hoofdstuk 5-8) en informatieopslag en -verwerving (hoofdstuk 2 en 9). ADDIS is iteratief ontwikkeld volgens een 'agile' proces, wat inhoudt dat gedurende vrijwel het hele project er een werkend systeem beschikbaar was en dat nieuwe versies regelmatig publiek werden gemaakt (zie ook appendix C en D). Hierdoor konden belanghebbenden in het project (en derden) doorlopend de juistheid en doelmatigheid van het systeem beoordelen. ADDIS is werkende software, gratis beschikbaar en de broncode is door iedereen in te zien en aan te passen (zie appendix B).

Contents

Title page **i**

Abstract **v**

Samenvatting **vii**

Contents **ix**

1 Introduction **1**
- 1.1 Project context 2
- 1.2 A brief history 2
- 1.3 Research questions 3
- 1.4 Outline 5

2 Deficiencies in the transfer and availability of clinical trials evidence **7**

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Deficiencies in the transfer and availability of clinical evidence in drug development and regulation. *BMC Medical Informatics and Decision Making*, 2012c. doi: 10.1186/1472-6947-12-95. (in press)

- 2.1 Background 8
  - 2.1.1 Motivation 8
  - 2.1.2 Systematic review 8
  - 2.1.3 Scope and objectives 9
- 2.2 Methods 9
- 2.3 Results 10
  - 2.3.1 Scientific literature 10
  - 2.3.2 Trial registration 11
  - 2.3.3 Systematic review 15

2.3.4 Regulatory assessment 16
2.3.5 Standards and data models 16
2.4 Discussion 17
2.4.1 Identified deficiencies 18
2.4.2 Proposed future situation 19
2.4.3 Research directions 20
2.4.4 Limitations 21
2.4.5 Conclusions 21

3 Automating network meta-analysis 23

G. van Valkenhoef, G. Lu, B. de Brock, H. Hillege, A. E. Ades, and N. J. Welton. Automating network meta-analysis. *Research Synthesis Methods*, 2012a. doi: 10.1002/jrsm.1054. (in press)

3.1 Introduction 24
3.2 Background 25
3.2.1 Consistency models for dichotomous variables 25
3.2.2 Continuous variables 28
3.2.3 Maximum likelihood estimators in single trials 28
3.3 Methods 28
3.3.1 Generating the model structure 29
3.3.2 Choosing priors 30
3.3.3 Choosing starting values 31
3.3.4 Worked example 32
3.4 Implementation 33
3.5 Results 36
3.6 Discussion 42
3.7 Appendix: proof 43

4 Automated generation of network meta-analysis inconsistency models 45

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Algorithmic parameterization of mixed treatment comparisons. *Statistics and Computing*, 22(5):1099–1111, 2012d. doi: 10.1007/s11222-011-9281-9

4.1 Introduction 46
4.2 Mixed treatment comparison models 47
4.2.1 Study level effects 48
4.2.2 Consistency models 49
4.2.3 Inconsistency models 52
4.3 Problem definition 52
4.3.1 Spanning tree selection 52
4.3.2 Baseline selection 58
4.3.3 Parameterization problem 59
4.4 The algorithm 60
4.5 Example 62
4.6 Evaluation of the running-time 64
4.7 Discussion 66

4.8 Appendix: definitions from graph theory 66

5 A stochastic multi-criteria model for drug benefit-risk analysis 69

T. Tervonen, G. van Valkenhoef, E. Buskens, H. L. Hillege, and D. Postmus. A stochastic multi-criteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428, 2011. doi: 10.1002/sim.4194

5.1 Introduction 70
5.2 Stochastic Multi-criteria Acceptability Analysis 71
5.3 A multi-criteria model for the therapeutic group of antidepressants . . 73
  5.3.1 Criteria 73
  5.3.2 Probability distributions of the criteria values 74
  5.3.3 Partial value functions 74
  5.3.4 Preference information 75
  5.3.5 Analyses 75
5.4 Discussion 79

6 Hit-and-Run for weight generation in multiple criteria decision analysis 83

T. Tervonen, G. van Valkenhoef, N. Baştürk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 2012. doi: 10.1016/j.ejor.2012.08.026. (in press)

6.1 Introduction 84
6.2 Weight constraints in SMAA 85
6.3 Hit-And-Run (HAR) for weight generation 88
  6.3.1 Sampling space transformation 88
  6.3.2 Line intersection 89
  6.3.3 Starting point 89
6.4 Convergence metrics 91
6.5 Computational tests 92
6.6 Conclusions 96

7 Complex preference information in benefit-risk assessment 99

7.1 Introduction 100
7.2 Two-dimensional analysis 100
7.3 Higher-dimensional problems 106
7.4 Discussion 107

8 Multi-criteria benefit-risk assessment using network meta-analysis 111

G. van Valkenhoef, T. Tervonen, J. Zhao, B. de Brock, H. L. Hillege, and D. Postmus. Multi-criteria benefit-risk assessment using network meta-analysis. *Journal of Clinical Epidemiology*, 65(4):394–403, 2012e. doi: 10.1016/j.jclinepi.2011.09.005

8.1 Introduction 112
8.2 Stochastic Multicriteria Acceptability Analysis 113
8.3 Mixed treatment comparison 114
8.4 MTC/SMAA for Benefit-Risk (BR) analysis 116
  8.4.1 Measurement scales 116

8.5 Application to anti-depressants 118
  8.5.1 Previous work 118
  8.5.2 Methods 119
  8.5.3 Results 120
8.6 Discussion 124
  8.6.1 Case study 125
  8.6.2 Limitations and future work 126

9 ADDIS: a decision support system for evidence-based medicine 129

G. van Valkenhoef, T. Tervonen, T. Zwinkels, B. de Brock, and H. Hillege. ADDIS: a decision support system for evidence-based medicine. *Decision Support Systems*, 2012f. doi: 10.1016/j.dss.2012.10.005. (in press)

9.1 Introduction 130
9.2 Background 131
  9.2.1 Clinical trial information systems 132
  9.2.2 Standards and data models 135
  9.2.3 Data extraction 137
  9.2.4 Evidence synthesis 138
  9.2.5 Decision models 138
9.3 The unifying data model 140
9.4 ADDIS decision support system 143
  9.4.1 Study import from ClinicalTrials.gov 144
  9.4.2 Evidence synthesis 146
  9.4.3 Benefit-Risk models 146
9.5 Discussion 152
  9.5.1 Limitations and future work 154

10 Discussion 157
10.1 Answers to research questions 157
  10.1.1 Automating network meta-analysis 158
  10.1.2 Decision analysis for drug benefit-risk assessment 159
  10.1.3 Using network meta-analysis in benefit-risk assessment 161
  10.1.4 Storing aggregated clinical trial results 161
10.2 Ongoing and future work 162
10.3 Conclusions 163

A Clinical trials information in drug development and regulation 165

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Clinical trials evidence in drug development and regulation: a survey of existing systems and standards. SOM Research Report 12003-Other, School of Management, Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands, 2012b. URL `http://irs.ub.rug.nl/dbi/4fcf224db9977`

A.1 Background 166
A.2 Clinical trial information systems 167
  A.2.1 Operational management 167
  A.2.2 Data warehousing 169

A.2.3 Trial registration . . . 169
A.2.4 Regulatory assessment . . . 170
A.2.5 Medicinal product information . . . 171
A.2.6 Standards and data models . . . 172
A.2.7 Controlled terminologies . . . 174
A.3 Discussion . . . 175

B Software deliverables 177
B.1 ADDIS . . . 177
B.2 GeMTC . . . 178
B.3 drugis.org common . . . 178
B.4 hitandrun . . . 179
B.5 odcread . . . 179
B.6 jags-jni . . . 180

C Product and Release Planning Practices for Extreme Programming 181

G. van Valkenhoef, T. Tervonen, E. O. de Brock, and D. Postmus. Product and release planning practices for extreme programming. In *Proceedings of the 11th International Conference on Agile Software Development (XP2010)*, Trondheim, Norway, 2010. doi: 10.1007/978-3-642-13054-0_25

C.1 Introduction . . . 182
C.2 Rolling Forecast for Product Planning . . . 182
C.3 Supporting Release Planning Model . . . 183
C.4 Real-Life Example . . . 185
C.5 Conclusions . . . 186

D Quantitative release planning in Extreme Programming 187

G. van Valkenhoef, T. Tervonen, B. de Brock, and D. Postmus. Quantitative release planning in extreme programming. *Information and Software Technology*, 53(11):1227–1235, 2011. doi: 10.1016/j.infsof.2011.05.007

D.1 Introduction . . . 188
D.2 Release planning model . . . 189
D.2.1 Planning in XP . . . 190
D.2.2 Model overview . . . 190
D.2.3 Discussion . . . 192
D.3 Nested knapsack formulation . . . 193
D.4 Theme valuation . . . 195
D.5 Velocity estimation . . . 196
D.6 Example . . . 197
D.7 Computational tests . . . 200
D.8 Conclusions . . . 200

Acknowledgements 203

Bibliography 205

CHAPTER 1

Introduction

Health care policy decisions often involve assessing the performance of alternative treatment options. Ideally, such decisions should be informed by all available high quality evidence, which is usually provided by clinical trials. However, there is a large gap between the available evidence published as articles in scientific journals and the integrated overview of current scientific knowledge that the decision maker requires. To close this gap, three challenging and time-consuming steps must be taken:

1. Data acquisition: collecting the relevant clinical trials
2. Evidence synthesis: statistically combining the evidence
3. Decision aiding: giving insight in the data and identifying trade-offs

Currently, each of these steps is performed separately or sometimes not at all. A limited budget or lack of expertise may mean that some steps are bypassed, leading to less transparent, lower quality decisions that may be based on only part of the relevant evidence. An integrated approach to evidence-based decision making is needed to increase the quality of decision making, to take into account all available evidence, and to enable transparency through explicit coupling of decision models with the underlying evidence from clinical trials.

This thesis shows how statistical methods for evidence synthesis and models for Multiple Criteria Decision Analysis (MCDA) can be applied to support benefit-risk decision making by drug regulatory authorities and other decision makers. The computational methods developed were implemented in an integrated decision support system, Aggregate Data Drug Information System (ADDIS). ADDIS enables basing decision models directly on an underlying database of aggregated clinical trials data.

In this manner, ADDIS improves the transparency and reproducibility of benefit-risk assessment by allowing decisions to be traced back to the underlying evidence from clinical trials. Importantly, automating these features enables an on-demand approach to decision modeling that was not previously possible.

In the following sections, first the context and history of the project that culminated in this thesis are briefly described. Then, the research problems that were addressed are explained. Finally, an outline of the thesis is given. The background literature is discussed in Chapter 2.

1.1 Project context

My PhD project was funded by the Escher project of the Top Institute Pharma (TI Pharma). TI Pharma is a Dutch public-private partnership, financing research projects with funding from government, pharmaceutical companies and academic institutes. The Escher project aims to generate new insight into the drug regulation process and to produce innovations that can lead to more efficient drug development, more transparent decision making, increased public trust in industry and regulators and the removal of unnecessary barriers to market entry of innovative new medicines. The project consists of three main work packages, with 16 PhD projects divided among the three. The first work package investigates the regulatory system itself, in order to identify unnecessary barriers to market access and opportunities for innovation, and consists of 6 PhD projects. The second work package is focussed on developing innovative methods of testing and monitoring the efficacy and safety of drugs, and consists of 8 PhD projects. The third work package is concerned with knowledge transfer and preservation, learning, and decision making based on clinical data. This final work package consists of 2 PhD projects, 3.1 and 3.2, of which this PhD project is work package 3.2.

The aim of Escher work package 3.2 is to develop an information system that streamlines the handling of clinical data submitted to regulators and the execution and reporting of the benefit-risk assessment that is based on that data. Regulatory data should transition from document-based repositories to structured databases of clinical and assessment data that enable cross-study analyses.

1.2 A brief history

When I joined the project in April 2009, the Escher project had been running for nearly a year. An overview of drug information systems had already been produced [Tervonen et al., 2010], and the team was able to point me towards the right literature, providing me with a flying start. In addition, several rounds of discussions with stakeholders (e.g. MSD, GSK, and assessors from CBG and EMA) were conducted, and a list of 16 global requirements had been composed. Because the project was originally formulated as a software development project without clear research related deliverables it was decided that the best course of action would be to develop a prototype as quickly as possible. The prototype could then function as a basis for discussion with the various stakeholders, and we expected that research questions would emerge naturally from its development. After a month of literature catch-up and intense discussions, we identified the three key elements for the prototype:

1. A structured database of clinical trials, with aggregated data
2. The synthesis of data from multiple clinical trials using meta-analysis
3. Formal decision modeling based on the clinical evidence using MCDA

The initial prototype, ADDIS 0.2, was released on 2009-06-30 and consisted of rudimentary implementations of each of the key elements. Several problems emerged immediately from our initial implementation: (1) standard meta-analysis techniques do not extend easily to decision problems involving more than two alternatives and more advanced techniques require manual model specification, and (2) the results of a meta-analysis are not directly transferable to a decision analytic environment. These problems gave rise to the research questions described in the next section.

During the initial months, we spent a substantial part of our time designing the software and writing code. It became clear that the resources available were not sufficient to address both the research questions and the software requirements. Fortunately, starting October 2009, we could attract several part-time software developers to take over most of the development workload, which permitted both research and software development to proceed in parallel, and provided me with the unique opportunity of managing my own software development team. Since the requirements were not clearly defined and preliminary, we decided to work according to an Agile software development process, which emphasizes responding to change rather than following a set-in-stone plan. We spent a fair amount of time researching Agile software development processes, which also resulted in two publications.

A collaboration with the Multi-Parameter Evidence Synthesis group at Bristol University was established in the first half of 2010 to support the work on advanced meta-analytic techniques. This collaboration gave rise to Chapter 3 and many improvements to the software, and is still ongoing.

1.3 Research questions

Based on the intended key elements of the ADDIS prototype, the main research question can be formulated as follows:

> How can a network of clinical trials be used to inform benefit-risk assessment in a formal decision modeling framework, and how can an information system support such decision modeling?

To apply decision modeling to a network of clinical trials, the results of these trials must first be combined in a methodologically sound and consistent framework. Network meta-analysis is a family of evidence synthesis methods that enables the simultaneous synthesis of complex networks of clinical trials comparing two or more alternatives [Lu and Ades, 2004, Salanti et al., 2008a]. However, performing a network meta-analysis requires manually specifying a complex statistical model. This is a potential source of errors and requires specific knowledge of the statistical software being used that is not directly relevant to network meta-analysis itself. More importantly, to build upon network meta-analysis in a decision support system, it must be relatively straightforward to apply. Therefore, the first sub-question is:

> How can the statistical models for network meta-analysis be generated automatically, based on the network of clinical trials data?

While network meta-analysis provides a consistent account of what the data tells us about the performance of the alternative drugs on a single clinical outcome, most drugs do have an impact on multiple clinical outcomes. Benefit-risk decisions usually involve at least one beneficial outcome (benefit criterion) and several adverse outcomes (risk criteria). Thus, a framework for providing insight in the data and identifying trade-offs between different criteria is needed. Supporting decisions in health care is complex, because the underlying evidence is inherently uncertain, and trade-offs are often difficult to quantify. Accepting MCDA as the proper framework for decision making involving multiple criteria [Keeney and Raiffa, 1976] raises the second sub-question:

> How can the MCDA framework be applied to drug benefit-risk assessment, while properly accounting for the uncertainty inherent in data obtained from clinical trials, and acknowledging the difficulty of precisely quantifying trade-offs?

Furthermore, building a decision model based on (network) meta-analysis creates an additional problem. Meta-analysis results in estimates of *relative* effects for reasons of statistical robustness, but *absolute* treatment effects are needed for decision making [Egger et al., 1997]. This is especially important if trade-offs between multiple criteria are to be considered. Therefore, the results of meta-analysis are not directly amendable to decision modeling, and the third sub-question is:

> How can the results of (network) meta-analysis be used to inform drug benefit-risk decision making in the MCDA framework?

Finally, once a sound and feasible framework for decision aiding based on networks of clinical trials has been established, it should be implemented in a decision support system to facilitate *on demand* decision modeling. Thus, the fourth sub-question is:

> How can aggregate clinical trial results be stored in an efficient and meaningful way, and how can automated decision modeling be applied to those results?

1.4 Outline

This thesis consists of an introduction, eight chapters (seven of which correspond to a published or submitted journal article), a discussion, an appendix providing additional background on clinical trial information systems (published as a research report), an appendix describing the software that was produced, and two appendices on software development methodology (published as a conference and a journal article).

Chapters 2–9 are the main body of the thesis. Chapter 2 presents a review of information systems and data representation standards dealing with aggregate data from clinical trials, and identifies problems in the dissemination and subsequent use of clinical trials results. The currently open problems provide the motivation for the development of ADDIS: there are clear gaps to be filled, and ADDIS addresses some of these problems.

Then, Chapter 3 and Chapter 4 show how network meta-analysis models can be generated automatically, a core ingredient in the ADDIS decision support system. Chapter 5 takes a step back and considers how to apply MCDA, specifically the Stochastic Multicriteria Acceptability Analysis (SMAA) method, to drug benefit-risk assessment in the context of a single clinical trial. Chapter 6 describes a method for sampling from the distribution of possible weights that satisfy preference information given by the decision maker. How this method can be used to better model the decision maker's preferences is briefly illustrated in Chapter 7. The pieces are put together in Chapter 8, which addresses the problem of using a SMAA model for drug benefit-risk assessment based on meta-analysis of a larger number of trials. Chapter 9 describes the ADDIS decision support system itself, putting all the pieces together in a single integrated workflow, and showing how ADDIS supports benefit-risk decisions. Overall conclusions and future research directions are discussed in Chapter 10.

Appendix A provides an account of clinical trials information systems used in drug development and regulatory submission. Then, Appendix B is an overview of the software that was delivered during the research project, consisting of the ADDIS software itself as well as several supporting libraries and programs. Finally, Appendices C and D present work on planning agile software development projects, specifically those using the eXtreme Programming (XP) methodology.

CHAPTER 2

Deficiencies in the transfer and availability of clinical trials evidence

A survey of existing systems and standards

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Deficiencies in the transfer and availability of clinical evidence in drug development and regulation. *BMC Medical Informatics and Decision Making*, 2012c. doi: 10.1186/1472-6947-12-95. (in press)

Abstract

Background: Decisions concerning drug safety and efficacy are generally based on pivotal evidence provided by clinical trials. Unfortunately, finding the relevant clinical trials is difficult and their results are only available in text-based reports. Systematic reviews aim to provide a comprehensive overview of the evidence in a specific area, but may not provide the data required for decision making.

Methods: We review and analyze the existing information systems and standards for aggregate level clinical trials information from the perspective of systematic review and evidence-based decision making.

Results: The technology currently used has major shortcomings, which cause deficiencies in the transfer, traceability and availability of clinical trials information. Specifically, data available to decision makers is insufficiently structured, and consequently the decisions cannot be properly traced back to the underlying evidence. Regulatory submission, trial publication, trial registration, and systematic review produce unstructured datasets that are insufficient for supporting evidence-based decision making.

Conclusions: The current situation is a hindrance to policy decision makers as it prevents fully transparent decision making and the development of more advanced decision support systems. Addressing the identified deficiencies would enable more efficient, informed, and transparent evidence-based medical decision making.

2.1 Background

2.1.1 Motivation

Health care policy decision makers such as drug regulatory authorities, reimbursement policy makers and guideline committees routinely evaluate the efficacy and safety of medicines, as well as other factors such as costs. Clinical trials provide the pivotal evidence for drug efficacy and safety. The ability to efficiently identify and make use of the results of existing clinical trials is critical to evidence-based policy decision making.

Until recently, journal publications were the only generally available source of trial designs and results. Thus, systematically reviewing the medical literature for the clinical trials that address a specific topic is of central importance to evidence-based health care policy [Chalmers, 2007, Sackett et al., 1996]. This provides decision makers with a coherent overview of the current evidence, and also helps to set the agenda for future clinical research [Mulrow, 1994, Sutton et al., 2009]. However, systematic reviewing is currently not feasible for most decision makers, because it is time consuming and expensive.

Therefore, most decision makers will have to rely on published systematic reviews. However, this is problematic because the review may not match the needs of the decision maker. Thus, even when a relevant systematic review is available, there may be a need to go back to the underlying trial data, especially for quantitative decision modeling. It may additionally be necessary to update or extend the review, or to combine several of them. Doing so also requires access to the underlying trial data, but these are not commonly reported. This is a serious limitation to the efficiency of both evidence-based decision making and systematic reviewing.

Thus, the quality of health care policy could be improved if systematic reviews could be performed for whatever decision is currently at hand, ideally even on demand. This would require enormous improvements to the manner in which clinical trials evidence is made available. Although efforts to standardize the information systems for the management and regulatory submission of clinical trials have been successful [Bleicher, 2003, El Emam et al., 2009, van Valkenhoef et al., 2012b], this has not so far resulted in similar improvements in the dissemination of clinical trial evidence. A comprehensive overview of the various information systems that store and process clinical trials information could identify the gaps in information transfer that limit the efficiency of systematic reviews and consequently health care policy decision making.

2.1.2 Systematic review

The need to identify and summarize the evidence for decision makers is evident from the sheer scale of the available information: PubMed alone indexes nearly 20 million publications from over 5,500 journals, and this is only a selected subset of the biomedical literature [Mulrow, 1994]. Systematic review addresses this need and consists of three steps: literature screening, data extraction, and reporting. In the first step, literature databases are searched, yielding a set of potentially relevant publications. These are screened for suitability, which results in them being included in, or excluded from, the review. Because literature searches are often inaccurate, thousands of publications may need to be screened. Moreover, to ensure comprehensiveness and avoid bias, multiple databases have to be searched [Crumley et al., 2005] and multiple publications of a single trial have to be identified as such. Once the relevant trials have been identified and the corresponding reports retrieved, the data have to be extracted from the reports. Finally, the collected data are summarized and combined (e.g. using meta-analysis), and reported in a journal article or a technical report. Typically only this final product is made available, even though making the results of the screening step and the extracted data available would greatly enhance the efficiency of future systematic reviews and decision making. Thus, to assess the efficiency of clinical trials results dissemination, systematic review should not be ignored.

The difficulty of performing a systematic review also impacts the quality of systematic reviews themselves: it leads to reviews that focus on a single treatment or a pair of treatments. Consequently, for one particular therapeutic indication many competing reviews may be available, that each provide only a small part of the overall picture [Caldwell et al., 2010]. This has led to 'overviews of reviews' or 'umbrella reviews' summarizing the results of several existing reviews [Ioannidis, 2009]. Umbrella reviews generally merely repeat the pooled summaries of treatment effects from the original reviews, but it has been argued that they may lead to misleading and inconsistent conclusions [Caldwell et al., 2010]. An approach based on the individual studies is therefore preferable but labor-intensive if the data are not available in a structured format.

2.1.3 Scope and objectives

The aim of this paper is to identify opportunities to enhance the efficiency of systematic review and evidence-based decision making, supported by a broad and useful overview of the current state of the art in the transfer and availability of clinical trial evidence. To these ends, we provide a critical overview of existing systems and standards that support the dissemination of clinical trial results.

Because publicly available clinical trial results are nearly always aggregated (at the population level) rather than reported per patient, and because most decision makers base their decisions on such data, we limit the scope of this paper to systems and standards for the aggregate level.

2.2 Methods

We included academic publications and websites of manufacturers or standardization bodies that describe information systems or standards that deal with the transfer and availability of aggregate-level results of clinical trials. We also considered review articles and peer-reviewed position papers related to such information systems.

We identified relevant publications through key word searches using Google, Google Scholar, ISI Web of Science and PubMed (last searched May 2011). We also

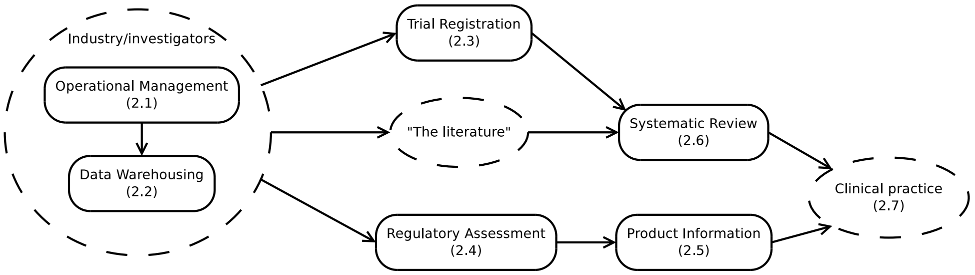


Figure 2.1: An overview of the processes dealing with clinical trials information and how they relate to each other.

screened the reference lists of included publications. In addition, through our participation in the Escher project of the Dutch Top Institute Pharma (TI Pharma), we were able to engage in discussions with many experts from the pharmaceutical industry, regulatory authorities, and academia.

Publications (both peer-reviewed articles and web pages published by companies or standardization bodies) were screened for eligibility using titles and abstracts (if applicable). Potentially relevant publications were read in full. If a peer-reviewed article and a web page conveyed (nearly) identical information, only the peer-reviewed work was included. Moreover, web pages were excluded if the source was not considered authoritative for the subject matter. Included publications were summarized using keywords, and especially important sections were highlighted for later reference. For each system and standard we collected the context in which it is used, its purpose, its defining features, the types of data it handles and/or produces, its connection to other systems or standards, and expected future developments.

2.3 Results

In this section, we present the identified systems grouped according to the processes they support. These are publication in the scientific literature (Section 2.3.1), trial registration (Section 2.3.2), systematic review (Section 2.3.3) and regulatory assessment (Section 2.3.4). Figure 2.1 shows how these processes relate to each other, to the operation of the trial itself, and to policy decision making. Standards and data models relevant to the dissemination of aggregate level results of clinical trials are discussed in Section 2.3.5.

2.3.1 Scientific literature

Pharmaceutical industry and other investigators may choose to summarize selected results of clinical trials in manuscripts submitted to peer-reviewed scientific journals. A clinical trial may result in any number of publications, from none to dozens. Unfortunately, such publications frequently lack sufficient information to allow the reader to judge whether the trial was rigorously conducted. The CONSORT statement [Begg et al., 1996, Schulz et al., 2010] aims to improve the situation by providing guidance on the proper reporting of clinical trials. Nevertheless, trial reporting is often still inadequate [Chan and Altman, 2005] and selective outcome reporting is common [Chan et al., 2004].

Reporting clinical trial results in text-based articles rather than properly structured data sets makes computational processing of the results practically impossible [Sim et al., 2000]. Moreover, although peer review is essential to guarantee the quality of such articles, publication in scientific journals scatters the results throughout the scientific literature. The problem of identifying and accessing clinical trial publications is addressed by abstract databases and search engines, most notably PubMed (`http://pubmed.com/`), backed by the MEDLINE database, and EMBASE (`http://embase.com/`), which maintains its own database; see Table 2.1 for their coverage. Both databases label abstracts using controlled vocabularies, allowing the restriction of searches to clinical trials, controlled clinical trials, or systematic reviews of clinical trials. However, not all abstracts might be labeled, which is why systematic reviewers often broaden their search beyond these categories. PubMed's Clinical Queries (`http://pubmed.com/clinical`) provide optimized search strategies that have been empirically validated for this use [Haynes et al., 2005]. It might be useful in these cases to annotate which abstracts *do not* belong to reports of clinical trials. Research is ongoing to improve the accuracy of search results through text processing of abstracts [Boudin et al., 2010] and to aid the screening process by ranking the search results [Karimi et al., 2010]. The Cochrane CENTRAL database of trial publications [Dickersin et al., 2002], kept up-to-date by the non-commercial Cochrane Collaboration (`http://www.cochrane.org/`), provides links to publications known to describe randomized controlled trials. It is quarterly updated from MEDLINE, EMBASE, and databases of specialized Cochrane groups [The Cochrane Foundation, 2010] Records that are identified as clinical trials are reported back to MEDLINE.

When clinical trial results were only published in scientific journals, the decision whether to publish the results was completely left to the investigator, which led to incomplete trial reporting and publication bias [Dickersin and Rennie, 2003]. For example, over half of the clinical trials that supported successful new drug applications made to the Food and Drug Administration (FDA) had still not been published 5 years after the medicines' market approval [Lee et al., 2008]. This is a serious problem that can lead to incorrect conclusions from a systematic review.

2.3.2 Trial registration

As early as 1986 the registration of trials in advance was proposed as a solution to publication bias [Simes, 1986]. The *trial bank* concept proposes to take the registration of trials even further by recording not only the existence of a trial, but also the study protocol (in advance) and the results (after completion) in a "machine readable" way [Sim, 1997, Sim et al., 2000]. For trial banks to be successful, all trials must be entered in a way that conforms to a single machine-readable data model [Sim, 1997, Sim et al., 2004]. The Global Trial Bank project was set up in 2005 to create a practically usable trial bank [Sim and Detmer, 2005], but in 2008 it was put on hold due to lack of

	all records	**since 2000**	**source**
PubMed/MEDLINE (1 Jan 2011)			
PubMed records	19,569,568	6,628,156	(1)
Identified clinical trial	457,378	171,025	(1)
Identified randomized controlled trial	293,963	156,496	(1)
Identified meta-analysis	25,723	21,146	(1)
Indexed journals	5,543	1,287	(2) 17 May 2011
EMBASE (17 May 2011)			
EMBASE records	~ 24 M		(3)
EMBASE journals	$\geq 7{,}500$		(3)
Cochrane library (17 May 2011)			
Clinical trials (CENTRAL)	645,086	286,418	(4) issue 2 of 4, Apr 2011
Cochrane reviews	4,621	4,621	(4) issue 5 of 12, May 2011
Other reviews	14,602	12,683	(4) issue 2 of 4, Apr 2011
All reviews (Cochrane + Other)	19,223	17,304	
Cochrane protocols	2,020	2,020	(4) issue 5 of 12, May 2011

sources

1. `http://www.nlm.nih.gov/bsd/licensee/baselinestats.html`
2. `http://www.nlm.nih.gov/tsd/serials/lji.html`
3. `http://embase.com/info/what-is-embase/coverage`
4. `http://thecochranelibrary.com/`

Table 2.1: Coverage of abstract databases PubMed and EMBASE, and the Cochrane Library.

funding [AMIA, 2010].

The FDA Modernization Act of 1997 made the US the first country to make trial registration a legal requirement. To implement this legislation, the ClinicalTrials.gov registry was launched in February 2000 [McCray and Ide, 2000]. The initial installment focused on providing a record of trials for enabling patient recruitment and investigator accountability. In 2004, both the World Health Organization (WHO) and the International Committee of Medical Journal Editors (ICMJE) released statements in support of the prospective registration of clinical trials. This policy has been widely adopted and now assures that the existence of most recent trials is known [Zarin et al., 2007]. Subsequently, various organizations, including the WHO, have called for a full disclosure of the trial protocol (including amendments) and results [Krleza-Jeric et al., 2005, Sim et al., 2006, Kaiser, 2008, Ghersi et al., 2008, Zarin and Tse, 2008, Chan, 2008, Sim et al., 2009]. In the US, recent legislation [FDA, 2007] has required protocol registration since December 2007, basic results reporting since September 2008, and Adverse Drug Events (ADEs) reporting since September 2009 [Wood, 2009]. Other governments with policies requiring prospective registration include the EU, India, Argentina, Brasil, Israel and South Africa [ICTRP, 2010].

To register a trial in ClinicalTrials.gov [Tse et al., 2009], researchers enter summary protocol information [ClinicalTrials.gov, 2009a] when their studies are initiated, and subsequently create the results section [ClinicalTrials.gov, 2009c] when the data collection for at least one primary outcome measure is complete. The ClinicalTrials.gov staff will review the results data after their submission. The data are reported in a structured tabular format and some meta-data, such as units of measurement or the use of standard vocabularies, can also be provided. Limited support for reporting statistical analyses is offered; these analyses are tied to specific results tables. Study protocols have long been available in XML format, and the retrieval of results in XML format was added in December 2011 [ClinicalTrials.gov, 2011].

Other countries have set up their own registries. Since 2004, the European Medicines Agency (EMA) has established clinical trial registration in accordance with the EU Directive 2001/20/EC through the EudraCT system. EudraCT was opened to the public only recently as the EU Clinical Trials Register, on 22 March 2011 [European Medicines Agency, 2011b], and the records are being released in a staggered fashion. Currently 17,102 [European Medicines Agency, 2012] of over 28,150 registered trials [European Medicines Agency, 2011a] are available. Another international registry is the Current Controlled Trials Ltd.'s ISRCTN registry, which has been in operation since 1998 [Current Controlled Trials Ltd., 2010]. It provides a semi-structured textual representation of the trial protocol, but no results. A number of countries have open national registries that generally record information in a way similar to ISRCTN. All of these registries are less sophisticated than ClinicalTrials.gov.

In order to unify trial registration world-wide, the WHO International Clinical Trials Registry Platform (ICTRP) was established following the Ministerial Summit on Health Research in November 2004. The goal of the ICTRP is to create "a network of international clinical trial registries to ensure a single point of access and the unambiguous identification of trials" [The Ministerial summit on health research, 2004]. This network of trial registries, the WHO Registry Network, was formally launched

Register	Studies	Indexed	Results
ClinicalTrials.gov (United States)	122,758	yes	yes (5,436)
European Union Clinical Trials Register	17,102	no	no
ISRCTN register (international)	10,465	yes	no
Japan Primary Registries Network	8,329	yes	no
Australian New Zealand Clinical Trials Registry	6,369	yes	no
The Netherlands National Trial Register	3,187	yes	no
Clinical Trials Registry India	2,499	yes	no
Iranian Registry of Clinical Trials	2,449	yes	no
Chinese Clinical Trial Register	2,004	yes	no
German Clinical Trials Register	831	yes	no
Cuban Public Registry of Clinical Trials	392	yes	no
South Korea Clinical Research Information Service	379	yes	no
Brazilian Clinical Trials Registry	131	no	no
Pan African Clinical Trial Registry	97	yes	no
Sri Lanka Clinical Trials Registry	71	yes	no

Table 2.2: ClinicalTrials.gov and the 14 WHO primary registries. The 'studies' column indicates the number of registered trials (per 19 March 2012), the 'indexed' column whether the register is indexed by the WHO search portal and the 'results' column whether the registry also enables results publication.

in 2007. In March 2012, 14 primary registries were listed on the ICTRP website (see Table 2.2). The ICTRP also provides a search portal that collects and indexes some basic information on trials from most of the primary registries and attempts to group trials that are registered in more than one registry in the search results. The search portal provides a textual description of the trial design as well as a link to the primary registry. Table 2.2 gives an overview of the WHO primary registries and ClinicalTrials.gov, with the number of included studies and whether or not they are indexed by the search portal. ClinicalTrials.gov is by far the largest registry, containing more than 7 times the number of trials available in the second largest registry (EU Clinical Trials Register), and 69% of all trials in the registries (not taking duplicates into account). Moreover, ClinicalTrials.gov is the only registry that registers results. In January 2010 it had published the results of 1,156 studies, which had increased to 5,436 studies by March 2012.

Although a decentralized system of federated registries (both national and multinational) seems cumbersome and may cause duplicate registration, there are important reasons why this method is to be preferred to a centralized approach [Grobler et al., 2008]: national registries, for example, are in the position to ensure complete registration in their region of influence and are perfectly aligned with the local political situation. As long as the different registries are sufficiently interoperable, an overarching organization such as the ICTRP can aggregate their databases.

The increased transparency enabled by trial registration offers new opportunities for evidence-based medicine and will likely lead to an increase in the number of systematic reviews that are undertaken [Honig, 2010]. However, the current registries contain only text-based or semi-structured information and lack a common coding system, for example for labeling interventions. The amount of protocol information registered is often insufficient to judge the validity of reported results and the problem of identifying all relevant studies has not yet been solved [Zarin et al., 2007]. In addition, the publicly available information may be incomplete or even "largely incomprehensible" [Wood, 2009]. The call for federated, open access, mandatory results databases continues [Kaiser, 2008, Ghersi et al., 2008, Zarin and Tse, 2008, Chan, 2008, Sim et al., 2009], and it is likely that the trend toward open and complete registration of results will continue.

2.3.3 Systematic review

The process of systematic review produces data and meta-data that is potentially useful for future reviewers and decision makers (see Section 2.1.2). The following assesses whether the existing information systems enable the dissemination of this information.

The Cochrane Collaboration provides several databases to support reviewers. Besides the CENTRAL database of clinical trials (see Section 2.3.1) the most relevant ones are the Cochrane Database of Systematic Reviews, in which Cochrane reviews are published [Starr and Chalmers, 2003], and the DARE database of other reviews. Table 2.1 provides statistics regarding the scope of the library. In contrast to the traditional journal publications of systematic reviews, which usually provide data in tables or figures, the Cochrane Reviews incorporate descriptions and results of the original studies. However, the published dataset has many data elements removed and the use of the data is restricted by license.

There are several software programs to aid in systematic review and meta-analysis, such as Comprehensive Meta-Analysis, MetaWin, MetaStat and MetaAnalyst. Moreover, many general-purpose statistical programs, such as SPSS Statistics, SAS Statistics, Stata, and R, have meta-analysis functionality. In general, dedicated meta-analysis software will provide easier data entry and management, while statistical programs will offer more powerful tools for analysis. The Cochrane Collaboration also provides the Review Manager software for performing systematic reviews (`http://ims.cochrane.org/revman`). Review Manager is unique in that it provides not only data analysis and management features, but includes functionality to write the full systematic review report. Indeed, the Review Manager file itself is submitted to the Cochrane library for review and eventual publication. Unfortunately, all of these systems lack sufficient meta-data to enable automated processing.

Finally, published systematic reviews are usually presented in a textual format without the underlying dataset, making it difficult to perform additional analyses that may be required for decision making. Thus, the inclusion of data on a new trial or new compound, as would be required for regulatory decision making, is also impossible. In conclusion, systematic review currently represents a missed opportunity to introduce additional structure to the available clinical trials information.

2.3.4 Regulatory assessment

After a pharmaceutical company develops a drug, it compiles the evidence collected from the discovery and development processes into a dossier that is submitted to the regulators who decide upon its market authorization. Submissions to the EMA and the national medicines boards in Europe are mainly text-based, containing aggregate-level results of clinical trials based on the applicant's statistical analysis. The FDA requests the submission of patient-level data, which is unlikely to become publicly available and as such is out of scope for this paper. The dossier forms the basis on which regulators assess the benefit-risk profile of a new drug. Although the clinical trial results are pivotal in this assessment, the decision is only indirectly based on them, as the decision making process is based on informal discussion between experts. While the decision may still be of high quality, this informal framework does not allow pharmaceutical companies or patients to discern how different pieces of the evidence weigh in on it.

The EMA publishes the European Public Assessment Reports (EPARs) of all centrally approved or refused medicines on its website. Note that this does not include all applications submitted to the EMA, as they can be withdrawn before a decision is reached [Eichler et al., 2010]. The EPAR contains information on all trials, but is completely textual without a semantic structure. Moreover, its information is directly derived from the submission by the applicant, while there is no standardization concerning what information should be provided, or in which format. Trials submitted to the EMA are required to be registered in EudraCT and will thus also be made known to the public through the EU Clinical Trials Register.

2.3.5 Standards and data models

A common standard of how clinical trials are performed and how their results are presented would make the process of systematic review more reliable and less laborious [Sim et al., 2000]. Some progress has been made by ClinicalTrials.gov, which currently registers and displays aggregate results. To do so, ClinicalTrials.gov has developed their own model, the Data Element Definitions (DED) [ClinicalTrials.gov, 2009a,c]. This model allows the reporting of aggregated outcome data and statistical analyses to some extent, but the information cannot be processed automatically because most fields are free text. This also means that finding all trials that are relevant to a specific patient condition is inaccurate, and thus requires overly broad search terms [Tu et al., 2011]. This lack of standardization and interoperability among registries and other databases should be addressed in the near future.

Several projects aim to enable general purpose re-use of clinical trials information, e.g. for cross-study analyses. We identified three such projects. The first is the Biomedical Research Integrated Domain Group (BRIDG) project, a collaboration between the Clinical Data Interchange Standards Consortium (CDISC), Health Level 7 (HL7), the National Cancer Institute (NCI) and the FDA that aims to bring together the common elements of their various standards into a complete data model for clinical trials [Biomedical Research Integrated Domain Group (BRIDG), 2010]. The BRIDG model is implementation-independent in the sense that it specifies the problem domain, not a specific solution. For example, unlike some other CDISC standards it does not specify the format in which to submit data to the FDA. BRIDG is subdivided between the protocol representation, study conduct, adverse event and regulatory perspectives. Unfortunately, a data analysis perspective is currently missing as there is no adequate standard for the modelling of statistical analyses. In short, the BRIDG model is accurate as regards the management of a single clinical trial, but not as regards cross-study analysis [Sim et al., 2010]. For example, the study population and eligibility criteria, outcomes and the measures used to assess outcomes do not have a sufficiently deep semantic structure.

To enable cross-study analyses and efficiently finding relevant trials, the Human Studies Database (HSDB) project aims to share fully machine understandable representations of study design information between institutions [Sim et al., 2010]. HSDB is developing the Ontology of Clinical Research (OCRe), which defines the concepts that should be accessible across the individual institutions' databases. At the time of writing, OCRe included a study design representation derived from BRIDG [Sim et al., 2010], a study design typology [Carini et al., 2009], the ERGO formal machine readable representation of eligibility criteria [Tu et al., 2011], and a model of study outcomes that separates the phenomena of interest from the variables that encode them [Sim et al., 2010]. While OCRe is a promising effort, its representation of study design is far from comprehensive, and it completely lacks a model for trial results.

Finally, the Ontology Based Extensible Conceptual Model (OBX) is another ontology for representing clinical trials [Kong et al., 2011, Scheuermann, 2010]. Its aim is to make the results of immunology studies available for data re-use and re-analysis. The OBX also incorporates study design representation ideas from BRIDG and the ClinicalTrials.gov DED [Kong et al., 2011]. While it appears successful in developing a broadly applicable data model for biomedical studies, and also allows the inclusion of trial results, it would appear that OBX suffers from similar shortcomings as regards the depth of modelling as BRIDG does.

All of the discussed models rely on an external coding system for their clinical content. Such coding systems, known as controlled terminologies of clinical terms, are an important first step in the application of information technology to medicine [Cimino, 1996]. There are many controlled terminologies for medicine, often developed for specific applications, but unfortunately there is as yet no standardization of which ones should be used, and there is no accurate mapping between them [Nadkarni and Darer, 2010, Vikstrom et al., 2007]. For example, in clinical research the Medical Dictionary for Regulatory Activities (MedDRA) is used to code ADEs, while the healthcare area prefers the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) dictionary. This hinders the interoperability of the various information systems being used.

2.4 Discussion

Having reviewed the information systems and standards dealing with the information from clinical trials, we will now summarize their deficiencies concerning the integration of clinical trials information from different resources, discuss how the status quo could be improved and identify directions for future research.

2.4.1 Identified deficiencies

Systems and standards oriented towards the management of single studies are relatively mature, but this is not the case for cross-study analysis. There are no known large, successful, and publicly available data warehouses, nor any standards that would enable cross-study analyses of aggregate level results. Considerable effort is required to harmonize the current clinical research standards. Important areas that require standardization are the representation of statistical analyses and aggregate results, as well as complex semantic structures such as patient eligibility criteria. None of the general purpose data models being developed are yet in widespread use, and from the perspective of capturing the designs and results of clinical trials in a reusable way, none of them are close to completion.

Although much effort is spent to publish the results of clinical trials, the current systems do not facilitate optimal use of the information. The journals and abstract databases that publish the trial results do not preserve the results' structure and thus require manual data extraction. Moreover, relevant articles are hard to identify and the retrieval of all available studies cannot be guaranteed. Public registries are meant to improve the efficiency and reliability of the identification of relevant studies, but the available data is not sufficiently structured to realize this. Moreover, the systems that currently deal with clinical trials results are not interlinked nor do they use interoperable standards. In short, there is not yet a comprehensive system of structured machine-understandable databases that contains descriptions of the design, execution, and summary-level results of individual trials. This situation hinders systematic review and makes cross-study analyses and data-mining prohibitively difficult. Thus, current infrastructure is focused on text-based reports of single studies, whereas efficient evidence-based medicine requires the automated integration of multiple clinical trials from different information resources.

Moreover, while systematic review collects and appraises the available evidence that is relevant to a certain question, the results are published in an unstructured format. This makes it hard to use the underlying data to inform evidence-based decisions, to verify the analyses, to update the review or to perform a combined analysis of several reviews for an umbrella review. The effort spent on literature screening and data extraction does not result in availability of this information for future reviewers, leading to duplication of effort.

Therefore, the current systems are unnecessarily burdensome and do not sufficiently facilitate reuse of the information. Figure 2.2 visualizes the current results dissemination process. Due to these shortcomings, decisions are not explicitly linked to the underlying evidence, leading to a lack of transparency, traceability, and reproducibility that is harmful for all stakeholders.

Importantly, the perceived lack of transparency in regulatory decision making may erode public trust in drug regulation and the pharmaceutical industry. More explicit quantitative decision models would enable a more transparent and repro-

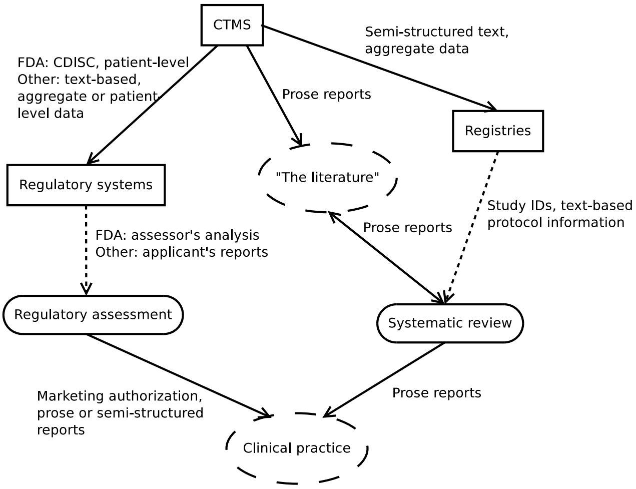

CTMS
FDA: CDISC, patient-level
Other: text-based,
aggregate or patient
level data
Prose reports
Semi-structured text,
aggregate data
Registries
Regulatory systems
"The literature"
Study IDs, text-based
protocol information
Prose reports
FDA: assessor's analysis
Other: applicant's reports
Regulatory assessment
Systematic review
Marketing authorization,
prose or semi-structured
reports
Prose reports
Clinical practice

Figure 2.2: In the current system of clinical trials results dissemination, data are collected in three separate systems (not including the organization that performs the trial).

ducible regulatory process, as well as a clearer communication of the requirements to the industry. However, for most real-world decisions it is currently too expensive to include all the evidence. This difficulty of accessing existing data is not only relevant to regulators and the industry, but also to reimbursement organizations, prescribing physicians and patients.

2.4.2 Proposed future situation

Now, we consider how these deficiencies might be addressed. Let us assume for a moment that a comprehensive machine-understandable standard were available for the design and aggregate level results of clinical trials. Then, it would be better for those submitting the data if both regulators and registries used this format, rather than a number of disparate formats. In addition, journal publications could easily be supplemented with data in this format.

Availability of a standard alone, however, is not enough to enable efficient access to the evidence. The data sets also need to be collected and made available in such a way that relevant clinical trials are easily identified. For this a collaborative (federated) system of databases should be established to capture all clinical trials data. Some of the stakeholders (e.g. regulators and registries) may require that data be submitted to a database that they control so that they can ensure the integrity of the data. This is fine as long as (1) the databases are interoperable and enable access to the information in the same format, (2) there is a single point of access through which the different databases can be identified and located, and (3) duplicate entries can be easily identified. It seems more likely for such a combination of databases to emerge from the current registries than a single centralized system.

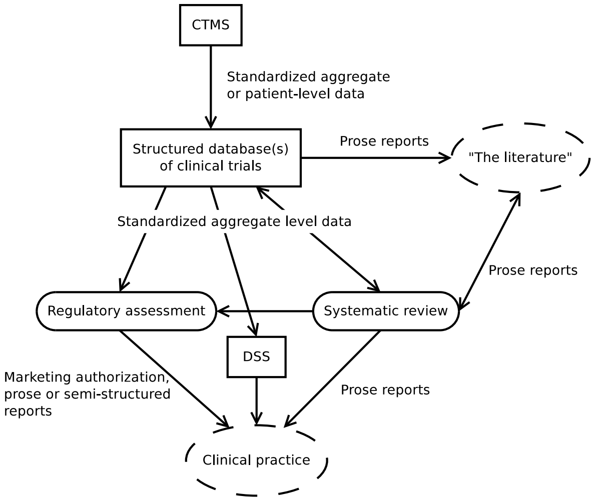

CTMS
Standardized aggregate or patient-level data
Structured database(s) of clinical trials
Prose reports
"The literature"
Standardized aggregate level data
Prose reports
Regulatory assessment
Systematic review
DSS
Marketing authorization, prose or semi-structured reports
Prose reports
Clinical practice

Figure 2.3: The alternative solution for clinical trials results dissemination proposed in this paper: harmonization of the different systems to create a unified platform for evidence-based decision making. Regulatory assessment has been merged into general policy decision making.

A comprehensive record of clinical trials in a machine-understandable format would make systematic review and consequently evidence-based decision making much more efficient. Decisions could then finally be explicitly linked to the underlying data (traceability). In addition, this could also enable a new generation of decision support systems for health care policy decision makers. The proposed future situation is visualized in Figure 2.3.

However, a suitable general-purpose data model is not likely to become available in the near future. Further, the usefulness of any data model should be demonstrated to the industry and other stakeholders before putting it into practice. We argue that the requirements for a general purpose data model for cross-study analysis and decision making are not sufficiently well known at the moment. Therefore, analysis tools and decision support systems that ensure data extraction done for the analysis can be shared with other researchers should be developed first, to illuminate these requirements.

2.4.3 Research directions

In order to attain the desired future system for aggregate-level clinical trial results dissemination, we identify concrete research directions for medical informatics, decision making and statistics researchers. Progress on each of these topics can be made in parallel.

- Computer-supported decision models for policy decision making based on clinical trials – to enable a direct and explicit link between the decision and the supporting evidence in drug regulation, reimbursement policy and guideline formulation

- Development of a platform to share structured systematic review data sets
- Discovery or creation of incentives for systematic reviewers to share the results of literature screening and data extraction
- Identification of the core data elements and modeling that are needed to increase the accuracy of literature searches
- Automated tagging and data extraction to facilitate transition to more structured data sets
- Development of search tools to integrate querying of abstract databases and registries
- Development of methods to identify duplicate trial publications and registrations
- Development of a comprehensive data model for clinical trials and their aggregate level results

2.4.4 Limitations

As with any review paper, there is a risk that relevant publications have not been identified, either because the search terms were not broad enough, or because relevant studies were not identified as such based on their title and abstract. We acknowledge that the broad scope of this particular review increases that risk.

The nature of the collected information necessitates a qualitative synthesis, and the identified deficiencies are at least partially subjective. The future that we propose is based on the premise that a standard for aggregate clinical trial data will become available. Unfortunately, it is unclear how and when this could be realized. Finally, the list of proposed research directions is sure to be incomplete, and we hope the present paper will ignite discussions on this topic.

2.4.5 Conclusions

We reviewed the existing systems and standards dealing with aggregate level results of clinical trials. The transfer of evidence to scientific journals, public registries, and regulators is a largely ad hoc and text-based affair. In part, this is because there are currently no data standards that enable cross-study analyses. We have argued that the lack of a standardized, federated system for results dissemination leads to gaps in the transfer and availability of evidence to the relevant decision makers. As long as such a system does not exist, systematic review will remain an incredibly inefficient ad-hoc process, and evidence-based decision making will remain unnecessarily difficult. We believe that these difficulties lead to a lack of transparency in health care policy decision making, which threatens public trust in the decision makers.

In the future, results registries and regulatory systems should be harmonized and federated to create a system of databases that forms the core of a more automated and efficient process of systematic review and evidence-based decision making. In addition, systematic reviews are currently a missed opportunity to introduce additional structure to the domain of clinical trials information, which should be addressed by more complete dissemination of their results. Although this vision is still far from realized, current trends seem to support this direction. Future work should not only focus on developing the 'ideal' data model for all of clinical research (justly called a monumental task) but start by creating useful tools to decision makers and systematic reviewers. Availability of such tools will lead to increased demand for an accessible evidence base and to a better understanding of its requirements.

CHAPTER 3

Automating network meta-analysis

G. van Valkenhoef, G. Lu, B. de Brock, H. Hillege, A. E. Ades, and N. J. Welton. Automating network meta-analysis. *Research Synthesis Methods*, 2012a. doi: 10.1002/jrsm.1054. (in press)

Abstract

Mixed Treatment Comparison (MTC) (also called network meta-analysis) is an extension of traditional meta-analysis to allow the simultaneous pooling of data from clinical trials comparing more than two treatment options. Typically, MTCs are performed using general purpose Markov Chain Monte Carlo (MCMC) software such as WinBUGS, requiring a model and data to be specified using a specific syntax. It would be preferable if, for the most common cases, both could be derived from a well-structured data file that can be easily checked for errors. Automation is particularly valuable for simulation studies in which the large number of MTCs that have to be estimated may preclude manual model specification and analysis. Moreover, automated model generation raises issues that provide additional insight into the nature of MTC. We present a method for the automated generation of Bayesian homogeneous variance random effects consistency models, including the choice of basic parameters and trial baselines, priors, and starting values for the Markov chain(s). We validate our method against the results of five published MTCs. The method is implemented in freely available open source software. This means that performing an MTC no longer requires manually writing a statistical model. This reduces time and effort, and facilitates error checking of the data set.

3.1 Introduction

Meta-analysis refers to statistical methods that combine evidence from multiple clinical trials in order to derive a pooled estimate of the relative effect of treatments. Traditional meta-analysis [Hedges and Vevea, 1998, Normand, 1999] has focused on pair-wise comparisons of treatments based upon summary measures of relative effect as reported in the original trials. Mixed Treatment Comparison (MTC) is a recently developed method that allows the simultaneous comparison of more than two treatments [Lu and Ades, 2004, Salanti et al., 2008a]. MTCs allow the use of both direct and indirect evidence for comparisons. In this paper, we focus on the Bayesian approach to MTC, which also allows the straightforward calculation of the rank-probabilities of a set of alternative treatments.

Specifying a Bayesian MTC model involves writing a Directed Acyclic Graph (DAG) model for general purpose Markov Chain Monte Carlo (MCMC) software such as WinBUGS [Lunn et al., 2000] or JAGS [Plummer, 2009]. In addition, prior distributions have to be specified for a number of the parameters and the data have to be supplied in a specific format. Together, the DAG, priors and data form a Bayesian Hierarchical Model (BHM). Moreover, due to the nature of MCMC estimation, over-dispersed starting values have to be chosen for a number of independent chains so that convergence can be assessed [Gelman and Rubin, 1992, Brooks and Gelman, 1998]. Currently, there is no software that automatically generates MTC models, although there are some tools to aid in the process. For example, the UK National Institute for Health and Clinical Excellence (NICE) Decision Support Unit offers a technical support document [Dias et al., 2011] that includes example WinBUGS code that applies to a broad range of data sets. However, most of the decisions mentioned above still have to be made, and having a fixed default value in the example code may lead to misleading results if the user is not aware of the need to modify this value for the situation at hand. Moreover, the format in which data are presented to BUGS can be hard to read (especially for large data sets) and does not facilitate error checking. For example, treatments are referred to by number rather than a name and data can be presented either in tabular format, in which it is difficult to keep track of which column corresponds to which variable, or in list format, in which one has to check that indices match between several lists.

The effort that is required to manually specify MTC models is for the most part unnecessary, and automated model generation would enable the analyst to focus on more interesting aspects of the problem. The input data can then be presented in a more structured format that facilitates error checking. Moreover, in some cases, the number of MTCs to be carried out may necessitate such an approach, for example in simulation based studies where each iteration requires estimating an MTC, or in decision support applications where many criteria need to be considered. To address this, we present how Bayesian MTC models can be generated automatically, an endeavor that also provides insight into the nature of MTC. Specifically, we show how homogeneous variance random effects consistency models for both dichotomous and continuous outcomes can be generated. To fully automate model generation, we show how to specify the DAG, priors that limit bias, and starting values that are unlikely to lead to misdiagnosing convergence. Generating the DAG for inconsistency models is discussed in van Valkenhoef et al. [2012d].

Note that what we present in this paper is the automated generation of MTC models that can be run in WinBUGS or JAGS. The analyst is still expected to choose the run length of the MCMC simulation, check whether the posterior has converged (and increase the run length if needed), assess the model fit, and interpret the results. For application in simulation studies our methods can easily be combined with an automated convergence checking routine that extends the run length as needed. However, there is always the risk of erroneously concluding that convergence has been reached so, if feasible, convergence checking is best done by visual inspection of the relevant plots [Brooks and Gelman, 1998].

3.2 Background

In the Bayesian framework, an MTC is implemented as a BHM and estimated using MCMC simulation [Lu and Ades, 2006, Salanti et al., 2008a]. This section explains the structure of the BHM and the considerations that have to be made when estimating it through MCMC simulation. MTC models are an extension of a Bayesian formulation of pair-wise meta-analysis. For clarity, we initially limit the discussion to random-effects homogeneous variance consistency models for dichotomous variables [Lu and Ades, 2004, 2006, Salanti et al., 2008a]. The extension to continuous variables [Salanti et al., 2008a] is discussed thereafter. We will assume that the outcome data are reported per arm, rather than as treatment contrasts against a common baseline.

3.2.1 Consistency models for dichotomous variables

For a dichotomous variable (the occurrence or non-occurrence of an event), for every clinical trial i, for each included treatment x, we have the sample size $n_{i,x}$, and the number of events $r_{i,x}$ that occurred. The events are assumed to arise from a binomial process with success probability $p_{i,x}$:

$$r_{i,x} \sim \mathrm{Bin}(p_{i,x}, n_{i,x}) \ . \tag{3.1}$$

The success probability can be transformed to the log odds scale through the

$$\mathrm{logit}(p) = \log \frac{p}{1-p}$$

function. On the log odds scale, relative effects are assumed additive and normally distributed, drastically simplifying the model. The inverse transformation, logit^{-1}, is used to define $p_{i,x}$ in terms of log odds scale random variables:

$$\mathrm{logit}(p_{i,x}) = \mu_i + \delta_{i,b(i),x} \Leftrightarrow p_{i,x} = \mathrm{logit}^{-1}(\mu_i + \delta_{i,b(i),x}) \ , \tag{3.2}$$

where $b(i)$ is the baseline arm chosen for i, μ_i is the effect of $b(i)$ in i and $\delta_{i,b(i),x}$ is the *random effect* of x relative to $b(i)$ in i. If $b(i) = x$, we set $\delta_{i,b(i),x} = 0$. Otherwise,

$$\delta_{i,b(i),x} \sim \mathcal{N}(d_{b(i),x}, \sigma^2) \ , \tag{3.3}$$

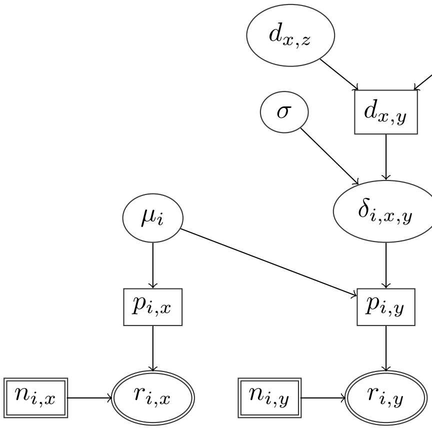


Figure 3.1: The basic structure of an MTC model is a DAG, here shown graphically for a subset of a full MTC model. Elliptical nodes represent density functions, whereas rectangular nodes represent deterministic functions. Nodes with a double border have data associated with them. Here, one study (study i) that includes two treatments (x and y) in a model with three treatments (x, y and z) is shown. The parameter $d_{x,y}$ is a functional parameter, defined as a deterministic function of the basic parameters $d_{x,z}$ and $d_{y,z}$.

where $d_{b(i),x}$ is the *relative effect* of x compared to $b(i)$, the quantity of interest, and σ^2 is the *random effects variance*, a measure of the heterogeneity between trials. Because we assume σ to be the same for all $d_{x,y}$, this is a *homogeneous variance* model. In such a model, the covariances between comparisons in multi-arm trials work out to $\sigma^2/2$ [Salanti et al., 2008a]:

$$\begin{pmatrix} \delta_{i,b(i),x} \\ \vdots \\ \delta_{i,b(i),z} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} d_{b(i),x} \\ \vdots \\ d_{b(i),z} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2/2 & \cdots \\ \sigma^2/2 & \ddots & \sigma^2/2 \\ \cdots & \sigma^2/2 & \sigma^2 \end{pmatrix}\right) \quad . \tag{3.4}$$

Finally, we assume that comparisons are *consistent* (or, more fundamentally, we assume that trials are exchangeable conditional on the true value of the hyper parameters $d_{\cdot,\cdot}$ and σ, and consistency is implied [Lu and Ades, 2009]). That is, if we compare x and y indirectly through z, the result will be consistent with the direct comparison:

$$d_{x,y} = d_{x,z} - d_{y,z} \quad . \tag{3.5}$$

The right hand side parameters are called the *basic parameters*, for which we estimate probability distributions. Any other relative effect can be calculated using the consistency assumption. Hence $d_{x,y}$, a *functional parameter*, is completely defined in terms of the basic parameters on the right hand side. There are ways to test whether consistency holds [Lumley, 2002, Lu and Ades, 2006, Dias et al., 2010, Lu et al., 2011], but these are beyond the scope of this paper.

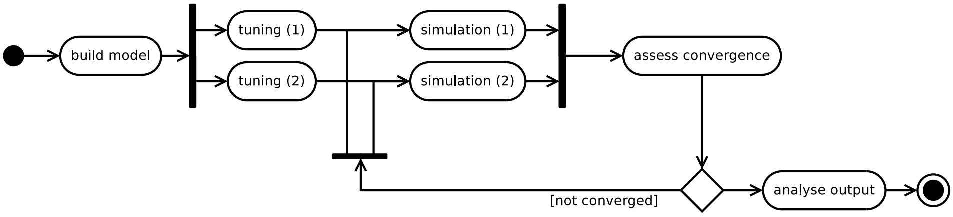


Figure 3.2: The process of running an MTC with two parallel chains. Notation: UML-2 activity diagram [Fowler, 2003].

The model as discussed above forms a DAG, as shown in Figure 3.1, with the $r_{i,x}$ nodes at the very bottom (these are *sinks*) and $n_{i,x}$, $d_{x,y}$, μ_i and σ at the top (*sources*). Now, because the model is Bayesian, we have to specify prior distributions for the source nodes (except $n_{i,x}$, which have a fixed value), that reflect our beliefs before seeing the data. In addition, to perform MCMC estimation of the BHM, we have to specify a starting point for each of the stochastic nodes. Stochastic nodes are nodes that have probability distributions such as (3.3) and unlike (3.2). The MCMC simulation then proceeds by making random jumps in the parameter space using a procedure that is *guaranteed* to converge (under certain regularity assumptions) on the posterior distribution in the limit of infinitely many iterations. Most MCMC software includes a *tuning* or *adaptive* phase [Spiegelhalter et al., 2003, Plummer, 2009, Graves, 2008], during which the sampling algorithms are optimized to ensure efficient exploration of the parameter space. While tuning, the MCMC simulation will not converge, so after this initial phase tuning is turned off to allow the model to converge. Then, we run the model only as long as is needed to accurately estimate the quantities of interest, discarding an initial sub-sequence of the samples called the *burn-in* period. In that case, we say approximate convergence has been reached.

To assess approximate convergence using the Brooks-Gelman-Rubin diagnostic [Brooks and Gelman, 1998], the model is run N_C times in parallel, with distinct starting points, each for $2N_I$ iterations. The starting points have to be over-dispersed relative to the target distribution for the assessment to be valid. The results of the last N_I iterations of all N_C runs (called *chains*), are then analyzed and the within-chain and between-chain variance are compared to estimate the Potential Scale Reduction Factor (PSRF). A PSRF close to 1 indicates approximate convergence has been reached. The first N_I iterations are called burn-in iterations, and are discarded. To conclude that approximate convergence has been reached in MTC, all of the parameters of interest (source nodes) should have a PSRF below a certain threshold α, and visual inspection of plots of the PSRF and time series should not contradict this conclusion. If this is not the case, the simulation phase should be extended. Although it has been suggested that values below 1.2 are acceptable [Brooks and Gelman, 1998], we suggest to set $1 < \alpha \leq 1.05$ to be conservative. For the type of model discussed here, that condition is usually achieved reasonably quickly (i.e. within 50,000 iterations), and so we can afford to be conservative. The full process of estimating an MTC is shown in Figure 3.2. Note that there are valid alternative work flows, that will not be described here.

3.2.2 Continuous variables

When considering continuous outcomes, the Equations 3.1 and 3.2 are replaced by a normal likelihood [Salanti et al., 2008a]. Now, for a trial i and treatment x we take the sample mean $m_{i,x}$, sample standard deviation $s_{i,x}$, and sample size $n_{i,x}$:

$$m_{i,x} \sim \mathcal{N}(\mu_i + \delta_{i,b(i),x}, s^2_{i,x}/n_{i,x}) \ . \tag{3.6}$$

Where, again, we model the observed effect in each arm in terms of a baseline effect μ_i for the trial and a random effect $\delta_{i,b(i),x}$. This model requires that all studies have measured the outcome on comparable scales, or that they have been transformed onto a common scale, for example the standardized mean difference.

3.2.3 Maximum likelihood estimators in single trials

Now, we briefly review the maximum likelihood estimators for both types of data, as they will be needed later on. First, define estimators $\hat{\theta}$ for the log odds and $\hat{v}^2$ for its variance:

$$\hat{\theta}_{i,x} = \text{logit}(\frac{r'_{i,x}}{n'_{i,x}}) \ ; \ \hat{v}^2_{i,x} = \frac{1}{r'_{i,x}} + \frac{1}{n'_{i,x} - r'_{i,x}} \ ; \tag{3.7}$$

$$r'_{i,x} = r_{i,x} + \frac{1}{2} \ ; \ n'_{i,x} = n_{i,x} + 1 \ ,$$

where $r'_{i,x}$ and $n'_{i,x}$ are corrected to ensure ratios are always defined. Usually the correction is only applied when the uncorrected ratio is undefined but, since we don't use these estimators for inference, the difference is not relevant. For continuous outcomes the estimators are:

$$\hat{\theta}_{i,x} = m_{i,x} \ ; \ \hat{v}^2_{i,x} = \frac{s^2_{i,x}}{n_{i,x}} \ . \tag{3.8}$$

For the relative effect of y when compared to x the estimators (for dichotomous or continuous data) are [DerSimonian and Laird, 1986]:

$$\hat{\delta}_{i,x,y} = \hat{\theta}_{i,y} - \hat{\theta}_{i,x} \ ; \ \hat{s}^2_{i,x,y} = \hat{v}^2_{i,y} + \hat{v}^2_{i,x} \ . \tag{3.9}$$

3.3 Methods

To automatically generate MTC models, the BHM has to be specified, consisting of the DAG, prior distributions and data. Moreover, over-dispersed starting values have to be chosen for N_C independent chains. In the following, we show how to specify the DAG for consistency models and how priors that limit bias, and starting values that are unlikely to lead to misdiagnosing convergence can be chosen to fully automate MTC model generation. Finally, we provide a worked example of the methods.

3.3.1 Generating the model structure

Generating the DAG for inconsistency models has previously been discussed [van Valkenhoef et al., 2012d], and involves finding the formulation that maximizes the number of potential inconsistencies that can be estimated, the Inconsistency Degrees of Freedom (ICDF). However, the problem is considerably simpler for consistency models, and rather than being forced to search for a DAG that maximizes the ICDF, we can try to specify one that has an easily understandable structure. Arguably, the most easily understood structure is the one where we parameterize each treatment effect relative to a common baseline (e.g. placebo). In such a model, the basic parameters form a star shaped graph, and all functional parameters are a linear combination of just two basic ones. The algorithm proposed for inconsistency models [van Valkenhoef et al., 2012d] does exactly the opposite: due to the specific search algorithm it uses to find a suitable parametrization, it is likely to end up with a line graph, or something close to it.

It has been shown that for inconsistency models, an incorrect choice of baseline treatments for the individual studies may result in some of the parameters being under-constrained [Lu and Ades, 2006, van Valkenhoef et al., 2012d]. In Appendix 3.7, we show that this problem cannot occur for consistency models. Thus, regardless of how the basic parameters and study baselines are chosen, the model is always well defined. Therefore, we are completely free to choose a parametrization that is easily understandable. In fact, in a consistency model it is not even necessary for the basic parameters to have been measured directly, and thus we could choose a star shaped graph for the basic parameters. However, the method we propose below to choose starting values depends on the basic parameters being measured directly as a simplifying assumption, so we restrict the basic parameters to the directly measured comparisons. Note that after the samples for the basic parameters have been obtained, it is straightforward to express the result relative to any of the included treatments by post-processing the samples, so this restriction does not hamper inference.

Given that the basic parameters are restricted to directly measured comparisons, and that they must form a spanning tree of the evidence graph, we would like for the linear equations that define the functional parameters to involve as few basic parameters as possible. The diameter of a spanning tree is the length of the longest of the paths between its vertices. Thus, the minimum diameter spanning tree provides the parametrization that minimizes the length of the longest equation. The minimum diameter spanning tree can be found efficiently [Hassin and Tamir, 1995]. This provides an automated way of specifying the basic parameters even if the evidence structure does not have a common comparator. For the study baselines, we just choose (in each study) the treatment that has the most connections in the spanning tree of the basic parameters, again with the aim of minimizing the length of equations. If there is more than one such treatment we may choose an arbitrary one, e.g. alphabetical ordering of the treatments can be used to break ties.

3.3.2 Choosing priors

Prior distributions have to be specified for all the basic relative effect parameters $d_{x,y}$, the baseline effects μ_i and the random effects variance σ^2. For the variance parameters, either a uniform or an inverse-Gamma prior may be chosen [Lu and Ades, 2004]. However, the uniform prior is more commonly used [e.g. Lu and Ades, 2006]:

$$\sigma \sim \mathcal{U}(0, u) \ ,$$

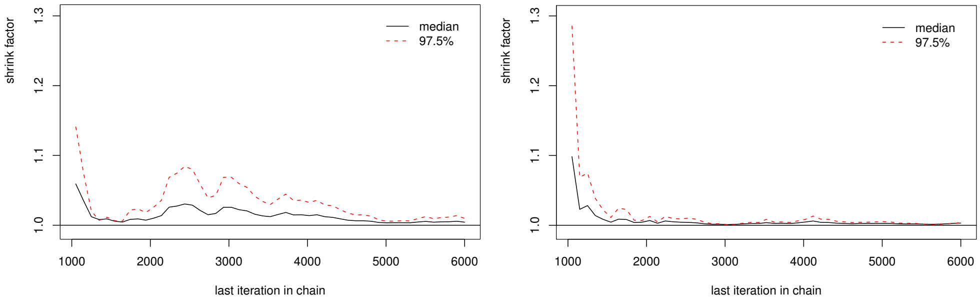
where u has to be appropriately chosen so that it is larger than the standard deviation that may exist. But how to do this has not been discussed in the literature. Normally, the analyst should provide this value, but doing so is often difficult: what would constitute an unreasonably large deviation in the log odds ratio for treatment response? No single value of u will be appropriate for every situation, and thus a simple heuristic that over-estimates the measurement scales based on minimal information is a reasonable compromise. It should be noted that in most cases the posterior is not sensitive to the variance prior, as long as it is sufficiently wide [Lu and Ades, 2004]. A simple technique is to calculate the maximum likelihood estimators $\hat{\delta}_{i,x,y}$ (Equation 3.9) for all possible combinations of i, x, and y and find the maximal one, δ^+. Setting $u = \delta^+$ guarantees $u > \sigma$ and this choice of u does not depend on the chosen DAG, so that alternative DAGs have the same prior.

For the relative effect parameters and the baseline effects, independent identical normal priors are usually specified [Lu and Ades, 2004, 2006, Salanti et al., 2008a], i.e.:

$$d_{x,y}, \mu_i \sim \mathcal{N}(\nu, \eta^2) \ ,$$

where $\nu = 0$ and $\eta^2 = 1000$ is often used for dichotomous outcomes [Lu and Ades, 2004, 2006, Salanti et al., 2008a]. The large variance is chosen so that the data will dominate the prior information in the final result (i.e. we specify a *vague* prior). Again, what constitutes a large variance depends on the scale of the data, and therefore we set $\eta^2 = (ku)^2$. Choosing 0 as the prior mean for $d_{x,y}$ is done so any bias due to the prior is conservative: the bias is in the direction of no effect. For μ_i a prior mean of 0 is more arbitrary. For dichotomous data, it introduces a slight bias towards $p_{i,b(i)} = 0.5$, which means values near 0 and 1 will be (slightly) shrunk towards 0.5. For continuous data, the bias is towards 0. Thus, it is important that η^2 is sufficiently large to minimize this bias.

To choose a value for the scaling factor k, we looked at a number of studies in the literature [Lu and Ades, 2004, Welton et al., 2009, Dias et al., 2010]. In those studies, η^2 was either 10^3 or 10^4, and the upper bound for the variance prior ranged from 2 to 50. If the η^2 were generated using a scaling factor k, the value of k would range from 3.16 to 15.8. Somewhat arbitrarily, we choose $k = 15$ as the default, which is in the upper range of values seen in published analyses, since the goal is to ensure the variance is large enough, regardless of the measurement scale. This choice is validated in Section 3.5 by comparing the posteriors we obtain against those reported in the original analyses.



(a) Undesirable (b) Desirable

Figure 3.3: Undesirable and desirable convergence behaviour illustrated. In (a), after 1,500 iterations we might incorrectly conclude the simulation has sufficiently converged. The more gradual but consistent decline of (b) is preferable.

3.3.3 Choosing starting values

Starting values have to be over-dispersed relative to the target density in order to ensure complete exploration of the parameter space and assess convergence. They have to be chosen for all stochastic nodes that are not bound to data (i.e., $\delta_{i,x,y}$, μ_i, $d_{x,y}$ and σ). Note that for the $d_{x,y}$ this includes only the basic parameters and not the functional ones. Starting values can be generated by creating an over-dispersed approximate distribution from which they are drawn [Gelman and Rubin, 1992]. Our strategy is to use analytic techniques for pair-wise comparisons [DerSimonian and Laird, 1986] to get a maximum likelihood estimate ξ and standard error τ for each node, then sample that node's starting values from $\mathcal{N}(\xi, c\tau^2)$ with $c \gg 1$. To estimate μ_i, take $\xi = \hat{\theta}_{i,b(i)}$ and $\tau^2 = \hat{v}^2_{i,b(i)}$ (Equations 3.7, 3.8). For $\delta_{i,x,y}$, $\xi = \hat{\theta}_{i,y} - \hat{\theta}_{i,x}$ and $\tau^2 = \hat{v}^2_{i,y} + \hat{v}^2_{i,x}$ (Equation 3.9). For the $d_{x,y}$, we do a pair-wise random effects pooling [DerSimonian and Laird, 1986] of all trials that include x and y to find ξ and τ^2. Note that in our generated models, the basic parameters are always directly measured (Section 3.3.1), and that we use a correction to account for zero cells (Section 3.2.3). When only one trial measures the basic parameter in question, the pooling method will estimate the between-trials error to be equal to the standard error of the mean of that one trial. Starting values for σ can be obtained by sampling from its prior distribution.

It is important that the starting values are sufficiently over-dispersed so that the PSRF is not close to 1 before approximate convergence has been reached, in which case it will increase later on when an unexplored part of the parameter space is found (see Figure 3.3). We also don't want c to be too large, as this would slow down the exploration of the parameter space significantly. However, we need not be overly concerned with this issue, as the models considered here tend to converge quickly, even if the starting values are not ideal. The value c is configurable in our implementation, but we set $c = 2.5$ by default.

Study	Treatment	Mean	Std. dev.	sample size
1	A	-1.22	3.70	54
	C	-1.53	4.28	95
2	A	-0.70	3.70	172
	B	-2.40	3.40	173
3	A	-0.30	4.40	76
	B	-2.60	4.30	71
	D	-1.20	4.30	81
4	C	-0.24	3.00	128
	D	-0.59	3.00	72
5	C	-0.73	3.00	80
	D	-0.18	3.00	46
6	D	-2.20	2.31	137
	E	-2.50	2.18	131
7	D	-1.80	2.48	154
	E	-2.10	2.99	143

Table 3.1: Mean off-time reduction data from 7 trials studying 5 treatments form Parkinson's disease [Dias et al., 2011].

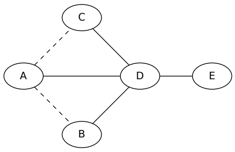


Figure 3.4: Evidence network for the Parkinson's disease trials, with basic parameters shown as solid lines and the functional parameters as dashed lines

3.3.4 Worked example

We now illustrate the methods with a worked example of 5 treatments for Parkinson's disease [Dias et al., 2011]. There is data on mean off-time reduction from 7 trials (see Table 3.1). The evidence network is shown in Figure 3.4. First, to determine the basic parameters, we find the minimum diameter spanning tree (the solid lines in Figure 3.4). In this case, the basic parameters are $d_{D,A}$, $d_{D,B}$, $d_{D,C}$, and $d_{D,E}$. This is the only paremeterization in which all functional parameters can be expressed in terms of at most two basic ones (i.e. treatment D is a common comparator). Then, we choose the study baselines by identifying the treatment with the most connections in the spanning tree (basic parameters). In this case, treatment D is connected to all four other treatments, so any trial that includes D will have D as the baseline treatment. For example, study 3 would be parameterized as:

$$\begin{pmatrix} \delta_{3,D,A} \\ \delta_{3,D,B} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} d_{D,A} \\ d_{D,B} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{pmatrix} \right) .$$

For the remaining trials, we choose the alphabetically first treatment as the baseline. For example, trial 1 would be parameterized as:

$$\delta_{1,A,C} \sim \mathcal{N}(-d_{D,A} + d_{D,C}, \sigma^2) \ .$$

Then, to choose the prior distributions, we calculate a maximum likelihood estimate for *all* relative effects in all studies, including $\hat{\delta}_{1,A,C} = -0.31$, $\hat{\delta}_{1,C,A} = 0.31$, etc. For three-arm studies, there are six such relative effects. In this case, the maximum is 2.3, so we set $u = 2.3$ and $\eta^2 = (15 \cdot 2.3)^2 = 1.2 \cdot 10^3$, leading to the priors:

$$\sigma \sim U(0, 2.3) \ ; \ d_{x,y}, \mu_i \sim \mathcal{N}(0, 1.2 \cdot 10^3) \ .$$

Finally, we use maximum likelihood estimators to derive sampling distributions from which to draw starting values. For the baseline effect μ_1 in study 1, where A is the baseline, we draw starting values from

$$\mathcal{N}(\hat{\theta}_{1,A}, c \cdot \hat{v}^2_{1,A}) = \mathcal{N}(-1.22, 2.5 \cdot 3.7^2/54) \ .$$

For the relative effect $\delta_{1,A,C}$ we draw starting values from

$$\mathcal{N}(\hat{\theta}_{1,C} - \hat{\theta}_{1,A}, c \cdot (\hat{v}^2_{1,A} + \hat{v}^2_{1,C})) = \mathcal{N}(-0.31, 2.5 \cdot (0.25 + 0.19)) \ .$$

Finally, for the basic parameters we perform random effects pooling of the relevant studies. For example, to determine the distribution of starting values for $d_{D,C}$, we pool studies 4 and 5. Then, the distribution is $\mathcal{N}(-0.04, 0.45)$.

3.4 Implementation

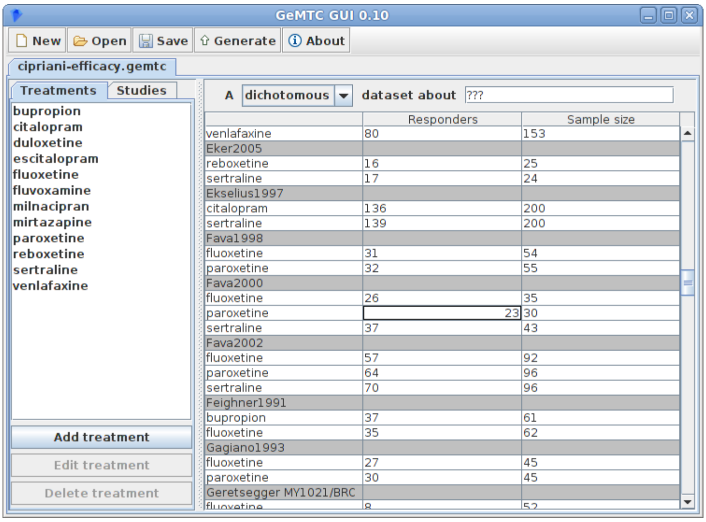
The methods presented in this paper are implemented in GeMTC (`http://drugis.org/gemtc`), enabling generation of JAGS and BUGS models. The data are stored in an XML format, an example of which is shown in Figure 3.5.

A simple graphical user interface is provided to facilitate data entry and manipulation of data files, as well as model generation. The main screen (see Figure 3.6) enables the user to create, load, and save datasets in the GeMTC XML format. Each dataset is opened in a tab named after the dataset ('`cipriani-efficacy.gemtc`' in Figure 3.6). In the left panel there are two tabs where the user can add, edit, and remove treatments and studies. A treatment consists of a short identifier and an optional description. A study consists of an identifier and at least two included treatments. At the top of the right panel, the user chooses the type of data (dichotomous or continuous), and a description for the dataset. The rest of the right panel consists of the data table, where data for each arm can be input in the appropriate format. For dichotomous measurements, this is the number of events and the sample size; for continuous measurements, the mean, standard deviation, and sample size. In the current version of GeMTC (0.12.1), data have to be input manually, as copy-paste support and import from various formats are not yet available.

After the dataset is complete, the statistical model can be generated by clicking 'Generate' in the tool bar. If the studies do not form a connected evidence network,

```
<network description="Response to treatment">
   <treatments>
      <treatment id="Pla">Placebo</treatment>
      <treatment id="Flu">Fluoxetine</treatment>
   </treatments>
   <studies>
      <study id="Fictional et al, 2009">
         <measurement treatment="Pla" responders="38" sample="97" />
         <measurement treatment="Flu" responders="42" sample="95" />
      </study>
   </studies>
</network>
```

Figure 3.5: The MTC input data format (XML) illustrated with an artificial data set containing one (fictional) study "Fictional et al, 2009" comparing two treatments (Placebo and Fluoxetine) on a dichotomous outcome (response to treatment)

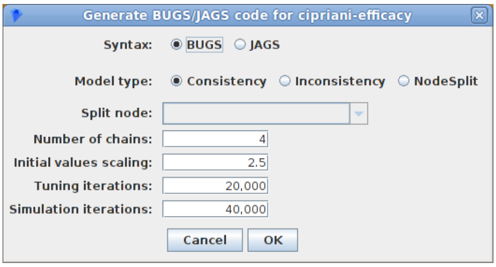

GeMTC GUI 0.10
New
Open
Save
Generate
About
cipriani-efficacy.gemtc
Treatments
Studies
Add treatment
Edit treatment
Delete treatment

bupropion
citalopram
duloxetine
escitalopram
fluoxetine
fluvoxamine
milnacipran
mirtazapine
paroxetine
reboxetine
sertraline
venlafaxine

A dichotomous dataset about ???

	Responders	Sample size
venlafaxine	80	153
Eker2005		
reboxetine	16	25
sertraline	17	24
Ekselius1997		
citalopram	136	200
sertraline	139	200
Fava1998		
fluoxetine	31	54
paroxetine	32	55
Fava2000		
fluoxetine	26	35
paroxetine	23	30
sertraline	37	43
Fava2002		
fluoxetine	57	92
paroxetine	64	96
sertraline	70	96
Feighner1991		
bupropion	37	61
fluoxetine	35	62
Gagiano1993		
fluoxetine	27	45
paroxetine	30	45
Geretsegger MY1021/BRC		
fluoxetine	8	52

Figure 3.6: Creating and editing an MTC data file using the GeMTC GUI



Syntax: ● BUGS ○ JAGS

Model type: ● Consistency ○ Inconsistency ○ NodeSplit

Split node:

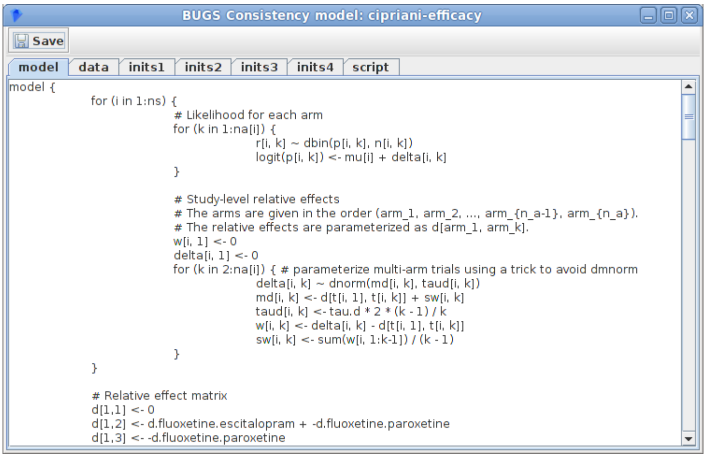
Number of chains: 4

Initial values scaling: 2.5

Tuning iterations: 20,000

Simulation iterations: 40,000

Figure 3.7: Configuring the model generation using the GeMTC GUI



```
model {
        for (i in 1:ns) {
                # Likelihood for each arm
                for (k in 1:na[i]) {
                        r[i, k] ~ dbin(p[i, k], n[i, k])
                        logit(p[i, k]) <- mu[i] + delta[i, k]
                }

                # Study-level relative effects
                # The arms are given in the order (arm_1, arm_2, ..., arm_{n_a-1}, arm_{n_a}).
                # The relative effects are parameterized as d[arm_1, arm_k].
                w[i, 1] <- 0
                delta[i, 1] <- 0
                for (k in 2:na[i]) { # parameterize multi-arm trials using a trick to avoid dmnorm
                        delta[i, k] ~ dnorm(md[i, k], taud[i, k])
                        md[i, k] <- d[t[i, 1], t[i, k]] + sw[i, k]
                        taud[i, k] <- tau.d * 2 * (k - 1) / k
                        w[i, k] <- delta[i, k] - d[t[i, 1], t[i, k]]
                        sw[i, k] <- sum(w[i, 1:k-1]) / (k - 1)
                }
        }

        # Relative effect matrix
        d[1,1] <- 0
        d[1,2] <- d.fluoxetine.escitalopram + -d.fluoxetine.paroxetine
        d[1,3] <- -d.fluoxetine.paroxetine
```

Figure 3.8: Generating BUGS model files using the GeMTC GUI

the user is advised of this and asked to correct the dataset. Otherwise, the user is presented with a settings dialog (Figure 3.7). Here, various parameters can be configured, such as the syntax type (BUGS or JAGS), the model type (only the consistency model is described in this paper), the scaling factor c for the initial values, and the number of chains, tuning iterations and simulation iterations for the MCMC simulation. Unfortunately, priors are currently not configurable from this screen, and the user will need to modify the generated BUGS or JAGS code to change the priors.

After configuring the parameters for model generation, the user can click 'OK' to generate the model. This opens a new window, which shows the various components of the model in separate tabs (see Figure 3.8). JAGS users can use the 'save' button to save these components to a series of files, after which the '`*.script`' file can be used to run the model in JAGS (some BUGS versions also support this). BUGS users can copy-paste the contents of each tab to BUGS and run the model from the BUGS user interface.

Generating the BUGS or JAGS models entails taking the data, applying the methods discussed in the previous section to determine the basic parameters and study baselines, choose the priors and generate starting values, and then generating BUGS or JAGS syntax code from that abstract representation. A number of files have to be generated: the model code, a data file, an initial values file for each chain and a script file that specifies how the model should be run (this last file is not necessary for WinBUGS). To generate the model code, we use a template based on [Dias et al., 2011], which we generalized to also apply to inconsistency models (though that is outside the scope of this paper). The template and an example of the resulting code are shown in Figure 3.9. The BUGS and JAGS model specification languages are not fully compatible, but we have been careful to ensure that the template is compatible with both. The data and initial values files are output in S-Plus/R format for both BUGS and JAGS, structured according to [Dias et al., 2011]. For BUGS, care has to be taken with matrices, since it stores matrices in row-major order, whereas JAGS (like S-Plus and R) stores matrices in column-major order. Generating the script file is, again, a simple matter of variable substitution on a template. However, in this case, separate templates are required for JAGS and BUGS since their scripting languages are entirely different.

3.5 Results

In this section, we validate the approach presented above by reproducing a number of previously published analyses. The first MTC is a smoking cessation network comparing three forms of counseling with self-help consisting of 24 trials [Lu and Ades, 2006]. The second MTC consists of 28 trials comparing 8 thrombolytic treatments after acute myocardial infarction [Lu and Ades, 2006]. The third, a comparison of the efficacy of 12 second-generation anti-depressants, includes 111 clinical trials with 24,693 patients in total [Cipriani et al., 2009]. These three MTCs synthesize dichotomous data on the log odds ratio scale. The remaining two MTCs were selected to evaluate our methods for continuous datasets. The fourth dataset concerns the effect of psychological interventions on coronary heart disease, and assessed eight outcomes, three of which were expressed as mean change from baseline: diastolic blood pressure, systolic blood pressure and total cholesterol [Welton et al., 2009]. The fifth MTC compares 5 treatments for Parkinson's disease on mean off-time reduction using 7 trials [Dias et al., 2011], which was included because there is relatively little data, meaning that priors may have a greater influence on the result. The first two MTCs (smoking cessation and thrombolytic treatments) were also analyzed by Dias et al. [2010].

In Table 3.2 we contrast the priors specified for the original analyses with those generated using the heuristics described in Section 3.3.2, using $k = 15$. As can be seen, the priors used in the literature vary widely, and the ones generated by our algorithm appear to be no less reasonable. In two cases (systolic blood pressure and total cholesterol) the generated priors are much narrower than those reported in the original paper, but the posterior is unaffected (see below).

Figure 3.10 shows the evidence graphs of the smoking cessation, thrombolytics, Parkinson and anti-depressants data sets. The basic parameters chosen by our algo-

```
model {
  for (i in 1:ns) {
    # Likelihood for each arm
    for (k in 1:na[i]) {
      %armLikelihood%
    }

    # Study-level relative effects
    w[i, 1] <- 0
    delta[i, 1] <- 0
    # parameterize multi-arm trials using a trick
    # to avoid using the multi-variate normal
    for (k in 2:na[i]) {
      delta[i, k] ˜ dnorm(md[i, k], taud[i, k])
      md[i, k] <- d[t[i, 1], t[i, k]] + sw[i, k]
      taud[i, k] <- tau.d * 2 * (k - 1) / k
      w[i, k] <- delta[i, k] - d[t[i, 1], t[i, k]]
      sw[i, k] <- sum(w[i, 1:k-1]) / (k - 1)
    }
  }

  # Relative effect matrix
  %relativeEffectMatrix%

  # Study baseline priors
  for (i in 1:ns) {
    mu[i] ˜ dnorm(0, %priorPrecision%)
  }

  # Variance prior
  sd.d ˜ dunif(0, %upper%)
  tau.d <- pow(sd.d, -2)

  # Effect parameter priors
  %parameters%
}
```

(a) Template

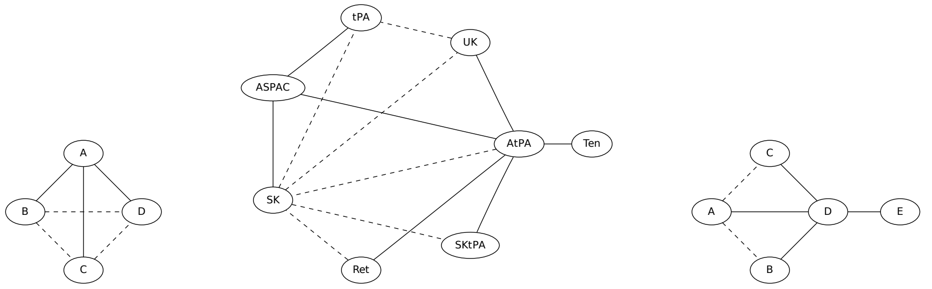
```
model {
  for (i in 1:ns) {
    # Likelihood for each arm
    for (k in 1:na[i]) {
      r[i, k] ˜ dbin(p[i, k], n[i, k])
      logit(p[i, k]) <- mu[i] + delta[i, k]
    }

    # Study-level relative effects
    w[i, 1] <- 0
    delta[i, 1] <- 0
    # parameterize multi-arm trials using a trick
    # to avoid using the multi-variate normal
    for (k in 2:na[i]) {
      delta[i, k] ˜ dnorm(md[i, k], taud[i, k])
      md[i, k] <- d[t[i, 1], t[i, k]] + sw[i, k]
      taud[i, k] <- tau.d * 2 * (k - 1) / k
      w[i, k] <- delta[i, k] - d[t[i, 1], t[i, k]]
      sw[i, k] <- sum(w[i, 1:k-1]) / (k - 1)
    }
  }

  # Relative effect matrix
  d[1, 1] <- 0
  d[1, 2] <- d.A.B
  d[2, 1] <- -d.A.B
  d[2, 2] <- 0

  # Study baseline priors
  for (i in 1:ns) {
    mu[i] ˜ dnorm(0, 9.0*10ˆ2)
  }

  # Variance prior
  sd.d ˜ dunif(0, 2.0*10ˆ0)
  tau.d <- pow(sd.d, -2)

  # Effect parameter priors
  d.A.B ˜ dnorm(0, 9.0*10ˆ2)
}
```

(b) Example

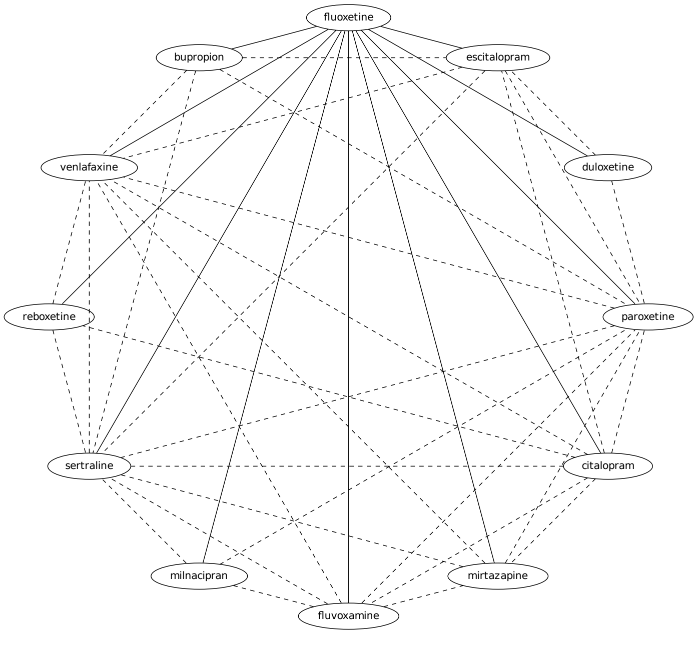
Figure 3.9: General BUGS/JAGS model template used to generate BUGS/JAGS code (a). Template variables (written as `%variableName%`) are replaced with the appropriate text by our code generation procedure. The resulting BUGS/JAGS code is illustrated for a pair-wise meta-analysis with dichotomous data (b). Note that the shown model (b) is specific to JAGS, as the priors are written in scientific notation as `9.0*10ˆ2`, whereas for BUGS they would be written as `9.0E-2`.

			Theirs			Ours	
Dataset	Outcome	Scale	u	η^2	$\hat{k}$	u	η^2
Smoking cessation [Lu and Ades, 2004]	Cessation	OR	2	10^3	15.8	3.52	$2.8 \cdot 10^3$
Smoking cessation [Dias et al., 2010]	Cessation	OR	10	10^4	10.0	3.52	$2.8 \cdot 10^3$
Thrombolytic drugs [Lu and Ades, 2004]	Death (30/35 days)	OR	2	10^3	15.8	1.36	$4.1 \cdot 10^2$
Thrombolytic drugs [Dias et al., 2010]	Death (30/35 days)	OR	10	10^4	10.0	1.36	$4.1 \cdot 10^2$
Anti-depressants [Cipriani et al., 2009]	Response	OR	NR	NR		1.98	$8.8 \cdot 10^2$
Coronary heart disease [Welton et al., 2009]	Diastolic blood pressure	MD	10	10^3	3.16	8.20	$1.5 \cdot 10^4$
	Systolic blood pressure	MD	50	10^3	0.63	8.60	$1.7 \cdot 10^4$
	Total cholesterol	MD	10	10^3	3.16	0.82	$1.5 \cdot 10^2$
	Cardiac mortality	OR	2	10^3	15.8	2.67	$1.6 \cdot 10^3$
	Non-fatal MI	OR	2	10^3	15.8	2.39	$1.3 \cdot 10^3$
	Total mortality	OR	2	10^3	15.8	2.47	$1.4 \cdot 10^3$
Parkinson's [Dias et al., 2011]	Off-time reduction	MD	5	10^4	20	2.30	$1.2 \cdot 10^3$

Table 3.2: Prior values employed in the literature ('Theirs'), and those based on our heuristics ('Ours'). u is the upper limit of the variance prior $\sigma \sim U(0, u)$, and η^2 is the variance for the normal priors $\mu, d \sim \mathcal{N}(0, \eta^2)$. For their priors, $\hat{k}$ is the value for k such that $\eta^2 = (ku)^2$. For our priors, η^2 is calculated using $k = 15$. MD = Mean Difference, OR = Odds Ratio, NR = Not Reported.


A
B
D
C
tPA
UK
ASPAC
AtPA
Ten
SK
SKtPA
Ret
C
A
D
E
B

(a) Smoking cessation (b) Thrombolytics (c) Parkinson


fluoxetine
bupropion
escitalopram
venlafaxine
duloxetine
reboxetine
paroxetine
sertraline
citalopram
milnacipran
mirtazapine
fluvoxamine

(d) Anti-depressants

Figure 3.10: Evidence graphs of the various data sets, with the basic parameters (spanning tree) shown as solid lines and the functional parameters as dashed lines

Parameter	Theirs	Ours	MC Error
d_{AB}	0.494 (0.399)	0.494 (0.405)	0.002
d_{AC}	0.844 (0.236)	0.838 (0.240)	0.001
d_{AD}	1.101 (0.437)	1.100 (0.442)	0.002
σ^2	0.731	0.751	0.003

Table 3.3: Results for the smoking cessation dataset [Lu and Ades, 2006], homogeneous variance consistency model. 'Theirs' are the posterior mean and standard deviation for the log odds ratio, as reported in the original paper. 'Ours' are the equivalent calculated based on our generated model for the same data. 'MC Error' is the Monte Carlo error of our estimates. The analysis was done using 4 independent chains, with 20,000 tuning (adaptive), 20,000 burn-in and 20,000 inference iterations.

Parameter	Theirs	Ours	MC Error
SK–AtPA	-0.219 (0.126)	-0.224 (0.136)	0.003
SK–tPA	-0.010 (0.088)	-0.020 (0.093)	0.002
SK–SK+tPA	-0.056 (0.134)	-0.057 (0.141)	0.001
SK–Ten	-0.212 (0.199)	-0.218 (0.214)	0.003
SK–Ret	-0.167 (0.142)	-0.171 (0.154)	0.003
SK–UK	-0.236 (0.232)	-0.233 (0.240)	0.008
SK–ASPAC	0.040 (0.102)	0.037 (0.107)	0.002
σ^2	0.020	0.023	0.001

Table 3.4: Results for the thrombolytics dataset [Lu and Ades, 2006], homogeneous variance consistency model. 'Theirs' are the posterior mean and standard deviation for the log odds ratio, as reported in the original paper. 'Ours' are the equivalent calculated based on our generated model for the same data. 'MC Error' is the Monte Carlo error of our estimates. The analysis was done using 4 independent chains, with 20,000 tuning (adaptive), 20,000 burn-in and 20,000 inference iterations.

rithm are shown as solid lines, and the functional ones as dashed lines. The coronary heart disease data sets are not shown, as they were analyzed in a pair-wise fashion. In most cases, the algorithm finds a single comparator against which all other treatments are compared. In the thrombolytics data set, however, no single comparator exists and a different solution is identified: some treatments are parameterized relative to ASPAC and others relative to AtPA.

Fixing $k = 15$ and $c = 2.5$, and using 4 chains with 20,000 tuning, 20,000 burn-in and 20,000 inference iterations, we reproduced the five published MTCs. The results are shown in Tables 3.3 through 3.7, where we show the posterior summaries reported in the original paper side by side with the results of our analysis. Convergence was assessed using the Brooks-Gelman-Rubin diagnostic and was adequate for all models. In all cases, the results of our generated models are very similar to those of the original analyses, and they are almost identical for the anti-depressants data (Table 3.5), as is to be expected since the priors should have very little influence on the posterior in such a dense dataset. There were no cases where our choice of prior led to truncation of the density for the variance parameter or to biased estimates for the relative effects,

Parameter	Theirs	Ours	MC Error
Fluoxetine–Bupropion	1.08 (0.90, 1.29)	1.08 (0.90, 1.30)	0.002
Fluoxetine–Citalopram	1.10 (0.93, 1.31)	1.10 (0.93, 1.31)	0.001
Fluoxetine–Duloxetine	0.99 (0.79, 1.24)	0.99 (0.78, 1.24)	0.002
Fluoxetine–Escitalopram	1.32 (1.12, 1.55)	1.32 (1.12, 1.55)	0.002
Fluvoxamine–Fluoxetine	1.02 (0.81, 1.30)	1.02 (0.80, 1.29)	0.003
Milnacipran–Fluoxetine	0.99 (0.74, 1.31)	0.99 (0.74, 1.32)	0.003
Mirtazapine–Fluoxetine	0.73 (0.60, 0.88)	0.73 (0.60, 0.88)	0.002
Paroxetine–Fluoxetine	0.98 (0.86, 1.12)	0.99 (0.86, 1.13)	0.002
Reboxetine–Fluoxetine	1.48 (1.16, 1.90)	1.48 (1.15, 1.91)	0.002
Sertraline–Fluoxetine	0.80 (0.69, 0.93)	0.80 (0.69, 0.94)	0.002
Venlafaxine–Fluoxetine	0.78 (0.68, 0.90)	0.78 (0.68, 0.90)	0.001

Table 3.5: Results for the efficacy of second-generation anti-depressants dataset [Cipriani et al., 2009]. 'Theirs' are the posterior median and 95% credibility interval for the mean difference, as reported in the original paper. 'Ours' are the equivalent calculated based on our generated model for the same data. 'MC Error' is the Monte Carlo error of our estimates. The analysis was done using 4 independent chains, with 20,000 tuning (adaptive), 20,000 burn-in and 20,000 inference iterations.

Outcome		Theirs	Ours	MC Error
Diastolic BP	d	-1.377 (-3.312, 0.6232)	-1.382 (-3.340, 0.617)	0.004
	σ	1.966 (0.223, 5.085)	1.966 (0.215, 5.088)	0.014
Systolic BP	d	-1.316 (-4.24, 2.326)	-1.318 (-4.184, 2.176)	0.008
	σ	3.438 (1.042, 8.21)	3.407 (1.063, 7.288)	0.016
Cholesterol	d	-0.3197 (-0.4975, -0.1316)	-0.320 (-0.498, -0.131)	0.000
	σ	0.277 (0.154, 0.4982)	0.277 (0.154, 0.495)	0.001

Table 3.6: Results for the blood pressure (BP) dataset [Welton et al., 2009], pair-wise random effects model. 'Theirs' are the posterior median and standard deviation for the mean difference, as reported in the original paper. 'Ours' are the equivalent calculated based on our generated model for the same data. 'MC Error' is the Monte Carlo error of our estimates. The analysis was done using 4 independent chains, with 20,000 tuning (adaptive), 20,000 burn-in and 20,000 inference iterations.

Parameter	Theirs	Ours	MC Error
d_{AB}	-1.84 (-2.91, -0.85)	-1.84 (-2.86, -0.88)	0.004
d_{AC}	-0.50 (-1.78, 0.75)	-0.50 (-1.74, 0.71)	0.007
d_{AD}	-0.53 (-1.77, 0.71)	-0.53 (-1.76, 0.67)	0.008
d_{AE}	-0.83 (-2.35, 0.69)	-0.84 (-2.31, 0.63)	0.009
σ	0.28 (0.01, 1.55)	0.28 (0.01, 1.39)	0.006

Table 3.7: Results for the Parkinson dataset [Dias et al., 2011], homogeneous variance consistency random effects model. 'Theirs' are the posterior mean and standard deviation for the log odds ratio, as reported in the original paper. 'Ours' are the equivalent calculated based on our generated model for the same data. The analysis was done using 4 independent chains, with 20,000 tuning (adaptive), 20,000 burn-in and 20,000 inference iterations.

nor did we detect slow convergence or spuriously low PSRFs.

3.6 Discussion

In this paper, we described how to generate MTC consistency models fully automatically based solely on the data set. We showed that parametrization of consistency models is indeed easy, and thus that we can optimize for understandability of the model (this is a marked difference with inconsistency models). Moreover, we proposed heuristics to safely choose defaults for priors and starting values. The methods were validated against published MTCs and the generated models give nearly identical results, but at a significant reduction in effort for the analyst.

The methods have been implemented in GeMTC (`http://drugis.org/gemtc`), which provides a graphical interface to manipulate data sets and can generate MTC models for either JAGS or BUGS. The software can also generate inconsistency models (based on previous work), and is used in the ADDIS decision support system (`http://drugis.org/addis`). An R package is being worked on, and is currently available in experimental form (`http://drugis.org/gemtc`). All of this software has been released under an open source license, and the datasets referred to in this paper are distributed with the software.

The presented results, though encouraging, do not guarantee correct results for all problem instances. Specifically, priors are chosen heuristically, and can not in any sense be shown to be ‘correct’. But if default values are to be given, we can at least try to ensure that they don’t result in overly precise estimates. This is what the given heuristics do and arguably that is an improvement over template code that gives a fixed default value. The prior for the random effects variance σ may be problematic in sparse datasets, as it can lead to over-estimation of the variance. Thus, one should be careful when using the software with sparse data sets. Naturally, if real prior information is available, the defaults should be replaced. Moreover, human judgment is still needed to choose the appropriate run length, assess convergence, and interpret the results. For use in simulation studies where each iteration requires the estimation of one or more MTCs, our methods can be combined with automated convergence checking to run the MTCs fully unsupervised. Wrongly concluding that convergence is adequate is a real risk in that case, but running a larger number of chains with properly over-dispersed starting values can help to minimize that risk.

The software vastly simplifies the task of performing an MTC by automating the process of model specification. Compared to existing ‘generic’ code for MTCs, the software presented in this paper facilitates error-checking of the data set, specifies vague priors appropriate for the problem rather than giving a fixed default, and automatically specifies multiple chains with over-dispersed starting values. The `mvmeta` package for the Stata statistical software enables MTC in a frequentist framework [White, 2011]. However, if there is no common comparator in the MTC this has to be handled by augmenting the data set with fictional arms with high variance, which is not very elegant and requires a decision as to what constitutes a sufficiently high variance. Moreover, one might prefer the Bayesian approach, and thus an automated solution for Bayesian MTC is still desirable.

The presented methods are limited to specific types of MTC model, and future work should address this limitation. First, we have assumed that arm-level outcome data is available, but this is often not the case. Continuous outcomes are often reported as relative effects compared to a control treatment, in which case the relative effects are correlated and these correlations have to be modelled in the likelihood function for multi-arm trials [Dias et al., 2011]. There are various formats in which such data may be reported, and the correlations may or may not be reported. This gives rise to some challenges in the data modeling and user interface, and several new variants of the likelihood function. We showed that study baselines can be chosen arbitrarily, so whether data are arm-based or contrast-based should not affect the basic model structure for consistency models. However, contrast-based data will create additional problems for inconsistency models.

Second, we did not discuss fixed effects models, but generating them entails a trivial modification of the template. A greater challenge lies in heterogeneous variance random effects models, as problems arise in sparse data sets, where some comparisons are informed by only one study. For those comparisons, the random effects variance can not be estimated, and the model generation code should detect these cases and potentially offer a solution. Moreover, for some datasets more specialized likelihoods are needed, such as for time-to-event data [Lu et al., 2007] and in some cases it may be necessary to adjust for covariates using a meta-regression model [Salanti et al., 2009]. Future work should investigate how these types of model can be generated automatically.

Finally, it has not been discussed in the literature how node split models for the assessment of inconsistency [Dias et al., 2010] can be generated. Automatically generating these models would be useful, as for each MTC analysis there will be a relatively large number of node split models that have to be specified: potentially one model for each comparison present in the dataset.

3.7 Appendix: proof

Theorem 3.1. *The choice of basic parameters and study baselines is arbitrary in consistency models*

Proof. A parameter is under-constrained if either (a) it is never used to define any random effect (Equation 3.3) or (b) it always co-occurs with another variable. We show here that this problem cannot occur for consistency models.

Let $G_i = (T_i, E_i)$ be the (fully connected) evidence graph for trial i, and $G = \bigcup_i G_i = (T, E)$ the (possibly sparse) evidence graph of the treatment network. Now, if R is any spanning tree of G (determining the $|T|-1$ basic parameters), and whatever $b(i) \in T_i$ we choose as the baseline in the individual trials, the basic parameters are fully constrained. To see this, note that a trial i is expressed as a star-shaped graph H_i in which all included treatments are compared to $b(i)$. If we take the union $H = \bigcup_i H_i = (T, E_H)$, this represents all comparisons that are explicitly expressed in the model. Since H is the union of spanning trees of the graphs that G is the union of,

H is connected. Each contrast $\{x, y\} \in E_H$ generates a (unique, undirected, simple) path between x and y in R, denoted as $\mathrm{path}_R(\{x, y\})$. Because H is connected, the set P of these paths visits all treatments T and thus uses all the basic parameters at least once. Now, if $\{x, y\}$ and $\{y, z\}$ are basic parameters, there is at least one path $p \in P$ that contains one, but not the other: for $\{y, w\} \in H$, $\mathrm{path}_R(\{y, w\})$ contains $\{x, y\}$ or $\{y, z\}$, but not both. Finally, if any two treatment contrasts co-occur, the entire path in R between them must co-occur. However, for any adjacent contrasts we can find a $p \in P$ so that they don't co-occur, so for any two basic parameters we can find a $p \in P$ so that they don't co-occur. In summary: each basic parameter is used at least once and no two basic parameters always co-occur. $\square$

CHAPTER 4

Automated generation of network meta-analysis inconsistency models

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Algorithmic parameterization of mixed treatment comparisons. *Statistics and Computing*, 22(5):1099–1111, 2012d. doi: 10.1007/s11222-011-9281-9

Abstract

Mixed Treatment Comparisons (MTCs) enable the simultaneous meta-analysis (data pooling) of networks of clinical trials comparing ≥ 2 alternative treatments. Inconsistency models are critical in MTC to assess the overall consistency between evidence sources. Only in the absence of considerable inconsistency can the results of an MTC (consistency) model be trusted. However, inconsistency model specification is non-trivial when multi-arm trials are present in the evidence structure. In this paper, we define the parameterization problem for inconsistency models in mathematical terms and provide an algorithm for the generation of inconsistency models. We evaluate running-time of the algorithm by generating models for 15 published evidence structures.

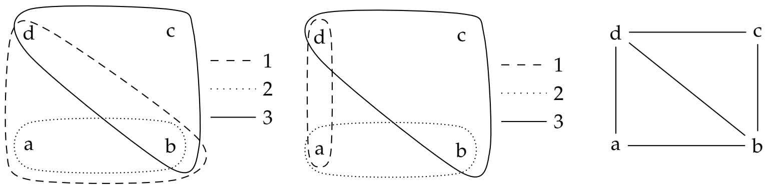
4.1 Introduction

Meta-analysis refers to statistical methods that summarize evidence from multiple studies (most commonly: clinical trials). Traditional meta-analysis [Hedges and Vevea, 1998, Normand, 1999] has focused on pair-wise comparisons of treatments based upon summary measures of relative effect as reported in the original studies. Several models to simultaneously compare more than two treatments have recently appeared in the literature [Sutton and Higgins, 2008], also leading to reported applications of the methodology [see Salanti et al., 2008b]. Such simultaneous comparisons are called Mixed Treatment Comparisons (MTCs), or network meta-analyses. MTCs allow the use of both direct and indirect evidence for comparisons, and to calculate the rank-probabilities of a set of alternative treatments with regard to a single evaluation criterion.

An MTC is implemented as a Bayesian hierarchical model and estimated using Markov Chain Monte Carlo (MCMC) simulation [Lu and Ades, 2006, Salanti et al., 2008a]. As in pair-wise meta-analysis, the goal is to combine evidence from multiple studies in order to derive a best estimate of the relative effect of treatments. MTC extends pair-wise meta-analysis by simultaneously estimating the relative effects of all possible pairs of the included treatments. Normally consistency is assumed, i.e., that direct and indirect evidence are in agreement. For example, if we have different studies comparing treatments a versus b, b versus c and a versus c, we add the constraint $d_{ac} = d_{ab} + d_{bc}$ to the model [Lu and Ades, 2006], where the d_{xy} are relative effects. This assumption of consistency does not necessarily hold and needs to be tested. To do this, an inconsistency model is formulated by relaxing the consistency constraint by introducing an Inconsistency Factor (IcF): $d_{ac} = d_{ab} + d_{bc} + w_{abca}$. Evidence can only be inconsistent if there are closed loops in the evidence structure [Lumley, 2002]: the IcF w_{abca} corresponds to the loop $abca$. Evidence consistency can be tested by individually assessing the null hypothesis that $w_C = 0$ for each IcF w_C [Salanti et al., 2008a], and further comparison of consistency and inconsistency models can be based on global goodness of fit [Lu and Ades, 2006].

No general formula or algorithm exists for evidence structures with multi-arm trials (trials with three or more arms – i.e., treatment groups) to determine the consistency equations that must be relaxed with IcFs to achieve correct model parameterization [Lu and Ades, 2006, Salanti et al., 2008a]. In addition, baseline treatments have to be chosen for the individual studies, which can prove to be problematic in the presence of multi-arm trials [Lu and Ades, 2006]. The absence of an algorithmic solution causes MTC model construction to be error prone and only applicable by experts in Bayesian modeling. Thus, MTC model generation would enable wider adoption of MTC and should allow greater confidence in the correctness of subsequently published MTCs. In this paper, we formally define the model generation problem for MTC inconsistency models and provide an algorithmic solution.

The remainder of the paper is structured as follows. First, an overview of MTC models and a mathematical formulation of their evidence structure is given in Section 4.2. Then, we give a precise definition of the parameterization problem as the problem of finding the spanning tree of the evidence structure that maximizes the



(a) structure I (b) structure II (c) evidence graph

Figure 4.1: Two examples of evidence structures used throughout the paper. Evidence structure I (a) contains three trials, where trial 1 compares a, b, and d, trial 2 compares a and b, and trial 3 compares b, c, and d. Evidence structure II (b) also contains three trials, but differs in that trial 1 compares only a and d. Both structures have an identical evidence graph (c).

number of IcFs, the Inconsistency Degree (IcD), while satisfying the constraint that every relevant parameter must be informed directly by at least one trial (Section 4.3). An algorithmic solution to the problem is given in Section 4.4. We give a detailed example of how parameterization is done in Section 4.5. In Section 4.6, we evaluate the feasibility of our algorithm on a number of published evidence structures. Finally, in Section 4.7, we discuss our results.

4.2 Mixed treatment comparison models

The Bayesian hierarchical model for an MTC evidence structure is specified following the general formulation in Lu and Ades [2006], which in turn extends that by Higgins and Whitehead [1996]. We shall only introduce the concepts that are relevant to the parameterization problem, and refer the interested reader to Lu and Ades [2006] for a full discussion. The evidence structure for any MTC consists of a number of studies, that together determine an undirected evidence graph in which the treatments are the vertices and the available comparisons are the edges. Since a trial S_i provides evidence for all possible comparisons between the included treatments $T(S_i)$, each study can be considered to provide a fully connected evidence graph $G(S_i) = (T(S_i), E(S_i))$. Here, $E(S_i)$ represents the estimates of relative effects that can be made based on the trial data. So a two-arm trial is a pair, a three-arm trial a triangle, a four-arm trial a fully connected 4-treatment graph, and so on.

Denote by $S = \{S_1, \ldots, S_n\}$ the set of n studies included in the MTC. The evidence graphs $G(S_i)$, $S_i \in S$ form an evidence structure, as illustrated in Figure 4.1(a) and Figure 4.1(b). These figures introduce two hypothetical examples that will be used throughout the paper to illustrate the introduced concepts. Structure I consists of two overlapping three-arm trials and one two-arm trial, while structure II has only one three-arm trial and two two-arm trials. The union of the individual study evidence graphs forms the MTC evidence graph:

Definition 4.1 (evidence graph)**.** *The graph $G(S)$ of all comparisons made in at least one* *of the trials in S is defined as:*

$$G(S) = (T(S), E(S)) = \left(\bigcup_{S_i \in S} T(S_i), \bigcup_{S_i \in S} E(S_i) \right) \ .$$

For example, the evidence structures I and II have the same evidence graph, shown in Figure 4.1(c). A graph corresponding to an MTC problem has to be connected. If it is not, S must be decomposed into two or more independent problems, corresponding to connected subgraphs of $G(S)$ that can be analyzed separately. Given the (connected) evidence graph $G(S)$, every edge in $E(S)$ becomes an effect parameter in the MTC model:

Definition 4.2. *Given the MTC problem S and an arbitrary ordering $\prec$ on $T(S)$ (e.g., alphabetical order of treatments), the set of effect parameters $D(S)$ is given by:*

$$D(S) = \{d(\{x, y\}) \mid \{x, y\} \in E(S)\}$$

where $d(\cdot)$ identifies a unique parameter with the set $\{x, y\}$:

$$d(\{x, y\}) = \begin{cases} d_{xy} & \text{if } x \prec y \\ d_{yx} & \text{if } y \prec x \ . \end{cases}$$

Furthermore, for directed edges (x, y), we define a signed function, that takes into account the direction of (x, y) relative to the parameter $d(\{x, y\})$:

$$d((x, y)) = \begin{cases} d(\{x, y\}) & \text{if } x \prec y \\ -d(\{x, y\}) & \text{if } y \prec x \ . \end{cases}$$

For example, if we adopt alphabetical ordering for $\prec$, we have

$$\begin{aligned} d((a, b)) &= d_{ab} \\ d((b, a)) &= -d_{ab} \end{aligned}$$

meaning we do not have to worry about the direction in which the evidence graph is traversed.

4.2.1 Study level effects

We first discuss how the relative effects at the study level are parameterized in terms of the parameters $D(S)$. For each $S_i \in S$, for each treatment $t \in T(S_i)$, we have a certain absolute effect μ_{it}. The way these absolute effects are defined depends on the type of model, and is not important for the current discussion; Salanti et al. [2008a] gives the formulation for both dichotomous and continuous data. Now, because we are interested in the *relative* effects, we choose a *baseline* treatment $x \in T(S_i)$. The baseline effect μ_{ix} is then a random variable for which we assume some prior distribution $\pi(\mu_{ix})$. For every other treatment $u \in T(S_i)$, $u \neq x$ the treatment effect is:

$$\mu_{iu} = \mu_{ix} + \delta_{ixu} \ ,$$

where δ_{ixu} is the random effect of treatment u relative to x. The distribution for the random effects is:

$$\begin{pmatrix} \delta_{ixu} \\ \vdots \\ \delta_{ixw} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} d((x,u)) \\ \vdots \\ d((x,w)) \end{pmatrix}, \Sigma \right) ,$$

where Σ is an appropriately defined variance-covariance matrix. For a full discussion of the study level absolute and relative effects, see [Lu and Ades, 2006, Salanti et al., 2008a].

From the definition of the relative effects, it is clear that they are transitive [Lu and Ades, 2009]: if $u, v, w \in T(S_i)$ are distinct treatments, then

$$\delta_{iuv} = \delta_{iuw} + \delta_{iwv} \ . \tag{4.1}$$

Based on this we conjecture that if a triangle of treatments is included in precisely the same set of studies, then this relation also holds for the estimates of the effect sizes:

Conjecture 4.3 (internal consistency). *Write $f((u,v),S')$ for the estimate of $d((u,v))$ based solely on the studies $S' \subset S$. Then if $u, v, w \in T(S_i) \ ; \ \forall S_i \in S'$,*

$$f((u,v),S') = f((u,w),S') + f((w,v),S') \ .$$

Note that this implies that each of the studies in S' has at least three arms.

4.2.2 Consistency models

Normally conclusions are drawn under the assumption of evidence consistency. Basically, this is a generalization of the conjecture in the sense that we assume that it holds regardless of the supporting studies:

Definition 4.4 (consistency). *Let $u, v, w \in T(S)$ be distinct treatments, then assuming consistency,*

$$d((u,v)) = d((u,w)) + d((w,v)) \ .$$

This can be justified by assuming exchangeability of the study level relative effects and taking expectations on both sides of Equation 4.1 [Lu and Ades, 2009]. The consistency assumption is essential, as it models the relationships between treatment contrasts and allows the model to borrow strength across the evidence structure [Lu and Ades, 2009]. More generally, a consistency equation can be written for any cycle and reference effect, as shown by the following lemma and corollary.

Lemma 4.5. *Given the evidence graph $G(S)$, let (w_1, w_n) be any pair of vertices of $G(S)$ and $p = (w_1, \ldots, w_n)$ a path of length $n-1$ between them, $n > 2$ (see Appendix 4.8). Then, under the assumption of consistency,*

$$d((w_1, w_n)) = \sum_{i=1}^{n-1} d((w_i, w_{i+1})) \tag{4.2}$$

Proof by induction. If $p = (u, w, v)$, then the lemma is just a restatement of the assumption. Now let the lemma hold for (w_2, w_n) and $p' = (w_2, \dots w_n)$, i.e.,

$$d((w_2, w_n)) = \sum_{i=2}^{n-1} d((w_i, w_{i+1})) \ .$$

Then for $p = (w_1, w_2, \dots, w_n)$, we get

$$\begin{aligned} d((w_1, w_n)) &= d((w_1, w_2)) + d((w_2, w_n)) \\ &= \sum_{i=1}^{n-1} d((w_i, w_{i+1})) \ . \end{aligned}$$

The first equality by the consistency assumption and the second by the induction hypothesis. □

Corollary 4.6 (consistency relation). *Given the evidence graph $G(S)$, a (simple) cycle $C \subseteq E(S)$, then if we take any edge $\{u, v\} \in C$, $u \prec v$ as a reference effect, we can write a consistency equation as follows:*

$$d_{uv} = d((u, v)) = \sum_{i=1}^{n-1} d((w_i, w_{i+1})) \ ,$$

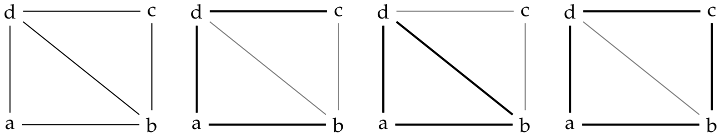
where $(w_1, \dots, w_n)$ is the directed path from $w_1 = u$ to $w_n = v$ consisting of the edges $(C - \{u, v\})$.

Thus, in a consistency model, the consistency relation defines d_{uv} completely in terms of the other comparisons in the cycle. In that case, d_{uv} is called a *functional* parameter [Lu and Ades, 2006]. For each functional parameter there must be a cycle in which it is the only one, otherwise a circular definition would result. Moreover, each cycle should have at least one functional parameter, or we do not assume full consistency. The right hand side parameters are called *basic* parameters and are defined through suitable distributions. The division of parameters into basic and functional ones is not arbitrary; it has previously been stated that the basic parameters should form a spanning tree [Lu and Ades, 2006]. This is proven by the following theorem.

Theorem 4.7 (basic parameters). *If we divide the parameter edges $E(S)$ into a set of basic parameters E_b and a set of functional parameters E_f, such that $E_b \cup E_f = E(S)$ and $E_b \cap E_f = \emptyset$, the basic parameters form a spanning tree $G_b = (T_b, E_b)$ of the evidence graph $G(S)$.*

Proof. It is sufficient to show (Appendix 4.8) that

1. G_b is a connected graph
2. G_b is acyclic
3. $T_b = T(S)$



(a) Evidence graph (b) Spanning tree (c) Cycle I (d) Cycle II

Figure 4.2: Choosing a spanning tree (b) for an evidence graph (a) determines the partition into basic and functional parameters. The spanning tree induces a set of fundamental cycles, (c) and (d), that determine the equations that define the functional parameters.

Proof of 1. Assume G_b is not connected. Then, since $G(S)$ is connected, there is an edge $e \in E(S)$, $e \notin E_b$ that connects two vertices not connected in G_b. Since $e = (u, v)$ does not correspond to a basic parameter, it must be a functional parameter. Hence, there must be a simple directed path from u to v in G_b, and therefore G_b must be connected.

Proof of 2. If $C \subseteq E_b$ is a cycle in G_b, then for an edge $e = (u, v) \in C$, $u \prec v$, Corollary 4.6 lets us write a consistency equation in terms of the other (basic) parameters in the cycle. Thus, if there would be a cycle in G_b, we would not be assuming full consistency.

Proof of 3. From the proof to the first part and $E_b \cup E_f = E(S)$. □

Corollary 4.8. *The functional parameters E_f are the non-tree edges (Appendix 4.8) corresponding to the spanning tree G_b, and the (simple) cycle C created by adding $e \in E_f$ to G_b generates a consistency relation, as considered in Corollary 4.6.*

The theorem and corollary imply that for any valid parameterization of $G(S)$, we will have $\dim(G(S)) = |T(S)|-1$ basic parameters and $\mathrm{nul}(G(S)) = |E(S)|-|T(S)|+1$ functional parameters (Appendix 4.8). For example, in Figure 4.2 we show a spanning tree of the evidence graph in Figure 4.1(c). Specifically, it is made up of $|T(S)| - 1 = 3$ basic parameters:

$$E_b = \{\{a, b\}, \{a, d\}, \{d, c\}\} \ ,$$

and thus there are $5 - 3 = 2$ functional parameters:

$$E_f = \{\{b, d\}, \{b, c\}\} \ .$$

Corresponding to the first functional parameter, $\{b, d\}$, we identify the cycle $badb$. This implies that:

$$d((b, d)) = d((b, a)) + d((a, d)) \ ,$$

and for the second functional parameter we get:

$$d((b, c)) = d((b, a)) + d((a, d)) + d((d, c)) \ .$$

4.2.3 Inconsistency models

The assumption of consistency does not necessarily hold and should be tested. Inconsistency can only occur if there are closed loops in the evidence structure [Lumley, 2002]. An inconsistency relation is obtained by expanding a consistency relation with an IcF, e.g., for a loop $abca$, we add w_{abca} [Lu and Ades, 2006]:

$$d_{ac} = d_{ab} + d_{bc} + w_{abca} \quad , \tag{4.3}$$

for which we again assume some distribution [Lu and Ades, 2006]. If multi-arm trials are included, some of the comparisons may be informed by only multi-arm trials, and evidence within a multi-arm trial is consistent by definition. For example, if we replace trials of a versus b, b versus c and a versus c with a three-arm trial a versus b versus c, the inconsistency model would not include w_{abca}. As we show in Section 4.3, the choice of basic parameters determines the number of IcFs that are included in the model.

4.3 Problem definition

The parameterization of an evidence structure requires partition of the parameters into *basic* and *functional* parameters, as presented in Theorem 4.7. However, when an inconsistency model is constructed for an evidence structure with multi-arm trials, the choices of the spanning tree and the individual study baselines are not arbitrary.

4.3.1 Spanning tree selection

If S contains only two-arm studies, then we may choose any spanning tree of $G(S)$ [Lu and Ades, 2006, Salanti et al., 2008a]. By contrast, when multi-arm studies are present, the choice of spanning tree is not arbitrary as some contrasts may be informed by only multi-arm trials, and the measurements within a multi-arm trial can not be inconsistent. For a cycle to potentially be inconsistent, it must be supported by at least three independent sources of evidence [Lu and Ades, 2006], which is formalized below in Theorem 4.13. To be able to do this, we introduce the concept of the *partition* of a cycle into comparisons with their supporting studies and the operation of *reduction*, which allows us to simplify a partition.

Definition 4.9 (elementary partition of C)**.** *Let C be a directed cycle in $G(S)$, represented by its set of (directed) edges. The elementary partition of C is (P, r), where $P = \{e \mid e \in C\}$, and $r(e) = \{S_i \in S \mid e \in E(S_i)\}$.*

Note that there are, for each (undirected) cycle, two possible elementary partitions, depending on the direction in which the cycle is traversed. Again using the evidence structure of Figure 4.1(a), an elementary partition of the cycle $abcda$ is (P, r), where:

$$\begin{gathered} P = \{(a,b),(b,c),(c,d),(d,a)\} \quad , \\ r((a,b)) = \{1,2\} \quad , \\ r((b,c)) = r((c,d)) = \{3\} \quad , \\ r((d,a)) = \{1\} \quad . \end{gathered}$$

Given an elementary partition (P, r) of an evidence cycle C, the inconsistency equation is given by:

$$w_C = F(P, r) = \sum_{(u,v) \in P} f((u, v), r((u, v))) \ , \tag{4.4}$$

with $f((u, v), S')$ defined as in Conjecture 4.3. This is a generalization of Equation 4.3. Based on the conjecture, it seems that if two adjacent comparisons have the same set of supporting studies, we should be able to simplify the equation. We call this reducing the partition:

Definition 4.10 (reduction). *Let (P, r) be a partition of C and $e_k = (u, w_1), \ldots, e_l = (w_n, v) \in P$ a sequence of pair-wise adjacent edges, such that $r(e_k) = \cdots = r(e_l)$. Then we may reduce this partition to (P', r'); where $P' = (P - \{e_k, \ldots, e_l\}) \cup \{(u, v)\}$, and*

$$r'(e) = \begin{cases} r(e) & \text{if } e \neq (u, v) \\ r(e_k) & \text{if } e = (u, v) \end{cases}$$

Note that the numbering of the e_i is arbitrary and that for any specific reduction step, we can always choose the numbering scheme such that the reduced sequence $e_k, \ldots, e_l$ does not contain the subsequence e_n, e_1. Given this definition, it is natural to think of two edges e_i and e_j as *independent* if $r(e_i) \neq r(e_j)$ (and dependent otherwise). For convenience, we will call any pair (P, r) that was obtained from the elementary partition of C by (repeated) application of reduction, a partition of C. For example, the previously discussed elementary partition (P, r) of the cycle $abcda$ can be reduced, because $r(b, c) = r(c, d) = \{3\}$. We get (P', r'), where

$$\begin{gathered} P' = \{(a, b), (b, d), (d, a)\} \ , \\ r'((a, b)) = \{1, 2\}, r'((b, d)) = \{3\}, r'((d, a)) = \{1\} \ . \end{gathered}$$

The following lemma shows that the inconsistency equation (Equation 4.4) is preserved under reduction of partitions:

Lemma 4.11. *Assume (P, r) is a partition of the cycle C, and (P', r') is obtained from (P, r) by a single reduction step. Then $F(P, r) = F(P', r')$.*

Proof. Let $e_1, \ldots, e_k$ be the edges reduced to e', then $r'(e') = r(e_1) = \cdots = r(e_k)$. Now, if P has n edges:

$$\begin{aligned} F(P, r) &= \sum_{i=1}^{n} f(e_i, r(e_i)) \\ &= \sum_{i=1}^{k} f(e_i, r(e_i)) + \sum_{i=k+1}^{n} f(e_i, r(e_i)) \\ &= f(e', r'(e')) + \sum_{i=k+1}^{n} f(e_i, r'(e_i)) \\ &= F(P', r') \ . \end{aligned}$$

Here, $f(e', r'(e')) = \sum_{i=1}^{k} f(e_i, r(e_i))$ holds by the same induction argument used for Lemma 4.5, but this time using Conjecture 4.3. □

Lemma 4.12. *If a partition (P, r) of C contains $k > 1$ independent pairs of adjacent edges (e_i, e_j), then there is a reduced partition (P', r') composed of k adjacently independent edges, such that $F(P, r) = F(P', r')$.*

Proof. By the previous lemma, a single reduction step will preserve $F(P, r)$, thus so will repeated reduction. It remains to be shown that there is a reduction with exactly k edges, each independent of its adjacent edges. To see this, number the edges $e_1, \ldots, e_n$ so that $r(e_1) \neq r(e_n)$. Create a strictly increasing index list $i(1), \ldots, i(k-1)$ so that $r(e_{i(j)}) \neq r(e_{i(j)+1}); \forall_{1 \leq j \leq k-1}$. Then if we set $i(k) = n$, this list enumerates all independent pairs of adjacent edges. We can reduce $e_1, \ldots, e_{i(1)}$ to e'_1, $e_{i(1)+1}, \ldots, e_{i(2)}$ to e'_2 and so on, until $e_{i(k-1)+1}, \ldots, e_{i(k)}$ to e'_k. Then (P', r') with $P' = \{e'_1, \ldots, e'_k\}$ and $r'(e'_j) = r(e_{i(j)})$ consists of k adjacent independent edges. The reduction is unique up to the numbering of the e'_j. □

The lemma leads to a simple test of when an inconsistency can occur in an evidence cycle, as given in the following theorem. We make the distinction between *potentially* inconsistent, which is a property of the evidence structure, and *actually* inconsistent, which depends additionally on the data. A cycle is potentially inconsistent if we can devise data so that it becomes actually inconsistent.

Theorem 4.13 (inconsistency cycle)**.** *Let C be a cycle of length n and suppose that the elementary partition (P, r) of C has m independent pairs of adjacent edges. Then, C is potentially inconsistent iff $m \geq 3$.*

Proof. Case I ($m < 3$) The first possibility is that all studies include the complete set of vertices in C ($m = 0$), and through internal consistency we have:

$$F(P, r) = \sum_{e \in P} F(e, r(e)) = 0 \ .$$

From Lemma 4.12, if $m = 1$, C is not a cycle. If $m = 2$, we derive, using Lemma 4.12:

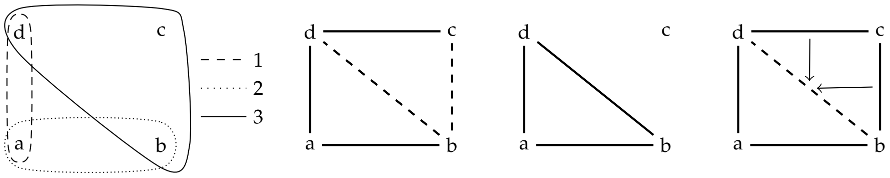
$$\begin{aligned} F(P, r) &= F(\{(u, v), (v, u)\}, r) \ ; \\ r((u, v)) &= R_1, \ r((v, u)) = R_2 \ . \end{aligned}$$

Thus, $F(P, r) \neq 0$ reduces to

$$\begin{aligned} f((u, v), R_1) + f((v, u), R_2) &\neq 0 \\ f((u, v), R_1) &\neq f((u, v), R_2) \ , \end{aligned}$$

which is just inter-study heterogeneity.

Case II ($m \geq 3$) Using Lemma 4.12, reduce the elementary partition to a partition where each pair of adjacent edges is independent. Since there are at least three distinct sets of supporting studies, the equations cannot be reduced as was done for $m < 3$, and this gives us sufficient freedom to choose data so that $w_C \neq 0$ without reducing to heterogeneity. Thus, C is potentially inconsistent. □



(a) Structure II (b) Parameterization (c) Cycle 1 (d) Cycle 2

Figure 4.3: An evidence structure (a) and spanning tree (b) in which two fundamental cycles reduce to the same set of equations. The bcd path in (d) collapses (through reduction) and leaves the same cycle as in (c), because (b, c), (b, d) and (c, d) are all supported only by study 3.

In evidence structure II (repeated in Figure 4.3(a)), if we consider the cycle $bcdb$, each of the comparisons is supported only by study 3, and hence $m = 0$, so according to the theorem, $bcdb$ is *not* potentially inconsistent. On the other hand, each of the comparisons in $abda$ is supported by a different study, so $m = 3$, making this cycle potentially inconsistent. The same holds for the longer cycle $abcda$, in which (b, c) and (c, d) are both supported by study 3, (d, a) by study 1 and (a, b) by study 2, also giving $m = 3$.

It would appear that this theorem allows us to count the number of inconsistency cycles for a given spanning tree. However, although the fundamental cycles for any spanning tree are independent, some may reduce to the same set of linear equations. An example of this is shown in Figure 4.3, where the cycles $abda$ and $abcda$ discussed previously have the same reduced partition, namely (P, r) with:

$$P = \{(a, b), (b, d), (d, a)\}$$
$$r((a, b)) = \{2\}, r((b, d)) = \{3\}, r((d, a)) = \{1\} \ .$$

Hence, to count the number of inconsistencies, we should count the number of distinct reduced partitions among the fundamental cycles. Moreover, any cycles that reduce to the same set of linear equations should be assigned the same IcF. The following definitions and lemma make this notion precise:

Definition 4.14. *Let $g((u, v)) = (u, v)$ or $g((u, v)) = (v, u)$ be a one-one correspondence $P_1 \to P_2$. Then the partitions (P_1, r_1) and (P_2, r_2) are equivalent if*

$$r_1(e) = r_2(g(e)) \ ; \ \forall e \in P_1 \ .$$

Lemma 4.15. *Let (P_1, r_1) and (P_2, r_2) be equivalent partitions under g. Then $F(P_1, r_1) = F(P_2, r_2)$ or $F(P_1, r_1) = -F(P_2, r_2)$, for $g((u, v)) = (u, v)$ or $g((u, v)) = (v, u)$ respectively.*

Proof. Assuming $g((u, v)) = (u, v)$, we have

$$f((u, v), r_1((u, v))) = f(g((u, v)), r_2(g((u, v)))) \ ,$$

and thus

$$\begin{aligned}F(P_1, r_1) &= \sum_{e \in P_1} f(e, r_1(e)) \\ &= \sum_{g(e) \in P_2} f(g(e), r_2(g(e))) = F(P_2, r_2) \ ,\end{aligned}$$

where the second equality holds because g is a one-one correspondence. If $g((u,v)) = (v,u)$, we have that

$$\begin{aligned}f((u,v), r_1((u,v))) &= -f((v,u), r_2((v,u))) \\ &= -f(g((u,v)), r_2(g((u,v)))) \ ,\end{aligned}$$

since $d((u,v)) = -d((v,u))$. Then clearly

$$F(P_1, r_1) = -F(P_2, r_2) \ .$$

□

Definition 4.16 (S-equivalence)**.** *Two cycles C_1 and C_2 are S-equivalent ($C_1 \sim_S C_2$) iff their maximally reduced elementary partitions (in the evidence structure S) are equivalent.*

By this definition the cycles $abda$ and $abcda$ shown in Figure 4.3 and discussed above are S-equivalent. This means that if we assign the inconsistency factor w to $abda$:

$$d((b,d)) = d((b,a)) + d((a,d)) + w \ ,$$

we should assign the same one to $abcda$:

$$d((b,c)) = d((b,a)) + d((a,d)) + d((d,c)) + w \ .$$

Note that according to Lemma 4.15, the direction in which we go around the cycle matters. In the above case the equivalence is due to $g((u,v)) = (u,v)$, so we use $+w$. If we traverse it in the other direction, $g((u,v)) = (v,u)$, we should use $-w$:

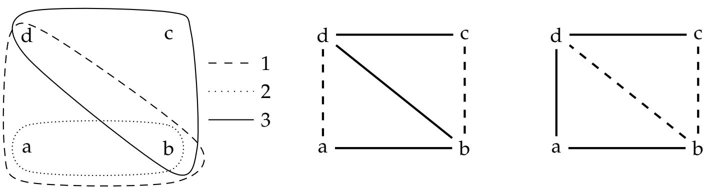
$$d((c,b)) = d((c,d)) + d((d,a)) + d((a,b)) - w \ .$$

Definition 4.17 (inconsistency degree)**.** *For an evidence structure S and spanning tree G_b, let $\mathbf{C} = \mathbf{C}(G(S), G_b)$ be the set of fundamental cycles. Then, $\mathbf{C}/\sim_S$ is the set of equivalence classes under $\sim_S$ in $\mathbf{C}$. The IcD of G_b is the number of equivalence classes that contain inconsistency cycles:*

$$\mathrm{icd}(S, G_b) = \sum_{X \in \mathbf{C}/\sim_S} \mathrm{icc}(S, C) \ ; \ C \in X \ ,$$

where $C \in X$ may be chosen arbitrarily and

$$\mathrm{icc}(S, C) = \begin{cases} 1 & \text{if } C \text{ is an inconsistency cycle} \\ 0 & \text{otherwise} \ . \end{cases}$$



(a) Structure I (b) Incorrect (c) Correct

Figure 4.4: An evidence structure in which a non-obvious choice of IcFs is required to arrive at the correct IcDF. The structure (a) contains three trials, where trial 1 compares a, b and d, trial 2 compares a and b and trial 3 compares b, c and d. The subfigures give an incorrect (b) and a correct (c) parameterization.

To clarify the meaning of the quotient set $\mathbf{C}/\sim_S$, consider again the situation of Figure 4.3. Since both cycles are equivalent, we have

$$\mathbf{C}/\sim_S = \{\{abda, abcda\}\}\ .$$

On the other hand, for the evidence structure in Figure 4.4(a) and the spanning tree of Figure 4.4(b), the cycles $abda$ and $bcdb$ are clearly not equivalent, so then

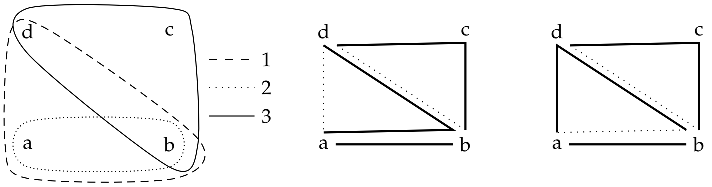
$$\mathbf{C}/\sim_S = \{\{abda\}, \{bcdb\}\}\ .$$

Definition 4.17 allows us to count the inconsistency degree of a spanning tree. For example, consider the evidence structure in Figure 4.4(a). Clearly, both Figure 4.4(b) and Figure 4.4(c) are parameterized so that the $\dim(G(S)) = 3$ basic parameters (solid edges) form a spanning tree, and the remaining $\mathrm{nul}(G(S)) = 2$ edges become the functional parameters. This implies that the IcDF is at most 2, the number of functional parameters. The spanning tree G_1 in Figure 4.4(b) has two fundamental cycles, namely $abda$ and $bdcb$. The first cycle is partitioned into $e_1 = (a, b)$, $e_2 = (b, d)$ and $e_3 = (d, a)$, with support $r(e_1) = \{1, 2\}$, $r(e_2) = \{1, 3\}$ and $r(e_3) = \{1\}$. Since the sets of supporting studies are all distinct, Theorem 4.13 leads us to conclude that $\mathrm{icc}(abda) = 1$. For the latter cycle, we have $e_1 = (b, d)$, $e_2 = (d, c)$ and $e_3 = (c, b)$, for which the supporting studies are $r(e_1) = \{1, 3\}$, $r(e_2) = \{3\}$ and $r(e_3) = \{3\}$. Thus, using Theorem 4.13, this reduces to heterogeneity on (b, d), so $\mathrm{icc}(bdcb) = 0$. Hence, in this parameterization $\mathrm{icd}(S, G_1) = 1$.

Now, consider the tree G_2 in Figure 4.4(c), with fundamental cycles $abda$ and $abcda$. We already know that $\mathrm{icc}(abda) = 1$. The partition of $abcda$ reduces to $e'_1 = (a, b)$, $e'_2 = (b, d)$, and $e'_3 = (d, a)$ with $r'(e'_1) = \{1, 2\}$, $r'(e'_2) = \{3\}$, and $r'(e'_2) = \{1\}$. All three edges are independent, and hence $\mathrm{icc}(abcda) = 1$. Moroever, the partitions of $abda$ and $abcda$ are not equivalent and thus $\mathrm{icd}(S, G_2) = 2$, the maximum possible. Hence, the choice of spanning tree determines the IcD:

Lemma 4.18. *The IcD* $\mathrm{icd}(S, G_b)$ *depends on the chosen spanning tree* G_b*.*

Theorem 4.19 (spanning tree selection problem)**.** *To parameterize the model correctly, we need to find a spanning tree* G_b *that maximizes* $\mathrm{icd}(S, G_b)$*. Then,* $\mathrm{icdf}(S) = \mathrm{icd}(S, G_b)$*.*



(a) Structure I (b) Incorrect (c) Correct

Figure 4.5: An evidence structure in which the choice of study baselines is not arbitrary. The subfigures (b)–(c) give an incorrect and a correct choice of baselines. The solid edges connect to the study baseline. In (b) baselines are 1:b, 2:a, 3:c and in (c) 1:d, 2:a, 3:c. The dotted edges are not connected to the baseline for that study and are thus not informed by direct evidence from that study.

Proof. $\mathrm{icd}(S, G_b)$ determines the number of independent inconsistency factors in the model. It has previously been shown that the IcFs $\mathbf{w}$ under one parameterization can be represented as linear combinations of the IcFs $\mathbf{w}'$ under another [Lu and Ades, 2006], assuming equal icd. However, from Lemma 4.18, not all spanning trees result in the same icd. Therefore, in order to be able to express any IcF as a linear combination of the chosen IcFs, a maximal set of independent IcFs must be chosen. □

Only one IcF should be created for each equivalence class of inconsistency cycles. Thus, whereas Lu and Ades [2006] claim that each IcF corresponds to exactly one functional parameter, actually each IcF may correspond to several. Their assertion that not every functional parameter need correspond to an IcF is confirmed by our work.

4.3.2 Baseline selection

The individual studies have to be parameterized in such a way that every comparison for which there is direct evidence (and which can be inconsistent) is expressed in the parameterization of at least one trial [Lu and Ades, 2006]. Again, this problem occurs only for multi-arm trials. For example, consider the structure in Figure 4.5(a); we might parameterize trial 1 with b as the baseline and trial 3 with c as the baseline (Figure 4.5(b)), having δ_{1ba}, δ_{1bd}, δ_{2ab}, δ_{3cb}, and δ_{3cd} as study parameters. Consider the cycle $abda$, where we have the inconsistency relation $d_{ad} = d_{ab} + d_{bd} + w_{abda}$. Now, d_{ad} is not informed directly by any of the study parameters δ_{ixy}, and hence the choice of d_{ad} is free, meaning that w_{abda} is also unconstrained. Hence, given this parameterization of the individual studies, the IcF w_{abda} cannot be estimated. A correct choice of baselines, covering all edges, is given in Figure 4.5(c).

Thus, in addition to choosing the basic parameters correctly, the study baselines must be chosen so that at least one study provides direct evidence where needed. That is, every cycle C for which $\mathrm{icc}(C) = 1$, all $|C|$ parameters should have direct evidence, while for cycles where $\mathrm{icc}(C) = 0$, only $|C| - 1$ need direct evidence. In case of an equivalence class of inconsistency cycles, only one of the cycles needs all $|C|$ parameters to have direct evidence. This is formalized as follows:

Definition 4.20 (evidence cover constraint). *Let* $X \in \mathbf{C}/\sim_S$ *be an equivalence class of cycles. Let every cycle* $C \in X$ *be represented by its edge-set. Define the indicator function* φ_X *that is* 1 *if the direct evidence constraint is satisfied by the edge set* E*:*

$$\varphi_X(E) = \begin{cases} 1 & \text{if } \forall_{C \in X}(|C \cap E| \geq |C| - 1) \wedge \\ & \quad \exists_{C \in X}(\text{icc}(C) = 0 \vee |C \cap E| = |C|) \\ 0 & \text{otherwise} \end{cases}$$

The individual studies are parameterized by choosing a baseline, and the effect parameters are the relative effects of all other treatments compared to the chosen baselines:

Definition 4.21 (baseline study graph). *Given a study* S_i *and a baseline* $b_i \in T(S_i)$*, the baseline study edge set is:*

$$E(S_i, b_i) = \{\{b_i, x\} \mid x \in (T(S_i) - b_i)\}$$

And the baseline study graph is given by $G(S_i, b_i) = (T(S_i), E(S_i, b_i))$*.*

Any given choice of baselines results in a graph representing the relative effects supported by at least one source of direct evidence:

Definition 4.22 (baseline evidence graph). *Given a baseline assignment*

$$B = \{(S_i, b_i) \mid S_i \in S\} \ ,$$

the baseline evidence graph is:

$$G(S, B) = (T(S), E(S, B)) = \left(T(S), \bigcup_{S_i \in S} E(S_i, b_i)\right) \ .$$

Thus, the baseline selection problem is to find the baseline assignment that simultaneously satisfies the evidence cover constraint (Definition 4.20) for all equivalence classes of fundamental cycles:

Definition 4.23 (baseline selection problem). *Given* S *and a spanning tree* G_b *for* S*, the baseline selection problem is to find a baseline assignment* B*, that satisties the constraint*

$$\varphi_X(E(S, B)) = 1 \ ; \ \forall X \in \mathbf{C}(G(S), G_b)/\sim_S$$

4.3.3 Parameterization problem

Together, the problems of maximizing the IcDF and selecting the baselines form the parameterization problem:

Definition 4.24 (Parameterization Problem). *To choose a spanning tree* G_b *of* S *that maximizes* $icd(S, G_b)$*, while allowing a solution* B *to the baseline selection problem.*

Algorithm 1 find-baseline-assignment, procedure to find a baseline assignment satisfying a certain goal condition.

```
Input: Evidence structure $S$, goal condition $\varphi(\cdot)$
Output: Baseline assignment, or undefined if none exists
 1: $b \leftarrow \emptyset, p \leftarrow \emptyset$
 2: for $S_i \in S$ do
 3:     if $|T(S_i)| = 2$ then
 4:         $b \leftarrow b \cup \{(S_i, \text{some-element-of}(T(S_i)))\}$
 5:     else
 6:         $p \leftarrow p \cup \{S_i \times T(S_i)\}$
 7:     end if
 8: end for
 9: $A \leftarrow p_1 \times \cdots \times p_n \times b_1 \times \cdots \times b_m$
10: for $B \in A$ do
11:     if $\varphi(B)$ then
12:         return $B$
13:     end if
14: end for
15: return undefined
```

4.4 The algorithm

With the problem precisely defined, we developed a naive, inefficient algorithm that is sufficiently fast to solve all problem instances encountered in practice. An open source implementation is available from `http://drugis.org/mtc`.

The baseline selection sub-problem (Definition 4.23) is solved through an exhaustive search over the space of possible assignments, as is shown in Algorithm 1. Before the search, an arbitrary baseline is assigned for the two-arm studies since either baseline will cover all included comparisons (lines 3–4). For the multiple arm studies, all possible baseline choices are constructed (lines 5–6). Then, these are combined with the two arm study assignments (line 9) to construct the space A of possible baseline assignments. Then, an exhaustive search over the space A is performed (lines 10–14). As soon as a valid baseline assignement is found, the search is terminated. Note that in practice the set A is not constructed beforehand, but the space of baseline assignments is explored with e.g. a depth-first search [Cormen et al., 2001].

The algorithm to solve the full parameterization problem is described as pseudo code in Algorithm 2. We start by trying to solve the baseline selection problem for the maximally constrained case in which all edges need direct evidence. The indicator procedure required in find-baseline-assignment for checking whether all edges are covered is presented in Algorithm 3. If no solution exists, any solution G_b to the parameterization problem will have $\mathrm{icd}(S, G_b) < \mathrm{nul}(G(S))$ (see Appendix 4.8).

Then, we use the standard algorithm presented in Gabow and Myers [1978] to iterate over all spanning trees of the evidence graph $G(S)$ (Definition 4.1). For each generated tree g, we determine the IcD $\mathrm{icd}(S, g)$ (Definition 4.17). The procedure

Algorithm 2 Parameterization of a mixed treatment comparison model as finding the spanning tree that maximizes the IcD while having a valid baseline assignment.

```
Input: Evidence structure $S$
Output: Solution $(G_b, B)$ to the parameterization problem
best-$g$ $\leftarrow$ undefined, best-$b$ $\leftarrow$ undefined
$b \leftarrow$ undefined
full-$b$ $\leftarrow$ find-baseline-assignment($S, \varphi_S$)
if defined(full-$b$) then
    $k \leftarrow \mathrm{nul}(G(S))$
else
    $k \leftarrow \mathrm{nul}(G(S)) - 1$
end if
for $g \in$ gabow-myers($G(S)$) do
    if not defined(best-$g$) or icd($S, g$) $>$ icd($S$, best-$g$) then
        if defined(full-$b$) then
            $b \leftarrow$ full-$b$
        else
            $b \leftarrow$ find-baseline-assignment($S, \varphi_{S,g}$)
        end if
        if defined($b$) then
            best-$g$ $\leftarrow g$
            best-$b$ $\leftarrow b$
        end if
    end if
    if icd($S$, best-$g$) $= k$ then
        return (best-$g$, best-$b$)
    end if
end for
return (best-$g$, best-$b$)
```

for computing IcD is given in Algorithm 4. If $\mathrm{icd}(S, g)$ is greater than the largest so far, we determine whether there is a solution to the baseline selection problem for this tree. In this case the find-baseline-assignment requires an indicator procedure for checking whether the parameterization satisfies the baseline selection constraints (Definition 4.23); this one is presented in Algorithm 5. If there exists a solution to the baseline selection problem, we record this spanning tree and its IcD as the best so far. We stop if for the best tree so far $\mathrm{icd}(S, g) = k$, the maximum possible, or if all spanning trees have been enumerated. For difficult problems this will be intractable, since there may be exponentially many spanning trees, and if the evidence structure has lower than maximal IcDF all of them have to be enumerated. However, it seems that most real-world problems are easy, as is shown by the computational tests in Section 4.6.

Using an exhaustive search to identify a baseline selection solves the baseline selection problem. Since the spanning tree search also (potentially) generates all pos-

Algorithm 3 φ_S, indicator procedure for checking whether all edges are covered. $E(S_i, b_i)$ is the baseline study graph (Definition 4.21).

```
Input: Set $B$ of pairs $(S_i, b_i)$: (study, baseline)
Output: true, if all edges are covered, otherwise false
$E \leftarrow \cup_{(S_i,b_i) \in B} E(S_i, b_i)$
return $E = E(S)$
```

Algorithm 4 icd, procedure for computing the IcD. $\boldsymbol{C}/\sim_S$ is the set of equivalence classes of cycles (Definition 4.17), and $\boldsymbol{C}(\cdot,\cdot)$ is the set of fundamental cycles (Appendix 4.8).

```
Input: Evidence structure $S$, Spanning tree $g$
Output: $\mathrm{icd}(S, g)$
$\boldsymbol{C} \leftarrow \boldsymbol{C}(G(S), g)$
$\mathrm{icd} \leftarrow 0$
for $X \in (\boldsymbol{C}/\sim_S)$ do
    $\mathrm{icd} \leftarrow \mathrm{icd} + \mathrm{icc}(\text{some-element-of}(X))$
end for
return icd
```

sible spanning trees and maximizes the IcD taking into account whether there is a solution to the baseline selection problem, the algorithm outlined here solves the parameterization problem (Definition 4.24).

4.5 Example

As a full example of finding a correct parameterization for an evidence structure, we consider a network of treatments for smoking cessation therapy comparing (a) nicotine replacement therapy, (b) bupropion, (c) varenicline and (d) placebo or no treatment [Wu et al., 2006]. The outcome of interest is smoking cessation at 12 months. For this outcome there are 78 studies with 4 different treatment comparisons: S_{ad} (66 studies), S_{bd} (6 studies), S_{abd} (3 studies) and S_{bcd} (3 studies). The evidence structure is

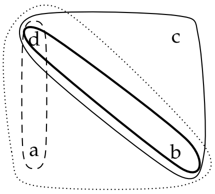


Figure 4.6: The evidence structure for the outcome 'smoking cessation after 12 months' from Wu et al. [2006]. a = nicotine replacement therapy, b = bupropion, c = varenicline, and d = placebo or no treatment.

Algorithm 5 $\varphi_{S,g}$, indicator procedure for checking whether the parameterization satisfies the baseline selection constraints (Definition 4.23). $E(S_i, b_i)$ is the baseline study graph (Definition 4.21), $\mathbf{C}/\sim_S$ is the set of equivalence classes of cycles (Definition 4.17), and $\boldsymbol{C}(\cdot,\cdot)$ is the set of fundamental cycles (Appendix 4.8).

```
Input: Set $B$ of pairs $(S_i, b_i)$: (study, baseline)
Output: true, if Definition 4.23 is satisfied, otherwise false
$E \leftarrow \cup_{(S_i,b_i) \in B} E(S_i, b_i)$
$\boldsymbol{C} \leftarrow \boldsymbol{C}(G(S), g)$
for $X \in (\boldsymbol{C}/\sim_S)$ do
    for $C \in X$ do
        if $|C \cap E| < |C| - 1$ then
            return false
        end if
    end for
    if icc(some-element-of($X$)) $= 0$ then
        return true
    end if
    for $C \in X$ do
        if $|C \cap E| = |C|$ then
            return true
        end if
    end for
end for
return false
```

shown in Figure 4.6, having $|T(S)| = 4$ treatments and $|E(S)| = 5$ comparisons. Thus, any correct parameterization will have $|E_b| = 4 - 1 = 3$ basic and $|E_f| = 5 - 3 = 2$ functional parameters.

The first step in the algorithm is to try and find a baseline selection that covers all edges. Given this structure, that is easy, e.g. a for the S_{ad} studies, b for S_{bd}, a for S_{abd} and c for S_{bcd} will suffice. We could also have chosen different baselines for studies of the same type, but that is not necessary here. Then, we set $k = |E_f| = 2$, meaning we will try to find an IcD equal to the number of functional parameters.

Now, we start iterating over the spanning trees of the evidence graph. Say the first spanning tree we are given is

$$g_1 = \{\{b, a\}, \{b, c\}, \{b, d\}\} \ .$$

Then the cycles to evaluate are *badb* and *bcdb*. For *badb* we get the partition

$$P = \{(b, a), (a, d), (d, b)\} \ ;$$

$$r((b, a)) = S_{abd},\ r((a, d)) = S_{ad} \cup S_{abd},\ r((d, b)) = S_{bd} \cup S_{bcd} \ .$$

This partition cannot be reduced any further, and there are 3 distinct sets of studies, so

$$\mathrm{icc}(S, badb) = 1 \ .$$

For $bcdb$, both (b, c) and (c, d) are supported only by S_{bcd}, and thus we have only 2 sets of supporting studies, so

$$\text{icc}(S, bcdb) = 0 \ .$$

The cycles are not equivalent, so they fall into two separate classes, and we get $\text{icd}(S, g_1) = 1$. This is the best so far, so we store g_1.

The IcD of g_1 is $1 < k = 2$, so we continue iterating over the spanning trees. The second spanning tree might be:

$$g_2 = \{\{b, a\}, \{a, d\}, \{d, c\}\} \ ,$$

having fundamental cycles $bdab$ and $bcdab$. We already know that $\text{icc}(S, bdab) = 1$, and we also recognize that $bcdab$ is a basically a longer version of $bdab$, so we will have to evaluate whether they are equivalent. We also recall that (d, c) and (c, b) are only supported by S_{bcd} and hence reduce to (d, b) with $r'((d, b)) = S_{bcd}$. The other two comparisons, (b, a) and (a, d) cannot be reduced. Thus we get a reduction for $bcdab$ that has the same comparisons as $bdab$, but a different set of supporting studies for (b, d): S_{bcd} for $bcdab$ and $S_{bd} \cup S_{bcd}$ for $bdab$. Hence $bdab$ and $bcdab$ are not equivalent in S: $bdab \not\sim_S bcdab$. Moreover, we also get

$$\text{icc}(S, bcdab) = 1 \ ,$$

so that $\text{icd}(S, g_2) = 2 = k$. Hence we have identified g_2 with the full baseline assignment identified earlier as the solution to the parameterization problem.

The example structure is structure number 18 in Table 4.1, and our implementation of the algorithm actually evaluates four spanning trees before it finds the correct one, rather than the two shown here.

4.6 Evaluation of the running-time

A review of published evidence networks [Salanti et al., 2008b] identified 18 different networks in the literature. Three of those were star-shaped, and have a trivial solution to the parameterization problem. For the other 15 networks, we extracted the evidence structure from the original papers and evaluated the running time of our algorithm, as well as the IcDF of each structure and the number of spanning trees that were generated before a solution was found. The results are summarized in Table 4.1, and we give the exact evidence structures in an online supplement. There are 3 structures with only two-arm trials, the remaining 12 have at least one three-arm trial. There is one structure that includes a four-arm trial.

All of the evidence structures were parameterized within 4 seconds (on a 3GHz processor), which is negligible compared to the time usually taken by the MCMC simulation used to estimate the models. The longest time taken was on structure 23, which contains the largest number (6) of distinct types of three-arm trials. Only 3 structures had non-maximal IcDF (23, 26, 28), namely $s - 1$, one less than the number of functional parameters. Note that if the IcDF would be $< s - 1$, our algorithm would need to enumerate all spanning trees to terminate. In only two cases more than one

	Structure							Result		
ref	v	e	s	n	x_2	x_3	x_4	i	N	t
18	4	5	2	78	2	2	0	2	4	0.3
19	4	5	2	34	5	0	0	2	1	0.2
20	4	4	1	12	3	1	0	1	1	0.1
21	4	4	1	10	3	1	0	1	1	0.1
22	4	5	2	21	5	0	0	2	1	0.1
23	10	21	12	43	16	6	0	11	1	3.8
24	9	14	6	54	12	3	0	6	1	0.5
25	7	16	10	22	13	4	0	10	1	0.5
26	9	16	8	18	10	3	1	7	1	1.0
3	6	9	4	14	7	1	0	4	13	0.4
27	7	8	2	25	8	0	0	2	1	0.2
28	16	22	7	34	15	4	0	6	1	0.9
29	7	8	2	10	8	1	0	2	1	0.1
30	8	10	3	14	9	1	0	3	1	0.2
31	10	12	3	14	10	2	0	3	1	0.2

Table 4.1: Performance of our algorithm on evidence structures from Salanti et al. [2008b]. Structures are listed in the same order as Figure 1 in Salanti et al. [2008b], omitting the first three; the first column (ref) gives the reference number in that paper. v is the number of included treatments, e is the number of comparisons, s is the number of functional parameters and n is the number of studies. The x_j indicate the number of different *types* of j-arm studies, e.g. if we have 2 ab studies and 3 bc studies, $x_2 = 2$. For the results, i is the IcD of the solution, N is the number of evaluated spanning trees and t is the time taken (seconds).

spanning tree needed to be explored. The number of distinct spanning trees the evidence graph had varied between three (structures 21 and 22) to 13611 (structure 23). All running times of > 0.5 seconds were observed for structures with non-maximal IcDF. In all three cases, this reflects a failed exhaustive baseline search for full evidence cover.

4.7 Discussion

In this paper, we defined the parameterization problem for MTC evidence structures and we provided an algorithm which can be used for automated model generation for MTC. We refine previous work [Lu and Ades, 2006] on identifying the IcDF by giving a precise problem definition, and point out the additional problem of equivalent cycles. An open source implementation of the algorithm is available (`http://drugis.org/mtc`). Although the worst-case complexity of our algorithm is exponential, it seems that real-world problems can be solved quickly. We evaluated running time of the algorithm with evidence structures from the literature, and all were solved within 4 seconds on a standard PC.

Future work should aim to develop more efficient algorithms, and further investigate the relationship between the spanning tree and baseline selection problems. In this paper, we took the pragmatic approach of defining the combined problem as finding the maximal IcD for which a baseline selection can be derived. The question remains whether there may be evidence structures for which the optimal spanning tree does not have an associated baseline assignment, and what would be the implications for the MTC method. There also seems to be a certain redundancy in the cycles *bdab* and *bcdab* of Figure 4.6 discussed in Section 4.5, even though they are not equivalent according to Definition 4.16. This is correct since the w-factors associated with these cycles are not provably equal. However, the w-factors should differ only by heterogeneity on the (b, d) comparison. Future work should address whether and, if so, how this should be incorporated in the parameterization of the evidence structure.

4.8 Appendix: definitions from graph theory

Definition 4.25 (spanning tree). – *Gabow and Myers [1978]*
In a connected, undirected graph G, a spanning tree *G_s is a subgraph having a unique simple path (a path containing each vertex at most once) between any two vertices of G. If G has t vertices, G_s has $\dim(G) = t - 1$ edges.*

Definition 4.26 (fundamental cycle set). – *Deo et al. [1982]*
The fundamental cycle set of a connected, undirected graph $G = (T, E)$ with respect to a spanning tree $G_s = (T, F)$ is generated from the set $E' = E \backslash F$, as follows:

$$\mathbf{C}(G, G_s) = \{C(G, G_s, e) \mid e \in E'\}\ \text{, with}$$

$$C(G, G_s, \{u, v\}) = \text{path}(v, u) \cup \{\{u, v\}\}\ \ ,$$

where $\mathrm{path}(v, u)$ *gives the (unique simple) path from* v *to* u *in* G_s*. The size of the set of non-tree edges,* $\mathrm{nul}(G) = |E'| = |E| - |T| + 1$ *is called the* nullity *of* G*, and determines the number of* independent *cycles in* G*. The set* $\mathbf{C}(G, G_s)$ *consists of independent cycles and* $|\mathbf{C}(G, G_s)| = \mathrm{nul}(G)$*, which means that the set of fundamental cycles is also a* cycle basis *of* G*.*

Definition 4.27 (path)**.** *A path is a sequence of directed edges, such that the target of each edge connects to the source of the next one:* $p = ((w_1, w_2), (w_2, w_3), \ldots, (w_{n-1}, w_n))$ *is a path of length* $n - 1$*, as counted by the number of edges. Often, the path* p *is conveniently written as* $(w_1, w_2, \ldots w_n)$*, which should be read as shorthand for the longer notation.*

CHAPTER 5

A stochastic multi-criteria model for evidence-based decision making in drug benefit-risk analysis

T. Tervonen, G. van Valkenhoef, E. Buskens, H. L. Hillege, and D. Postmus. A stochastic multi-criteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428, 2011. doi: 10.1002/sim.4194

Abstract

Drug benefit-risk analysis is based on firm clinical evidence regarding various safety and efficacy outcomes. In this paper, we propose a new and more formal approach for constructing a supporting multi-criteria model that fully takes into account the evidence on efficacy and adverse drug reactions. Our approach is based on the Stochastic Multi-criteria Acceptability Analysis (SMAA) methodology, which allows us to compute the typical value judgments that support a decision, to quantify decision uncertainty, and to compute a comprehensive benefit-risk profile. We construct a multi-criteria model for the therapeutic group of second-generation antidepressants. We assess fluoxetine and venlafaxine together with placebo according to incidences of treatment response and three common adverse drug reactions by using data from a published study. Our model shows that there are clear trade-offs among the treatment alternatives.

5.1 Introduction

Drug Benefit-Risk (BR) analysis is daily business for health care professionals. Health authorities, prescribing physicians, pharmacists, reimbursement policy makers, and employees of insurance companies all more or less explicitly evaluate the safety and efficacy of different medicinal compounds. Although the exact scope of the analyses conducted by these evaluators is different (e.g. in clinical practice the decision concerns an individual patient, whereas in policy making the general population or a subset of the population that has some particular characteristics is considered), they all must examine and weight the clinical evidence regarding the magnitudes of benefit and risk, taking into account the quality and precision with which these magnitudes have been estimated.

The benefit/risk ratio (which is calculated from the difference in risk and difference in benefit between therapies) has been proposed as a simple aggregate measure of the BR trade-off for a single efficacy criterion and a single risk criterion. Although such a measure is easy to interpret and implement in clinical practice, drug BR analysis typically includes multiple benefit and risk criteria and consequently must include value judgments [McGregor and Caro, 2006, Claycamp, 2006, Garrison, Jr. et al., 2007]. In such a setting, the use of Multi-Criteria Decision Analysis (MCDA) is more appropriate as it provides a framework for systematic and replicable analyses of complex decision problems involving value trade-offs.

The use of MCDA in the context of drug BR analysis was first proposed by Mussen et al. [Mussen et al., 2007]. Their work includes a general framework for constructing a multi-criteria decision model for BR assessment of new drugs by regulatory authorities. Although it is an important seminal work in the field, they score alternative drugs on the different benefit and risk criteria solely based on point estimates. Thus, uncertainty associated with sampling variation inherent to criteria measurements obtained in experimental or observational studies is ignored. In addition, the approach suggested by Mussen et al. [Mussen et al., 2007] requires Decision Makers (DMs) to provide exact weights for describing the relative importance of the different criteria. Although detailed weight elicitation during model construction can help the DMs to understand the problem, in many real-life situations DMs are not able to (or do not want to) give exact preference information. Also, a group of DMs may not reach a consensus about the weights [Tervonen and Figueira, 2008]. Felli et al. [Felli et al., 2009] provided a similar application of MCDA in drug BR analysis. Instead of using continuous measurements, they proposed to use categorical value scales for all BR attributes included in the model. Although it makes the model easier to apply in different contexts, there is a substantial risk of losing information by mapping measurements from a continuous scale to ordinal categories.

To overcome the limitations of the two previous approaches, we propose to use Stochastic Multi-criteria Acceptability Analysis (SMAA) [Lahdelma et al., 1998, Lahdelma and Salminen, 2001, Tervonen and Figueira, 2008] as a new and more elaborate approach to drug BR analysis. Our choice of the SMAA methodology is supported by its proven applicability in risk assessment [Tervonen et al., 2009b,a] and reported real-life analyses [Hokkanen et al., 1999, Tervonen et al., 2008, Kangas et al., 2006, 2003, Hokkanen et al., 1998, Kangas and Kangas, 2003, Lahdelma et al., 2002] alike. To demonstrate its applicability in drug BR analysis, we will apply the SMAA-2 method [Lahdelma and Salminen, 2001] to evaluate the potential benefits and risks of two commonly prescribed second-generation antidepressants in the setting of a published placebo-controlled trial [Nemeroff and Thase, 2007].

5.2 Stochastic Multi-criteria Acceptability Analysis

SMAA-2 [Lahdelma and Salminen, 2001] considers a discrete, multi-criteria decision problem consisting of a set of m alternatives that are evaluated in terms of n criteria. The vector of criteria values corresponding to alternative i is denoted by $\mathbf{C}^i = (C_1^i, \ldots, C_n^i)$, where C_k^i represents the performance of alternative i on criterion k. Instead of using deterministic values, the criteria values are assumed to be random variables with joint density function $f_{\mathbf{C}^i}(\mathbf{c}^i)$ in the evaluation space $X \subseteq R^n$.

It is assumed that the DM's preferences for any point $\mathbf{c} \in X$ can be represented by the real-valued value function $u(\mathbf{c}, \cdot)$. Although SMAA-2 can be applied with any type of value function, it is generally assumed that the criteria satisfy the independence conditions [Keeney and Raiffa, 1976] for applying the additive value function:

$$u(\mathbf{c}, \mathbf{w}) = w_1 \cdot u_1(c_1) + \cdots + w_n \cdot u_n(c_n).$$

The additive value function is normalized by $u(\mathbf{c}^{'}, \mathbf{w}) = 0$ and $u(\mathbf{c}^{''}, \mathbf{w}) = 1$ for arbitrarily chosen $\mathbf{c}^{'}, \mathbf{c}^{''} \in X$, such that $\left(c^{''}_{\{k\}}, \mathbf{c}^{'}_{\{\bar{k}\}}\right) \succ \left(c^{'}_{\{k\}}, \mathbf{c}^{'}_{\{\bar{k}\}}\right)$, $\forall k \in \{1, \ldots, n\}$. The symbol $\succ$ denotes the strict preference relation, and $(c_Y, c_{\bar{Y}})$ refers to the partition of c according to a subset Y of the criteria and its complement $\bar{Y}$. For example, if $n = 5$ and $Y = \{1, 3, 5\}$, $c_Y = (c_1, c_3, c_5)$ and $c_{\bar{Y}} = (c_2, c_4)$. The weights w_k, normalized so that they sum to one, rescale the values of the partial value functions, normalized by $u_k(c^{'}_k) = 0$ and $u_k(c^{''}_k) = 1$, in such a way that a unit increase in the scaled function (i.e. the swing from $c^{'}_k$ to $c^{''}_k$) indicates the importance of the criterion [von Nitzch and Weber, 1993]. For example, $w_s > w_t$ implies that if the DM is currently at $\mathbf{c}^{'}$ and could choose between moving to $\left(c^{''}_{\{s\}}, \mathbf{c}^{'}_{\{\bar{s}\}}\right)$ or $\left(c^{''}_{\{t\}}, \mathbf{c}^{'}_{\{\bar{t}\}}\right)$, he or she would rather move to $\left(c^{''}_{\{s\}}, \mathbf{c}^{'}_{\{\bar{s}\}}\right)$.

Instead of using the value function to rank the alternatives for an elicited weight vector $\mathbf{w}$, which is the traditional approach in multi-attribute value theory, the SMAA methodology has been developed for situations where the weights are random variables with a joint density function $f_{\mathbf{W}}(\mathbf{w})$ in the feasible weight space

$$\Omega = \left\{ \mathbf{w} \in R^n : \mathbf{w} \geq 0 \text{ and } \sum_{j=1}^{n} w_j = 1 \right\}.$$

Total lack of preference information is represented by a uniform weight distribution in Ω, i.e., $f_{\mathbf{W}}(\mathbf{w}) = 1/\text{vol}(\Omega)$. In practice, it may be possible to elicit some preference information from the DM, such as a partial or complete ranking of the criteria.

This information can easily be incorporated into the model by restricting the feasible weight space accordingly [Tervonen and Lahdelma, 2007].

Define $\Xi = \left(\mathbf{C}^1, \ldots, \mathbf{C}^m\right)$, and let $f_\Xi(\xi)$ denote the joint density function of Ξ. For given realizations ξ of Ξ and $\mathbf{w}$ of $\mathbf{W}$, the rank of each alternative is defined as an integer from the best rank (= 1) to the worst rank (= m) by means of a ranking function

$$rank(i, \xi, \mathbf{w}) = 1 + \sum_{k=1}^{m} \rho(u(\mathbf{c}^k, \mathbf{w}) > u(\mathbf{c}^i, \mathbf{w})),$$

where $\rho(true) = 1$ and $\rho(false) = 0$. SMAA-2 is then based on analyzing the stochastic sets of favorable rank weights

$$\Omega_i^r(\xi) = \{\mathbf{w} \in \Omega : rank(i, \xi, \mathbf{w}) = r\}.$$

Any weight $\mathbf{w} \in \Omega_i^r(\xi)$ results in such values for the different alternatives that alternative i obtains rank r.

The main decision aiding measure in SMAA-2 is the *rank acceptability index*, denoted by b_i^r. It describes the share of all possible values of the weight vector $\mathbf{W}$ and the joint random vector Ξ for which alternative i is ranked at place r. Its value can be interpreted as the probability that alternative i is ranked at place r, where 0 indicates that the alternative will never obtain rank r and 1 indicates that alternative i will always obtain rank r. The rank acceptability index b_i^r is computed numerically as a multidimensional integral over the criteria distributions and the favorable rank weights as

$$b_i^r = \int_{\xi \in \Xi} f_\Xi(\xi) \int_{\mathbf{w} \in \Omega_i^r(\xi)} f_{\mathbf{W}}(\mathbf{w})\, d\mathbf{w}\, d\xi.$$

The preferred (best) alternatives are those with high acceptabilities for the best ranks.

In addition to the rank acceptability indices, the SMAA methods allow to describe the typical preferences of a DM supporting each efficient alternative (i.e. all alternatives with a non-zero first rank acceptability index). These so-called *central weight vectors* can be presented to the DM to help him or her understand what kind of weights would favor a certain alternative, without providing factual preference information. The central weight vector of an alternative is defined as the expected center of gravity of all possible weight vectors that rank the alternative at the first place. It is computed numerically as a multidimensional integral over the criteria distributions and the favorable first rank weights using

$$\mathbf{w}_i^c = \int_{\xi \in \Xi} f_\Xi(\xi) \int_{\mathbf{w} \in \Omega_i^1(\xi)} f_{\mathbf{W}}(\mathbf{w}) \mathbf{w}\, d\mathbf{w}\, d\xi / b_i^1.$$

The *confidence factor* p_i^c is the probability for an alternative to obtain the first rank when the central weight vector is chosen. The confidence factor is computed as a multidimensional integral over the criteria distributions using

$$p_i^c = \int_{\xi \in \Xi : rank(i, \xi, \mathbf{w}_i^c) = 1} f_\Xi(\xi)\, d\xi.$$

Name	Preference direction	$c_k^{'}$	$c_k^{''}$
Efficacy	↑	0.28	0.63
Nausea ADRs	↓	0.50	0.04
Insomnia ADRs	↓	0.31	0.08
Anxiety ADRs	↓	0.17	0.00

Table 5.1: Criteria, preference directions, and scaling vectors. All criteria are measured as incidences

Confidence factors can similarly be calculated for any given weight vector. The confidence factors indicate whether the criteria values are sufficiently accurate to discern the efficient alternatives. Alternatives with low first rank acceptability indices and low confidence factors for their central weight vectors are unlikely to be considered the most preferred one by any DM. In contrast, a very high confidence factor indicates that if a DM finds his or her preferences to correspond to an alternative's central weight vector, the alternative is almost certainly the one with highest preference [Lahdelma and Salminen, 2006a]. Central weights of alternatives with low confidence factors (< 0.50) should be interpreted with care, as even when a DM finds his central weight vector to correspond with his preferences, there might be other alternatives that achieve higher first rank acceptability with those weights.

If there is no preference information, the decision making is aided mainly through central weight vectors and confidence factors. When preference information is incorporated, the rank acceptability indices can be used to find the "best" alternative and to quantify the risks related to uncertainties surrounding outcomes.

5.3 A multi-criteria model for the therapeutic group of antidepressants

To demonstrate the applicability of SMAA in drug BR analysis, we constructed a model for the therapeutic group of antidepressants using efficacy and safety data from a published study [Nemeroff and Thase, 2007]. If patients are not harmed by deferral of therapy, it is important to have a non-active control included in the analysis to put the relative performances of the different active compounds into context with what is seen without a treatment [Temple and Ellenberg, 2000]. For depressive disorder, there is no evidence that treatment delay or assignment to placebo results in permanent harm [Walsh et al., 2002]. Placebo was therefore explicitly included as one of the alternatives in the constructed BR model.

5.3.1 Criteria

The original placebo-controlled trial compared efficacy and safety of venlafaxine and fluoxetine [Nemeroff and Thase, 2007]. From this study, we selected treatment response, defined as an improvement from baseline of at least 50% on the Hamilton Depression Rating Scale (HAM-D), as our benefit criterion. To obtain our risk crite-

Criterion	Placebo	Fluoxetine	Venlafaxine
Efficacy	37/101	45/100	51/96
		RD 0.08 (-0.05, 0.22)	RD 0.16 (0.03, 0.30)
Nausea ADRs	8/102	22/102	40/100
		RD 0.14 (0.04, 0.23)	RD 0.32 (0.21, 0.43)
Insomnia ADRs	14/102	15/102	22/100
		RD 0.01 (-0.09, 0.11)	RD 0.08 (-0.02, 0.19)
Anxiety ADRs	1/102	7/102	10/100
		RD 0.06 (0.01, 0.11)	RD 0.09 (0.03, 0.15)

Table 5.2: Incidence rates of HAM-D responders and three ADRs as reported in the original study [Nemeroff and Thase, 2007], with their risk differences (RD, given as mean and 95% confidence interval) versus Placebo (calculated by the authors based on the original data)

Criterion	Venlafaxine	Fluoxetine	Placebo
Efficacy	$52, 46$	$46, 56$	$38, 65$
Nausea ADRs	$41, 61$	$23, 81$	$9, 95$
Insomnia ADRs	$23, 79$	$16, 88$	$15, 89$
Anxiety ADRs	$11, 91$	$8, 96$	$2, 102$

Table 5.3: Beta distributions of the criteria values (parameters are given as a_k^i, b_k^i)

ria, we asked an expert in the field of antidepressants to select three Adverse Drug Reactions (ADRs) that she considered to be most relevant from a drug safety perspective. The resulting criteria for evaluating the two drugs and placebo are summarized in Table 5.1, and the data reported in the original study are shown in Table 5.2. There is a certain overlap between efficacy and insomnia, because improved efficacy can lead to less insomnia. For sake of simplicity, we disregarded this possible source of double-counting and assumed the criteria to be independent.

5.3.2 Probability distributions of the criteria values

The observed incidences $\frac{r_k^i}{n_k^i}$ of treatment response and ADRs can considered to be realizations from binomially distributed variables with success probability C_k^i (i.e. $r_k^i \sim \mathrm{Bin}(n_k^i, C_k^i)$). Assuming independence of the $m \cdot n$ success probabilities, we modeled $C_k^i \sim \mathrm{Beta}(a_k^i, b_k^i)$. Following a Bayesian approach with a flat $\mathrm{Beta}(1,1)$ prior, the Beta parameters a_k^i and b_k^i were set equal to $r_k^i + 1$ and $n_k^i - r_k^i + 1$, respectively. The resulting parameter values are summarized in Table 5.3.

5.3.3 Partial value functions

It is generally helpful to limit the region $Z \subseteq X$ over which preferences must be assessed to as small a region as possible, taking into account the observed ranges in the criteria values [Keeney and Raiffa, 1976]. Our approach was therefore to bound Z

Drug	b_i^1	b_i^2	b_i^3
Venlafaxine	0.08	0.14	0.78
Fluoxetine	0.17	0.71	0.12
Placebo	0.75	0.16	0.09

Table 5.4: Rank acceptability indices from the analysis without preference information

by the interval hulls (the interval hull of k intervals is defined as the smallest possible interval that contains all of these k intervals) of the 95% probability intervals of the m success probabilities associated with each of the criteria. This ensures that even if the underlying independence assumptions are not valid for the complete range of theoretically achievable values in X, the additive value function will still be a good approximation for the subset Z in which the criteria values are most likely to fall. The two points $c^{'}$ and $c^{''}$ required to scale the (partial) value function(s) were set equal to the least and most preferable values in Z, respectively, and are listed in Table 5.1. The partial value functions $u_k(c_k)$ were assumed to be linear, meaning that they were defined as $u_k(c_k) = \frac{c_k - c_k^{'}}{c_k^{''} - c_k^{'}}$ if the preference direction is increasing and $u_k(c_k) = \frac{c_k^{'} - c_k}{c_k^{'} - c_k^{''}}$ if the preference direction is decreasing. For example, for nausea $c_{\mathrm{nau}}^{'} = 0.50$, $c_{\mathrm{nau}}^{''} = 0.04$, and the preference direction is decreasing, so $u_{\mathrm{nau}}(c_{\mathrm{nau}}) = \frac{0.50 - c_{nau}}{0.46}$.

5.3.4 Preference information

We performed three analyses: one without preference information, and two scenarios with criteria rankings elicited from an expert in the field of antidepressants. For the scenario-based analyses, we considered a scenario of mild depression and a scenario of severe depression. For both scenarios, we asked the expert to identify the criterion that she considered to be most important, i.e. would foremost increase from the worst to the best value, given the range of the scales as depicted in Table 5.1. Then we asked for the second one, etc. This process is similar to swing weighting in multi-attribute value theory [Belton and Stewart, 2002]. However, since no exact weights are elicited, it requires less effort from the DM.

Let us denote by $\succ$ the strict preference relation for unit increases in the partial value functions of the criteria. The elicitation process resulted in the following ranking for mild depression: Nausea $\succ$ Anxiety $\succ$ Efficacy $\succ$ Insomnia. For severe depression the ranking was similar with the exception of efficacy being the most preferred criterion (i.e. Efficacy $\succ$ Nausea $\succ$ Anxiety $\succ$ Insomnia).

5.3.5 Analyses

The three analyses were conducted using the open source JSMAA software [Tervonen, 2010] v0.8 for Monte Carlo estimation of SMAA models. All analyses were executed with 10,000 Monte Carlo iterations, thereby giving the results sufficient accuracy (95% confidence error margins of ±0.01) [Tervonen and Lahdelma, 2007].

The rank acceptability indices resulting from the analysis without preference information are listed in Table 5.4 and visualized as a column chart in Figure 5.1. These

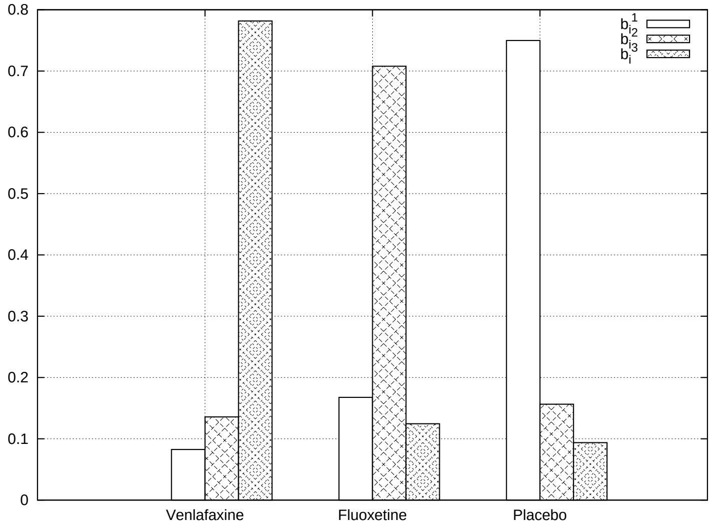


Figure 5.1: Rank acceptability indices for the model without preference information

		w_i^c			
Drug	p_i^c	Efficacy	Nausea	Insomnia	Anxiety
Venlafaxine	0.48	0.58	0.11	0.15	0.15
Fluoxetine	0.35	0.37	0.16	0.30	0.17
Placebo	0.96	0.18	0.28	0.25	0.29

Table 5.5: Central weights and corresponding confidence factors from the analysis without preference information

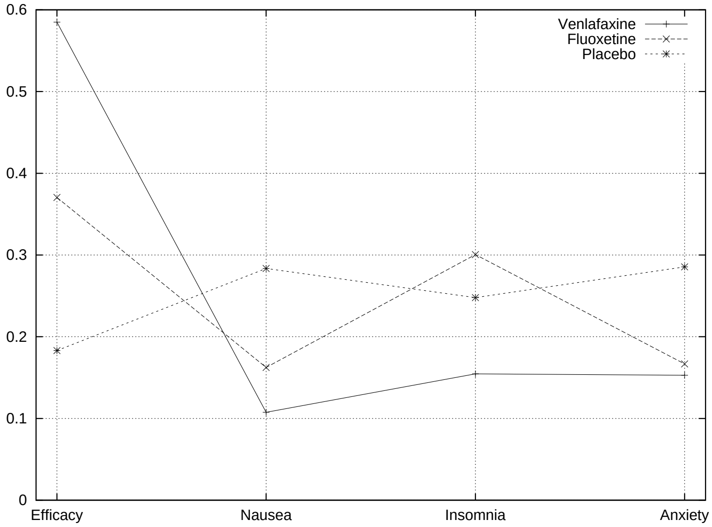


Figure 5.2: Central weight vectors for the model without preference information

indices show that each of the drugs is the preferred one given some preferences. Thus, all of them should be considered for further analysis. In a situation like this, the decision can be aided through the central weight vectors (see Table 5.5 and Figure 5.2). By looking at the central weights, we can see clear trade-offs among the three alternatives. For example, if the DM displays an a priori preference for venlafaxine, then based on the BR profiles expressed through the central weights, apparently efficacy has the highest relative importance. If the DM accepts the independence conditions underlying the additive model, he or she should find increasing efficacy from the worst scale value (0.28) to the best one (0.63) more important than improving any of the ADR criteria from their worst to best scale values.

By contrasting a DM's preferences for scale swings (Table 5.1) with the central weights presented in Table 5.5, the DM can quickly decide which drug is preferable in the current situation. For example, if the only preference information available is that the DM considers the scale swing of anxiety (0.17 to 0.00) less important than the scale swing of insomnia (0.31 to 0.08, see Table 5.1), then he or she should prefer fluoxetine as it is the only alternative for which the central weight of anxiety is considerably lower than the central weight of insomnia. In addition, the confidence factors (Table 5.5) quantify the risk associated with the decision. For example, if a DM finds fluoxetine's central weight vector to correspond with his or her preferences, the confidence factor (0.35) shows that the clinical data is too uncertain to make a truly informed decision.

Rank acceptability indices from the scenario of mild (severe) depression are presented in Table 5.6 (Table 5.7) and illustrated in Figure 5.3 (Figure 5.4). Placebo obtains a very high first rank acceptability for the scenario of mild depression, and it obtains

Drug	b_i^1	b_i^2	b_i^3
Venlafaxine	0.00	0.02	0.97
Fluoxetine	0.01	0.96	0.02
Placebo	0.99	0.01	0.00

Table 5.6: Rank acceptability indices for the scenario of mild depression

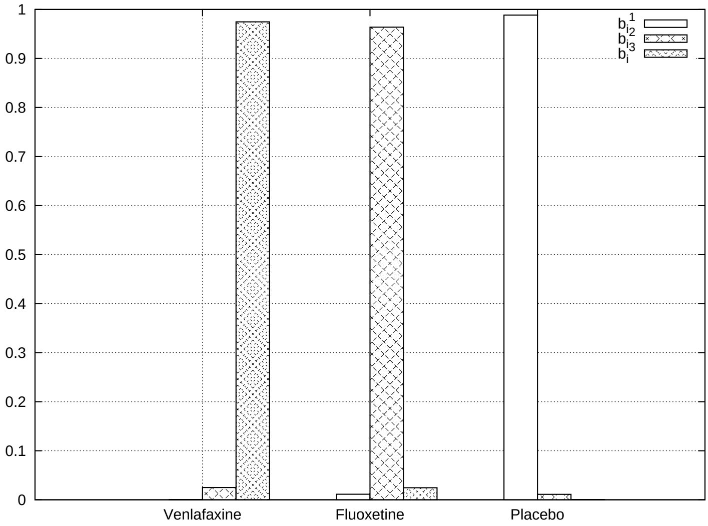


Figure 5.3: Rank acceptability indices from the scenario of mild depression

Drug	b_i^1	b_i^2	b_i^3
Venlafaxine	0.25	0.25	0.50
Fluoxetine	0.29	0.48	0.23
Placebo	0.46	0.28	0.27

Table 5.7: Rank acceptability indices for the scenario of severe depression

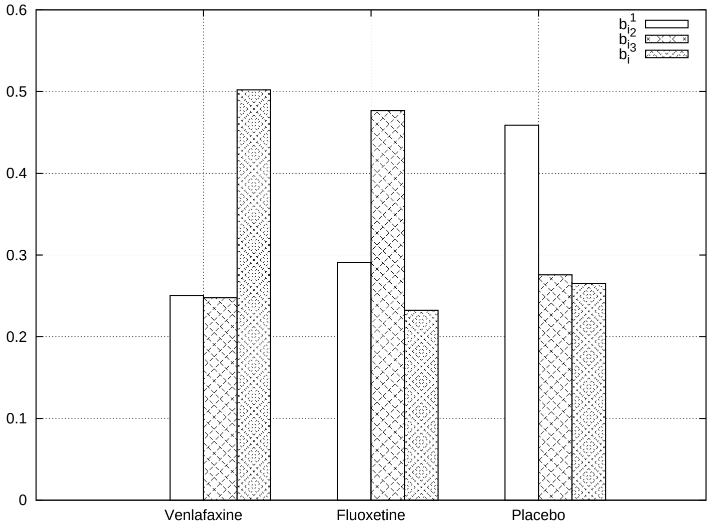


Figure 5.4: Rank acceptability indices from the scenario of severe depression

a reasonable rank profile for the scenario of severe depression. The rank profiles of fluoxetine and venlafaxine, in contrast, are very sensitive to the preferences as both of them obtain extremely low (≤ 0.01) first rank acceptabilities for the scenario of mild depression, but reasonable ones (0.25 and 0.29) for the scenario of severe depression.

5.4 Discussion

Drug BR analysis has multiple uses, ranging from regulatory decision making to supporting decisions of a practicing physician. The MCDA-based approach suggested in this paper can be adapted for most contexts. We constructed a stochastic multi-attribute model for the therapeutic group of antidepressants by using data from a published placebo-controlled trial. Despite the fact that few of the differences among the three alternatives were significant from a frequentist perspective, our results show that there are clear trade-offs among the two active compounds and placebo when the uncertainty regarding the criteria measurements is taken into account. This can be seen from the central weight vectors of the analysis without preference information, and from the rank acceptability indices for the scenarios of mild and severe depression that differed only in the preference rank of the efficacy criterion relative to the risk criteria.

Compared to the MCDA-based approaches proposed by Mussen et al. [Mussen et al., 2007] and Felli at al. [Felli et al., 2009], the use of SMAA has two main advantages. The first advantage of the SMAA methodology is the possibility to include the sampling variation that is inherent in criteria measurements that are based on clinical trials. Ignoring the uncertainty surrounding the criteria values, as is done by Mussen et al. [Mussen et al., 2007] and Felli et al. [Felli et al., 2009], makes it difficult to assess how much the different drugs differ on the selected criteria. For example, a systematic review of 10 second-generation antidepressants [Hansen et al., 2005] concluded that when looking at the point estimates and the corresponding 95% confidence intervals, the drugs "probably do not differ substantially for the treatment of major depressive disorder" and that choosing the most appropriate treatment is therefore difficult. However, a more recent review [Cipriani et al., 2009] was able to provide more concrete results through a network meta-analysis, a Bayesian approach to evidence synthesis that fully takes into account the uncertainty in the effect estimates. In addition to performing all possible comparisons, the authors provided rank probability plots that clearly showed that some drugs are "better" than others on specific criteria. Unfortunately, rank probability plots for individual criteria provide little guidance when more than two criteria are considered. As our results have shown, applying the SMAA method enables one to clearly assess the existing trade-offs. In addition, the ability of our approach to propagate uncertainty to the results (in terms of rank acceptability indices and confidence factors) allows one to quantify the risks associated with any decision that is based on the results of the BR analysis.

The second advantage of our approach over the two existing ones is the possibility to characterize typical trade-offs supporting a drug BR profile without knowing or eliciting the (exact numerical) preferences beforehand. The possibility to use our model without any preferences as well as with scenario-based ordinal preferences lowers the effort required to apply the model in different situations, and also increases the transparency of the decision making process. An analysis without preference information is useful when it is not feasible to elicit preferences or when the potential merits of a drug have to be assessed across a wide range of preferences. This latter situation occurs, for example, in policy decision making, where the policy maker's decision affects the complete target population. The central weight vectors could then be used to see whether there are likely scenarios (in terms of criteria weights) that will lead to the selection of a certain drug. The selection scenario could be, for example, a prescription decision, and the actual decision being aided is whether the drug should be granted a marketing authorization. The scenario-based rank acceptability indices can be used in operational support of decisions depending on drug BR analysis. For example, if the BR analysis is used for aiding a prescription decision for a patient with severe depression, our results show that both venlafaxine and fluoxetine are viable choices because of their relatively high acceptabilities for the best ranks. Also, if due to external factors (local reimbursement policy, patient profile including allergies, etc) a drug with a low first rank acceptability is prescribed (such as either of the active compounds in case of mild depression), the prescriber should be sensitive to changes in the external environment as drugs with "better" BR profiles may have become available.

Although our example of the second-generation antidepressants clearly demonstrated the usefulness of the proposed approach, in some decision making contexts other approaches might be more appropriate. When the DMs have the time and motivation to engage in decision conferencing [Phillips, 2007], a traditional multi-attribute value/utility theory approach can be more suitable. However, although such a facilitated environment might help the DMs to explore the problem in more detail, it can also introduce additional bias as the preference elicitation is heavily guided by the facilitator. In any case, we acknowledge that social aspects play an important role in group decision making, and future research should explore the applicability of our model in real-life pharmaceutical group decision making contexts, such as policy decision making.

Instead of having a different model for each therapeutic group, one could also consider constructing a more generic model by using the dimensions of an existing utility instrument, such as the EQ-5D or the Health Utilities Index. Although such instruments are suitable for calculating QALYs in the context of cost-effectiveness analysis, there is an important drawback when using them for drug BR analysis: their dimensions are defined in terms of generic health attributes – such as physical functioning, social functioning, and vitality – and may therefore not be very sensitive and responsive to the disease of interest. So, although our results have shown that there are clear trade-offs among the considered alternatives, the relative differences in safety and efficacy may not be large enough to significantly change a patient's health status when this is measured in terms of generic health attributes.

The results from our example should be interpreted with care for three reasons. First, ideally evidence from all existing studies should be taken into account, rather than just a single trial. Future research should therefore consider our model together with evidence synthesis methods. As discussed previously, an appropriate method in such cases would be network meta-analysis (also known as the Mixed Treatment Comparisons model) [Salanti et al., 2008a, Sutton and Higgins, 2008] as it allows to take into account all evidence simultaneously. If a full network meta-analysis were performed, the random samples from the full joint posterior distribution of the effect estimates could be fed directly into the benefit-risk model. In this case, however, the possible inconsistencies in the network of trials would have to be evaluated, which brings additional level of complexity to the model. Second, the model is relevant only with respect to the data within the trial. For decisions depending on comprehensive benefit-risk profiles (e.g. drug marketing authorization decision), it can serve only as a starting point for further discussion as there can be additional qualitative information that is not included in the model. For example, our model excludes drug-drug interactions that might differ among the alternatives. Finally, the preferences were elicited from a single expert, and might not represent consensus among a larger group of experts.

To conclude, we presented a new MCDA-based approach to drug BR analysis with an example application to the therapeutic group of second-generation antidepressants. In contrast to previous models, our model is based on the SMAA methodology, which allows us to take into account the sampling variation that is inherent in criteria measurements that are based on clinical trials and/or observational studies. In addition, by making the trade-offs among the analyzed drugs explicit, we separated clinical data from subjective judgments, thereby increasing the transparency of the decision making process. Finally, the constructed model is specific to the therapeutic group of antidepressants. It would appear that the underlying concepts are general, but future research should assess the applicability of the SMAA methodology to other therapeutic groups.

Acknowledgements

We thank Barbara van Zwieten for her expert opinion and information about antidepressants. In addition, we acknowledge Kit Roes, Hein Fennema, Hans van Leeuwen, and Marcel Hekking of Schering-Plough, as well as the anonymous referees, for their valuable remarks and suggestions.

CHAPTER 6

Hit-and-Run enables efficient weight generation for simulation-based multiple criteria decision analysis

T. Tervonen, G. van Valkenhoef, N. Baştürk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 2012. doi: 10.1016/j.ejor.2012.08.026. (in press)

Abstract

Models for Multiple Criteria Decision Analysis (MCDA) often separate per-criterion attractiveness evaluation from weighted aggregation of these evaluations across the different criteria. In simulation-based MCDA methods, such as Stochastic Multicriteria Acceptability Analysis, uncertainty in the weights is modeled through a uniform distribution on the feasible weight space defined by a set of linear constraints. Efficient sampling methods have been proposed for special cases, such as the unconstrained weight space or complete ordering of the weights. However, no efficient methods are available for other constraints such as imprecise trade-off ratios, and specialized sampling methods do not allow for flexibility in combining the different constraint types. In this paper, we explore how the Hit-And-Run sampler can be applied as a general approach for sampling from the convex weight space that results from an arbitrary combination of linear weight constraints. We present a technique for transforming the weight space to enable application of Hit-And-Run, and evaluate the sampler's efficiency through computational tests. Our results show that the thinning factor required to obtain uniform samples can be expressed as a function of the number of criteria n as $\varphi(n) = (n - 1)^3$. We also find that the technique is reasonably fast with problem sizes encountered in practice and that autocorrelation is an appropriate convergence metric.

6.1 Introduction

Multiple Criteria Decision Analysis (MCDA) methods consider a set of alternatives that is evaluated in terms of a set of criteria in order to choose the best one, to rank them from best to worst, or to sort them into ordered categories [Roy, 1996]. The alternatives are evaluated with respect to the chosen preference model, and often some type of preferential independence is assumed. This allows the evaluation to be decomposed in two parts: (i) evaluation with respect to individual criteria, and (ii) aggregation of the per-criterion scores to describe the alternatives' overall attractiveness interpreted according to the chosen preference model.

Although various MCDA methods allow criteria values to be specified in an imprecise format, e.g., as probability distributions in Multi-Attribute Utility Theory (MAUT) or indirectly through thresholds in ELECTRE methods, they require exact weights to quantify the relative contribution of the individual criteria to the alternatives' overall attractiveness. The Stochastic Multicriteria Acceptability Analysis (SMAA) family of methods [Tervonen and Figueira, 2008] extends the traditional MCDA preference models by allowing to take into account uncertainty in all model parameters, including the weights. The indices describing the decision problem are estimated through Monte Carlo simulation, and each iteration includes sampling uniformly distributed weights from a convex polytope defined through a set of linear weight constraints [Tervonen and Lahdelma, 2007]. For each of the underlying preference models, several types of constraint are available to restrict the feasible weight space in a theoretically meaningful way. However, in practical applications of SMAA, the use of such constraints is limited by the lack of efficient sampling algorithms, which results in insufficient discrimination between the decision alternatives for some problem instances.

One approach to sample from a uniform distribution over the target polytope is to draw samples from a uniform proposal density over a polytope that approximates the target, such as a uniform density over a rectangular hyperbox or a Dirichlet distribution. The Dirichlet distribution is uniform over the simplex when the concentration parameter is set to 1, and has been applied in the MCDA setting by [Jia et al., 1998]. However, since the proposal density only approximates the target density, such methods require a rejection step. In general, the rejection rate increases exponentially with the dimension of the sampling space, thus making this approach infeasible for higher dimensions. Alternatively, it is possible to simulate weights using different Markov Chain Monte Carlo (MCMC) methods. There is often a trade-off between the mixing rate and the acceptance rate of the sampler. For uniform joint and conditional distributions, a standard single-state Gibbs sampler is applicable. The rejection rate in this case is 0 by definition and weights can be simulated iteratively satisfying linear bounds and ratio constraints. This iterative sampling method typically leads to high correlations between draws and slow mixing [Amit and Grenander, 1991, Besag et al., 1995]. It is possible to improve the mixing properties by simulating the weights jointly using random walk algorithms.

Hit-And-Run (HAR) is an MCMC sampling algorithm that, unlike other random walk algorithms, is known to mix rapidly from any interior point [Lovász, 1999, Lovász and Vempala, 2006]. HAR has two clear advantages over the alternatives. First, it provides block samples from the uniform weight distribution, typically leading to high mixing rates. Second, application of HAR on the transformed parameter space avoids the rejection step in the standard block samplers under linear constraints. The rejection rate of the algorithm is 0 by definition. Both the mixing rate and the acceptance rate are improved compared to the existing samplers in this context. However, using HAR for efficient MCMC sampling from the restricted weight space as encountered in SMAA is not trivial as the weight space needs to be transformed before the algorithm can be applied.

In this paper, we consider the application of HAR sampling to weight generation in SMAA and other simulation approaches in MCDA that require uniform sampling from a restricted weight space [Butler et al., 1997, Jia et al., 1998]. Our contribution is twofold. First, we present a technique for transforming the n dimensional weight space to an $n-1$ dimensional sampling space. We then evaluate through computational tests how the thinning factor required to obtain a sample equivalent to 10,000 uniform draws depends on the dimensionality of the problem. A thinning factor φ indicates that from the sequence of generated samples, we store only the φ-th sample. Because HAR sampling generates a series of dependent samples, thinning reduces the memory required to store the samples by accepting a relatively small loss of information about the target density.

The remainder of this paper is structured as follows. In Section 6.2 we discuss the weight constraints that are typically encountered in SMAA analyses and further motivate the value of efficient weight sampling methods for interactive decision aiding. Section 6.3 describes the HAR algorithm and discusses its application to weight generation. Section 6.4 discusses metrics to assess sample uniformity. Section 6.5 presents the results from the computational tests, and Section 6.6 concludes and provides directions for future research.

6.2 Weight constraints in SMAA

Consider a discrete multi-criteria decision problem consisting of a set of m alternatives that are evaluated in terms of n criteria. The vector of criteria values corresponding to criterion j is denoted by $x_j = (x_j^1, \ldots, x_j^m)^T$, where x_j^i denotes the performance of alternative i on criterion j.

Generally, the alternatives are first evaluated with respect to the individual criteria to obtain criterion-specific attractiveness scores, after which some kind of aggregation procedure is applied to combine the criterion-specific scores into an overall measure of preference or value [Choo et al., 1999]. Let the functions $f_j(x_j^i)$ and $g_j(x_j^i, x_j^k)$ give the criterion-specific score of alternative i on criterion j (value or utility based approaches) and the criterion-specific score of the pair (i, k) on criterion j (outranking based approaches), respectively. For the purpose of this paper, we assume that the evaluation of the alternatives with respect to the individual criteria has already been completed, so that we can focus on specifying the aggregation functions $f(f_1(x_1^i), \ldots, f_n(x_n^i))$ and $g(g_1(x_1^i, x_1^k), \ldots, g_n(x_n^i, x_n^k))$.

To simplify the assessment of $f()$ and $g()$, it is generally assumed that these functions are additive:

$$\begin{aligned} f(f_1(x_1^i), \ldots, f_n(x_n^i)) &= \sum_{j=1}^{n} w_j f_j(x_j^i) \\ g(g_1(x_1^i, x_1^k), \ldots, g_n(x_n^i, x_n^k)) &= \sum_{j=1}^{n} w_j g_j(x_j^i, x_j^k), \end{aligned}$$

where w_j denotes the weight of criterion j. Although the algebraic shape of $f()$ and $g()$ is the same, the meaning of the weights depends on the underlying preference model. In compensatory approaches such as Multi-Attribute Value Theory (MAVT) or MAUT, the weights are scaling factors that ensure that unit increases in the functions $f_j()$ are commensurate [Keeney and Raiffa, 1976]. In non-compensatory approaches such as ELECTRE or PROMETHEE, by contrast, the weights do not represent value trade-offs between criteria scale swings but should rather be interpreted as the amount of voting power that is allocated to each of the criteria [Vansnick, 1986, Belton and Stewart, 2002]. For a more thorough discussion of the interpretation of weights in MCDA, we refer to Choo et al. [1999].

The traditional approach in MCDA is to establish exact values for the weights by applying a dedicated weight elicitation technique, such as SWING weighting, which is then followed by extensive sensitivity analyses to explore how robust the obtained results are to small changes in the weights and the criteria values. In many real-life decision making contexts, however, decision makers do not feel confident with providing exact numerical values for the weights and/or the criteria values. In such situations, a method from the SMAA family can be applied to compute, for a wide range of weights and criteria values, the probability that an alternative is placed at a certain rank (ranking problems; Lahdelma and Salminen, 2001) or that it belongs to a certain category (sorting problems; Tervonen et al., 2009a). The estimation of these indices is achieved through Monte Carlo simulation, and in each iteration criteria weights sampled from their joint probability distribution are required. To generate these values, it is typically assumed that the weights are uniformly distributed within the convex polytope defined through a set of weight constraints.

Without loss of generality, we assume that the weights are non-negative and normalized so that they sum to one. When no preference information is available, the feasible weight space is an $n-1$ dimensional simplex in n dimensional space:

$$W_n = \left\{ w \in R^n : w \geq 0 \text{ and } \sum_{j=1}^{n} w_j = 1 \right\}.$$

In practice, it is often possible to elicit some preference information from the decision maker, such as a partial or complete ranking of the criteria weights. Such information can be included in the model by restricting the feasible weight space accordingly (see Figure 6.1 for an example). Weight constraints in SMAA models should be elicited by taking into account the underlying preference model. In MAVT or MAUT, meaning-

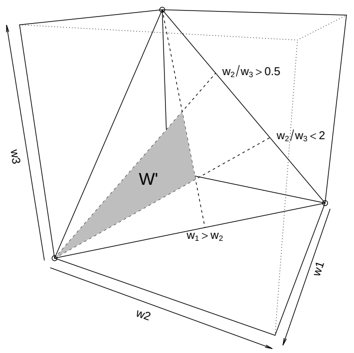


Figure 6.1: The feasible weight space W' (gray polygon) for a 3-criteria problem with the partial ranking $w_1 > w_2$ and the trade-off ratio constraint $w_2/w_3 \in [0.5, 2]$. The solid triangle is the feasible weight space without preference information (W_3).

ful constraints are ordinal rankings of the weights ($w_i > w_j$) and intervals for trade-off ratios between weights ($w_i/w_j \in [(w_i/w_j)^{\min}, (w_i/w_j)^{\max}]$). In case of ELECTRE, lower- and upper bounds ($w_j^{min} \leq w_j \leq w_j^{max}$) may be considered as well.

Although efficient weight generation techniques exist for sampling from the unconstrained weight space as well as from the weight space constrained through a complete ordering of the weights or weight lower-bounds [Tervonen and Lahdelma, 2007], no methods have been proposed for sampling weights constrained with imprecise trade-off ratios or with a combination of constraints. While the use of ordinal rankings of the weights or weight lower-bounds may result in high discrimination for some problem instances, for others the preference information portrayed by these constraints may not be precise enough to sufficiently differentiate the decision alternatives. For small problem instances, a lack of discrimination with the currently included weight constraints can be resolved in an iterative and interactive way by first eliciting more precise preference information from the decision maker and then applying rejection sampling to sample the weights accordingly. However, for rejection sampling the hit rate decreases exponentially with the number of criteria, so it is intractable even for a moderate number of criteria (i.e. ≥ 10). Hence, the ability to efficiently sample weight vectors for arbitrary combinations of the different types of weight constraint is critical to the real-world application of SMAA to larger problem instances. The HAR sampler introduced in the next section is particularly suited for this purpose.

6.3 Hit-And-Run (HAR) for weight generation

In HAR sampling [Smith, 1984], the Markov chain is initialized with a starting point within the polytope. At each iteration a random direction is generated by sampling from the unit hypersphere, implemented efficiently by generating independent normal variates and normalizing them to form the sample point [Marsaglia, 1972]. The random direction together with the current position generates a line set, and its intersection with the boundary of the polytope generates a line segment from which the next point is drawn uniformly. The generated Markov chain converges on the uniform distribution over the polytope in nondeterministic polynomial time $O^*(n^3)$. So to apply the HAR sampler to weight generation in the constrained weight space, the sampling space must be defined appropriately, line intersections must be computed, and a starting point contained in the polytope must be chosen. The HAR sampler and the transformations described here are implemented as a package for the R statistical software, and are available at `http://cran.r-project.org/web/packages/hitandrun/`.

6.3.1 Sampling space transformation

The feasible weight space with constraints defines a convex polytope $W' \subseteq W_n$. The $(n-1)$-simplex W_n is coincident with the hyperplane $W_n^* = \left\{w \in R^n : \sum_{j=1}^n w_j = 1\right\}$. Thus, the volume of the feasible weight space is essentially 0, which means that if we perform MCMC sampling in n dimensions, the probability of hitting inside W' is also 0. Therefore, we transform the simplex so we can sample directly in $n-1$ dimensions.

In the following, let I_n be the $n \times n$ identity matrix. The centroid of W_n is at $(1/n, \ldots, 1/n)^T$, so if we translate the plane W_n^* by $(-1/n, \ldots, -1/n)^T$, it forms an $n-1$ dimensional subspace $V \subset R^n$. We obtain an orthonormal basis $\{v^1, \ldots, v^{n-1}\}$ of V by first defining a basis of V and then performing orthogonalization and normalization. A basis can be defined by choosing $n-1$ vectors, so that for the kth vector the nth component is -1, the kth component is 1, and the others are 0. For a 2-simplex in R^3 (as shown in Figure 6.1) such a basis would be $\{(1, 0, -1)^T, (0, 1, -1)^T\}$. Now, to map an arbitrary point $x \in R^{n-1}$ to a point in the target space $w \in W_n^*$, we apply an affine transformation: a change of basis followed by a translation. To do this, we use the homogeneous coordinate representation $x = (x_1, x_2, \ldots, x_{n-1}, 0, 1)^T$:

$$w = TBx$$

where B is the $(n+1) \times (n+1)$ augmented change-of-basis matrix and T the $(n+1) \times (n+1)$ translation matrix:

$$B = \begin{pmatrix} v_1^1 & \cdots & v_1^{n-1} & \sqrt{1/n} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ v_n^1 & \cdots & v_n^{n-1} & \sqrt{1/n} & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \; ; \; T = \begin{pmatrix} & \Large I_n & & 1/n \\ & & & \vdots \\ & & & 1/n \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

To preserve the uniform distribution of the samples, we need the transformation to preserve distances between points: it must be *isometric*. Specifically, in Euclidian space, a transformation f is isometric if $||y-x|| = ||f(y) - f(x)||$ for all x, y. First note that B is unitary ($B^T B = I_{n+1}$) because it is the homogeneous coordinate representation of an orthonormal change of basis matrix. As (i) unitary matrices are isometric [Berberian, 1999], and (ii) translation is isometric, it follows that the proposed transformation is also isometric. Therefore uniform samples obtained in R^{n-1} are also uniform when transformed to W_n^*.

To complete the transformation, we show how the linear constraints that define the polytope $W' \subseteq W_n$ are defined in $n-1$ dimensions. Let us denote the constraint set defining W_n as follows:

$$Cw \leq b \;\; ; \;\; C = -1I_n, \, b = (0, 0, \ldots, 0)^T$$

$$\sum_{i=1}^{n} w_i = 1$$

Since we sample directly from the plane W_n^*, the equality constraint is satisfied by definition. The weight constraints presented in Section 6.2 are linear and can therefore be represented as additional rows in C and b. Then the constraints can be expressed in $n-1$ dimensions as:

$$Ax \leq b \;\; ; \;\; A = CTB$$

since $Ax = C(TBx) = Cw$.

The transformation is illustrated for a three-dimensional weight space with ordinal constraints ($w_1 > w_2 > w_3$) in Figure 6.2. First, the linear constraints that define the polytope in the n dimensional weight space are transformed to the $n-1$ dimensional sampling space. Then, HAR is applied to generate an (approximately) uniform sample in the sampling space. Finally, the sampled points are mapped to the weight space.

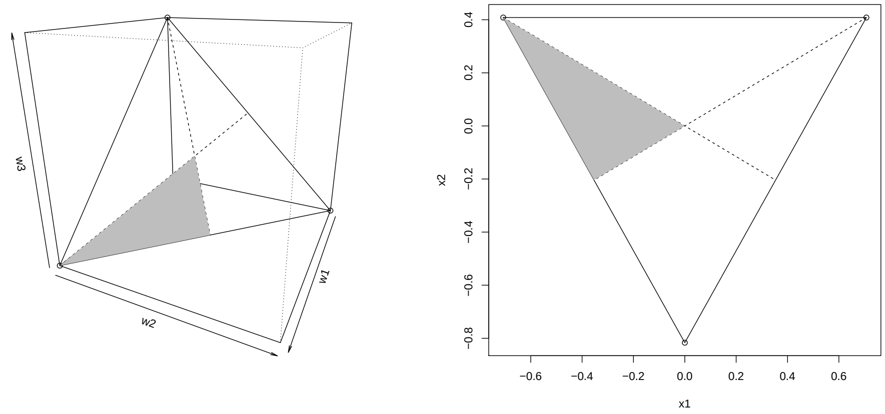
6.3.2 Line intersection

Given a point x (in homogeneous coordinates) in the polytope and a direction vector $d = (d_1, \ldots, d_{n-1}, 0)$, the line through x along d is $x + ld$. We want to find the interval $L = [L_0, L_1]$ such that $A(x + ld) \leq b$ iff $l \in L$. Then $lAd \leq b - Ax$, where either side is a vector: $lu \leq v$. Since all v_i are non-negative, positive u_i give the upper bound L_1 and negative u_i the lower bound L_0. If u_i is zero the direction d is parallel to the i-th constraint and provides no information on the bounds. The lower and upper bound are thus given by:

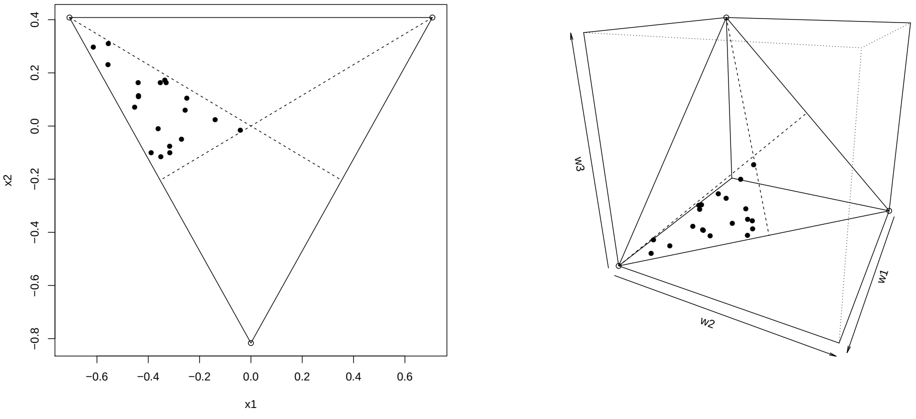
$$L_0 = \max_{i:u_i<0} \frac{v_i}{u_i} \;\; ; \;\; L_1 = \min_{i:u_i>0} \frac{v_i}{u_i} \quad .$$

6.3.3 Starting point

The starting point can be defined in several ways, either deterministically or in a pseudo-random manner. The underlying principle is the same for all techniques:



(a) Polytope in 3D (weight space) (b) Polytope in 2D (sampling space)



(c) Samples in 2D (sampling space) (d) Samples in 3D (weight space)

Figure 6.2: The sampling space transformation illustrated in 3D space. The preference information (expressed as linear constraints) defines a 2D convex polytope in the 3D weight space (a). The linear constraints are transformed to a 2D space (b). Then, HAR is used to sample from the polytope (c). Finally, the samples are transformed back to the 3D weight space (d).

we first determine a set of points within the polytope, and then take their weighted average. Extreme points along each dimension $k \in \{1, \ldots, n-1\}$ can be found by solving the following linear programs:

$$\begin{array}{ll|ll} \text{maximize} & x_k & \text{minimize} & x_k \\ \text{subject to} & Ax \leq b & \text{subject to} & Ax \leq b \end{array}$$

The obtained solutions are not necessarily the vertices of the polytope; if such are desired, they can be efficiently enumerated using the Avis-Fukuda pivoting algorithm [Avis and Fukuda, 1995].

If we choose a deterministic starting point, a point close to the centroid can be approximated by taking the mean of the extreme values along a subset of the dimensions. The exact centroid is the mean of all the vertices. To generate a pseudo-random starting point, we can use any set of interior points and take their weighted average with randomly generated weights.

6.4 Convergence metrics

With HAR, as with all MCMC samplers, it is important to assess whether a generated sample is representative of the target distribution, and this is determined by the mixing of the chain together with the number of iterations. We assume that the samples drawn from the restricted weight space will subsequently be used in simulation-based decision analysis, where the computed indices are insensitive to sample autocorrelation: as long as the draws uniformly cover the restricted weight space, the mixing itself is irrelevant for the computation of the indices describing the results of the decision analysis. In most MCMC applications, it is important to disregard an initial subsample from the Markov chain because the starting values may have been very unlikely, meaning that the initial iterations may have been used to move away from a low density region, leading to disproportionate representation of that part of the space and thus to biased results. This is not an issue in our setting as the sampling distribution is uniform and HAR mixes fast from any starting point [Lovász and Vempala, 2006].

Many commonly used convergence diagnostics for MCMC perform poorly because they are based on the first and second moments of the distribution (for each dimension separately) and depend on the target distribution being continuous. The constrained weight polytope can be quite irregularly shaped, and the target distribution is discontinuous at its boundaries. Moreover, these diagnostics do not assess uniformity of the samples directly, but rather the stability of certain moments or quantiles. Thus, to measure convergence to the target distribution, a metric is needed that takes into account the geometry of the polytope and measures the uniformity of the samples.

Our main metric for assessing convergence to the target distribution is the Friedman-Rafsky two-sample Minimum Spanning Tree (MST) test [Friedman and Rafsky, 1979, Smith and Jain, 1984] that compares the obtained sample Y with a sample X from the target distribution. Thus, this test is useful for evaluating the convergence of HAR when a sample from the target distribution can be obtained efficiently, but cannot be used to assess convergence in a general setting. The test assesses whether X and Y were drawn from the same distribution by constructing a MST for $X \cup Y$ and counting the number of within- and across-sample edges. Finally, a z-value for the null-hypothesis that both samples are from the same distribution is computed, which can be compared against quantiles of the normal distribution. A z-value in the lower tail indicates that Y is concentrated in sub-regions of the polytope (aggregation), whereas a z-value in the upper tail indicates Y is more regularly spaced than expected (regularity). For our purposes, only the lower-tail alternative hypothesis of aggregation is relevant.

Because the MST-based metric is not generally applicable, we also evaluate three other metrics for assessing convergence. The first is the Coefficient of Variation (COV), σ/μ, of the nearest-neighbour distances $\Gamma = \{\min_{j \neq i} |y_i - y_j| \ : \ y_i \in Y\}$, where μ and σ are the mean and standard deviation of the corresponding distance for draws from the sample distribution. Lower values of COV indicate more regular spacing of the sample points. The second metric, the Standardized Component-wise Error (SCE), is based on the fact that the mean of a sample from the target distribution will closely approximate the centroid h. Thus, $e = \mu - h$ is a measure of how close Y is to the target distribution. The component-wise errors e_i depend on the shape and dimension of the polytope, so they need to be standardized in order to be useful as a convergence metric. Thus, we define the SCE as e_i/s_i, where s_i is the sample standard deviation of the i-th vector component in Y. Finally, we use the sample autocorrelation at a given lag τ as a convergence metric. This measure is defined as:

$$R(\tau) = \frac{E[(y_t - \mu)(y_{t+\tau} - \mu)]}{\sigma^2},$$

where $t \in \{1, \dots, |Y| - \tau\}$. Like the SCE, autocorrelation is calculated individually for each component of the weight vector. Both metrics are scale invariant, and therefore the component-wise scores can be aggregated into an overall score simply by taking the maximum.

6.5 Computational tests

We assessed the thinning factors $\varphi_a(n)$ required to achieve uniformity of HAR sampling with constraints representing complete ordinal preference information ($w_1 > w_2 > \cdots > w_n$). This class of problem instances enables generating the sample X as required by the MST test by using an efficient algorithm [Tervonen and Lahdelma, 2007] and calculating the centroid $h = (h_1, \dots, h_n)$ as required by the SCE test as

$$h_i = \frac{1}{n} \sum_{j=i}^{n} \frac{1}{j} \ .$$

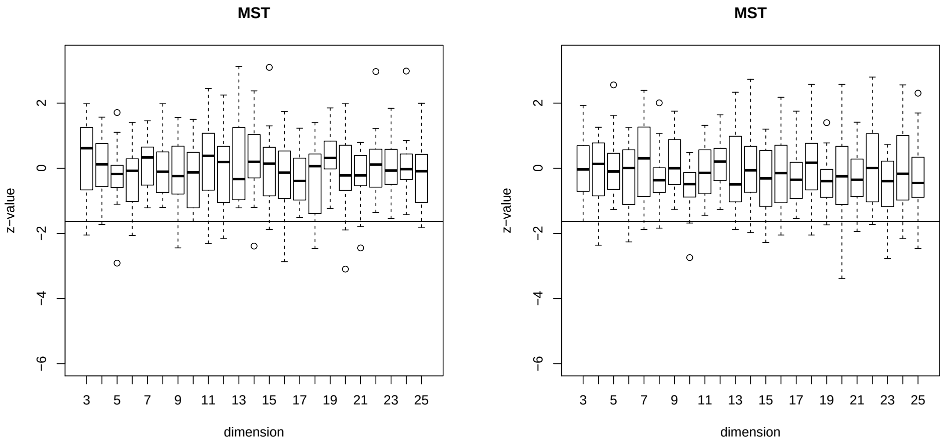
The amount of thinning required depends on the shape of the convex polytope W', with more conical shapes requiring higher thinning. The ordinal information causes

W' to be at least as conical as what is expected with other realistic weight constraints because the decision makers often express imprecise weight information with similar precision for all criteria / pairs of criteria.

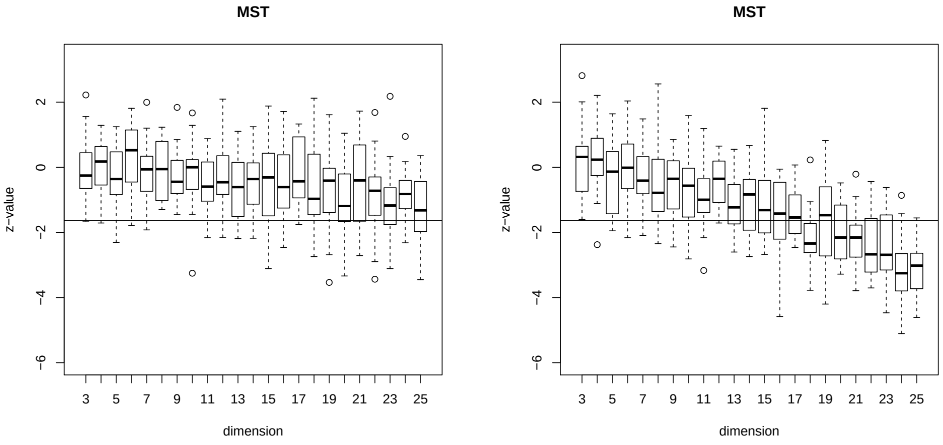
As an initial exploratory test, we constructed a large chain for $n \in \{3, \ldots, 15\}$ and recorded the minimum sufficient thinning factors required to obtain a z-value ≥ -1.64 from the MST test and a maximum SCE < 0.05. Previous research suggests that HAR mixes with $O^*(n^3)$ iterations [Lovász, 1999]. Because the algorithm reduces to uniform sampling for $n = 2$ (since we sample in $n - 1$ dimensions), the required thinning for $n = 2$ should be 1. Thus we fitted $\varphi_a(n) = a(n - 1)^3 + (1 - a)$ to our exploratory test data, which suggested that $a \geq 0.2$. However, there was a large degree of uncertainty in this estimate because we used only a single chain for each dimension and censoring occurred for some dimensions. Based on visual inspection of autocorrelation plots of the exploratory test data, we chose to use a lag of $\tau = 25$ for the autocorrelation metric in the validation tests. We also performed the exploratory tests with the standard Gibbs sampler, and found it to exhibit slower convergence than HAR.

We subsequently generated 20 HAR samples of $10,000$ weight vectors (sufficient for SMAA analyses; Tervonen and Lahdelma [2007]) for each $n \in \{3, \ldots, 25\}$ and for each thinning factor $\varphi_a(n)$ with $a \in \{0.125, 0.25, 0.5, 0.75, 1.0\}$. For the MST test, a separate benchmark sample X was generated for each HAR sample Y to marginalize the impact of random properties of a single benchmark sample. This led to respectively 68%, 38%, 14%, 7% and 8% of the HAR samples being rejected (Figure 6.3). Thus, either $a = 0.75$ or $a = 1.0$ could be appropriate, as they are close to the optimal 5% rejection rate. As the shape of the sampling space affects the rate of convergence, we recommend to use $a = 1.0$. The running times that were required to generate $10,000$ samples with thinning factors $\varphi_{1.0}(n)$ on an Intel Xeon X3440 2.53GHz CPU are shown in Figure 6.4. To illustrate the necessity of an MCMC approach, running times for rejection sampling are also shown: with $n = 3$ criteria, one in three samples are accepted; with $n = 10$ criteria one in 1.8 million, giving a running time of several hours for our implementation. The plotted sampling times show that using HAR reasonably high dimensionality problems can be analyzed almost interactively, i.e. the median sampling times for 5, 10, and 15 dimensions were 0.92, 22, and 136 seconds, respectively.

We also computed the other test metrics (COV, SCE, and autocorrelation at lag $\tau = 25$) for both the HAR samples and the benchmark samples. The COV metric converged to stable values long before the sample converged to uniformity and therefore it is not useful, so we omit the detailed results for this metric. Figure 6.5 shows how the SCE values for the HAR samples compared to those for the benchmark samples. The autocorrelation is plotted in Figure 6.6. The SCE and autocorrelation metrics appear to be useful and are correlated with the MST z-value ($\rho = -0.60$ and $\rho = -0.79$, respectively). Since autocorrelation can be calculated based on the sample alone, while computation of the SCE requires enumerating the vertices of the polytope, the former is potentially the most useful metric. Figure 6.7 shows a scatter plot of the autocorrelation value against the MST z-value, the Receiver Operating Characteristic (ROC) curve of autocorrelation as a predictor of $z < -1.64$. An autocorrelation cutoff



(a) benchmark (b) HAR, $a = 1.0$



(c) HAR, $a = 0.5$ (d) HAR, $a = 0.25$

Figure 6.3: Boxplot of MST z-values from validation tests with 20 sets of 10,000 samples for each n. The horizontal line at $z \approx -1.64$ corresponds to p-value 0.05, below which the null hypothesis of uniformity is rejected. Subfigure (a) shows how the MST test performs when comparing two uniform samples.

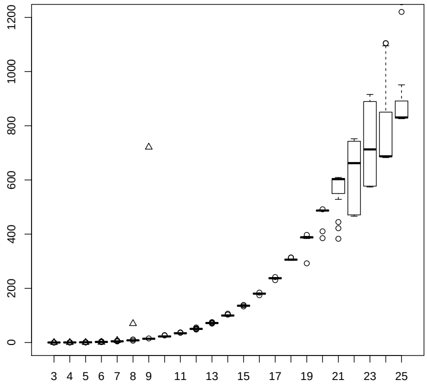


Figure 6.4: Boxplot of execution times (in seconds, y-axis) of 20 runs for each dimension n (x-axis) to generate $10,000$ samples at thinning $\varphi_{1.0}(n)$. The triangles show the running time for a rejection sampler for up to $n = 9$; with $n = 10$ criteria rejection sampling took $8 \cdot 10^3$ seconds.

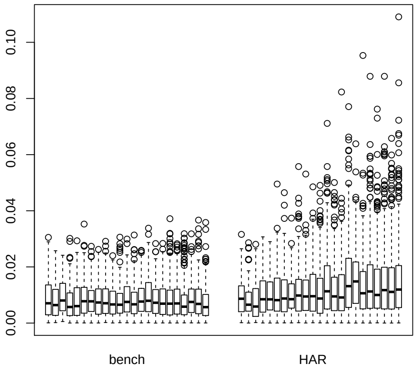


Figure 6.5: Box plot for the SCE metric, for both the benchmark sets and the test sets generated with thinning $\varphi_{1.0}(n)$. The horizontal axis shows the number of criteria n.

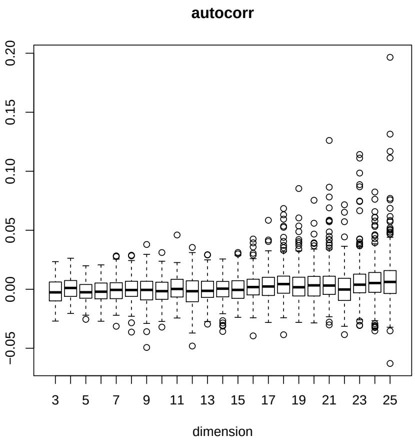


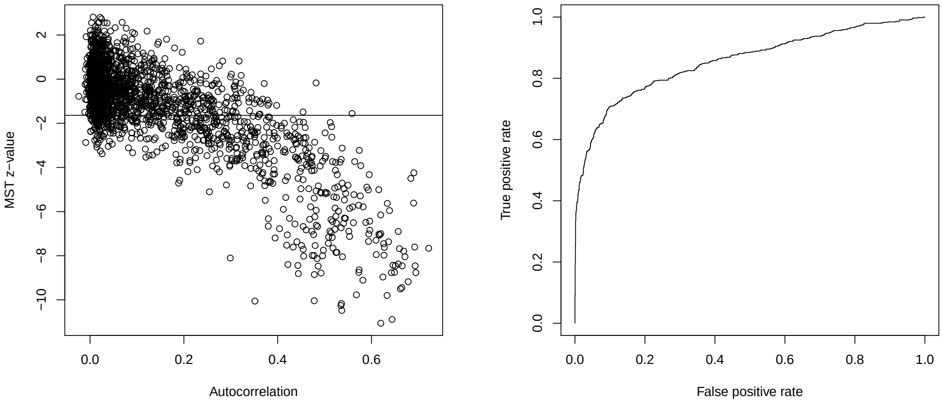
Figure 6.6: Box plots for autocorrelation at lag $\tau = 25$ for test sets generated with thinning $\varphi_{1.0}(n)$. The horizontal axis shows the number of criteria n.

value of 0.05 correctly rejected 82% of the samples that failed the MST test. Note that the samples that passed the autocorrelation test but failed the MST test are still likely to be sufficiently uniform to be used in simulation-based decision analysis as small deviations from uniformity are unlikely to affect the model outcomes. The code used to generate all the test results and the complete set of summary statistics are available as online supplements.

6.6 Conclusions

In this paper, we considered the application of HAR to sample uniformly from a subset of the n-simplex defined by linear constraints. Our contribution was in presenting the transformation to $n-1$ dimensions for efficient MCMC sampling and in assessing the thinning factor required to achieve acceptable deviation from a uniform distribution over the constrained weight space. The transformation is not specific to HAR, and can also be used to apply other MCMC samplers to the weight sampling problem. The computational tests showed that HAR is quite fast in small ($n \leq 15$) problems for 10,000 samples. The tests also showed that a thinning factor of $\varphi(n) = (n-1)^3$ is almost always sufficient to be unable to reject the null hypothesis of uniformity with the MST-test at the 0.05 confidence level. We also assessed other measures of convergence to the target distribution and found that the autocorrelation at lag 25 is the most appropriate to use in absence of a representative sample from the target distribution (meaning that the MST test is inapplicable). However, autocorrelation does not measure convergence directly and future research could yield better tests.

When used in the context of simulation-based decision analysis, HAR is slower than the techniques presented by Tervonen and Lahdelma [2007] for sampling



(a) Scatter plot

(b) ROC curve

Figure 6.7: Autocorrelation at delay $\tau = 25$ is predictive for failing the MST test.

weights representing missing or ordinal preferences. However, HAR is far more flexible as it also allows for polynomial-time generation of upper bounded and ratio interval constrained weights that previously had to be generated with rejection sampling. Especially the ratio interval weights are important, as the meaning of a weight in MAVT models is that of trade-off ratios, and imprecision should therefore be modeled in a compatible manner.

The sampling technique evaluated in this paper assumes that the sampling space is convex, and this is fulfilled by all weight constraints within the additive model. However, some preference models include other parameters as well, and in that case the space of feasible preference parameters can be concave, e.g. the general monotone value functions of robust ordinal regression [Greco et al., 2008, 2010]. Future research should investigate whether MCMC sampling can be applied for the efficient sampling of the preference parameters in such spaces as well.

CHAPTER 7

Complex preference information in benefit-risk assessment

Abstract

The previous chapter showed how, using hit-and-run, weights can be efficiently sampled from the feasible weight space restricted by arbitrary linear constraints. This enables new constraints to be considered, and allows more flexible combining of constraints. This chapter illustrates how this can be used to better support drug benefit-risk decision making. To this end, possible preference scenarios are explored in the context of an existing study of the benefit-risk profile of two anti-thrombolytic drugs. In this case study, the anti-thrombolytic may have the beneficial effect of reducing the risk of deep vein thrombosis (DVT), but it may have the harmful effect of increasing the risk of a major bleed. The simple 2×2 structure of this case study allows a simple illustration of the methods. However, we argue that the case has a more natural interpretation as a 2×3 or even 3×3 problem, and show how Stochastic Multicriteria Acceptability Analysis (SMAA) with imprecise preferences can be used to aid a decision in that case. It will be discussed how the strengths of using SMAA and hit-and-run become clear in higher dimensional problems.

7.1 Introduction

The benefit-risk assessment of medicines is a complex problem often involving multiple criteria that represent efficacy and safety concerns. Clinical trials provide the pivotal evidence for such assessment, and therefore the criteria measurements are inherently uncertain. Moreover, it may be difficult to precisely articulate preferences. For simple problems with two criteria and two alternatives (i.e. a 2×2 problem), a simple approach based on stochastic simulation and two-dimensional plotting that allows for full uncertainty in the criteria measurements and preferences can be applied [Lynd and O'Brien, 2004]. However, when more criteria and alternatives need to be considered, an approach based on Multiple Criteria Decision Analysis (MCDA) is more appropriate. Early applications of MCDA to benefit-risk analysis [Mussen et al., 2007, Felli et al., 2009] ignored the uncertainty inherent in the domain. The Stochastic Multicriteria Acceptability Analysis (SMAA) model proposed in this thesis (Chapter 5) combines the problem structuring approach of MCDA with stochastic simulation for taking into account uncertainty in the criteria measurements as well as partial or imprecise preference information.

While SMAA in theory enables arbitrary preference information to be used, in practice this was limited by the lack of an efficient sampling algorithm. The hit-and-run algorithm enables the efficient sampling of weights with arbitrary linear constraints (Chapter 6). This enables much greater flexibility in dealing with partial, imprecise, or complex preference information.

In this chapter, we will first illustrate the SMAA approach in a 2×2 decision problem. This helps to clarify the meaning of different weight constraints by explicitly visualizing them. Next, we extend the analysis to a 2×3 problem to show how the SMAA analysis generalizes to multi-dimensional problems. The implications of different types of preference information are illustrated by visualizing the weight space and by analyzing the SMAA decision metrics. Finally, we discuss the implications of the more flexible approach to preference information.

7.2 Two-dimensional analysis

The case study is based on an analysis of the benefit-risk profile of two drugs for the prophylaxis of deep vein thrombosis (DVT) following major trauma [Lynd and O'Brien, 2004]. The analysis was based on the results of a clinical trial comparing heparin and enoxaparin for both efficacy and safety [Geerts et al., 1996]. The safety concern is that administering anticoagulants to trauma patients already at an elevated risk of bleeding might cause additional major bleeding episodes. The benefits can be assessed as either prevention of proximal DVT, or of all DVT (i.e. both proximal and distal) [Lynd and O'Brien, 2004]. Proximal DVT is more often associated with the development of serious complications.

In the original analysis, the authors derive beta distributions for the risk of a major bleed, the risk of any DVT, and the risk of proximal DVT for both heparin and enoxaparin from the original trial data. The original trial data as well as the parameters

	Data		Beta distribution		
Event	r	r/n	α	β	median (95% CI)
Heparin ($n = 136$)					
Any DVT	60	0.441	60	76	0.441 (0.359–0.525)
Proximal DVT	20	0.147	20	116	0.145 (0.093–0.211)
Distal DVT	40	0.294	40	96	0.293 (0.221–0.373)
Major bleeds	1	0.007	1	135	0.005 (0.000–0.027)
Enoxaparin ($n = 129$)					
Any DVT	40	0.310	40	89	0.309 (0.233–0.392)
Proximal DVT	8	0.062	8	121	0.060 (0.027–0.109)
Distal DVT	32	0.248	32	97	0.247 (0.178–0.326)
Major bleeds	5	0.038	5	124	0.036 (0.013–0.078)

Table 7.1: The original trial data given as number of events r and proportion of events r/n. Estimated beta distribution parameters α and β and the characteristics of the estimated beta distribution given as the median and 0.025 and 0.975 quantiles.

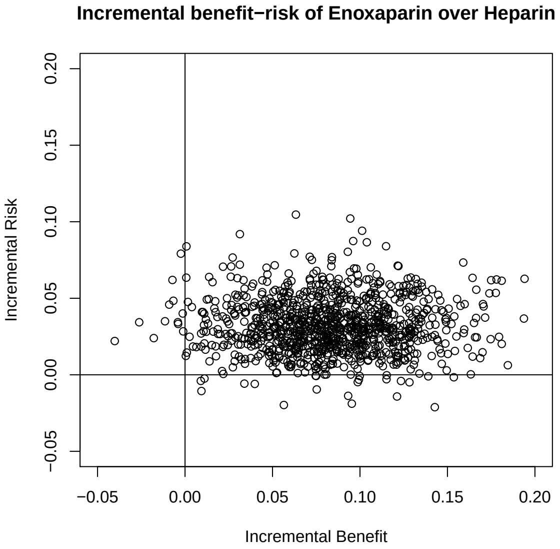


Figure 7.1: The benefit-risk plane showing incremental benefit (reduction in risk of proximal DVT) versus incremental risk (increase in risk of major bleeds). Enoxaparin is associated with an increase in benefit as well as an increase in risk.

and characteristics of the estimated beta distributions are given in Table 7.1, where we have additionally calculated the values for distal DVT. They sampled points from these distributions, and for each event they calculated the difference between the risk for enoxaparin compared to heparin, the *incremental* risk. A plot of the incremental risk of proximal DVT versus the incremental risk of a major bleed was used to show that there is a clear trade-off between the criteria (reproduced in Figure 7.1).

To illustrate different types of constraint, we show how SMAA can be applied to this case. The alternatives and criteria, as well as the measurements (beta distributions) are the same as for the original analysis (Table 7.1). To construct the partial value functions, we determine bounds for the likely values of the criteria measurements from the confidence intervals given in Table 7.1. For proximal DVT, we set the range from 0.0 (best value) to 0.25 (worst value), and for major bleeding from 0.0 (best value) to 0.1 (worst value). The partial value function $u_b(c_b)$ for major bleeding then maps the extremes of the scale to $u_b(0.0) = 1.0$ and $u_b(0.1) = 0.0$, and linearly interpolates between the two: $u_b(0.05) = 0.5$. The partial value function $u_p(c_p)$ does the same for proximal DVT. For fixed weights $\mathbf{w}$ and criteria measurements $\mathbf{c}$, the overall utility is then determined by

$$u(\mathbf{w}, \mathbf{c}) = w_b u_b(c_b) + w_p u_p(c_p) \ ,$$

where $w_b = 1 - w_p$. The ratio w_p/w_b expresses how much more important is decreasing the incidence of proximal DVT from 0.25 to 0.0 than decreasing the incidence of major bleeding from 0.1 to 0.0. Then, by fixing w_p to a certain value (between 0.0 and 1.0) and sampling from the measurement distributions, we can determine for what proportion of the samples heparin has higher utility than enoxaparin. This is what is done for a SMAA analysis with exact weight information. The first-rank acceptability of enoxaparin is plotted for different values of w_p in Figure 7.2.

In a preference free SMAA analysis the first-rank acceptability for enoxaparin is 0.54, shown as a dotted line in Figure 7.2. This value is derived by integrating the ranks over all possible weight vectors, as well as the probability distributions of the criteria measurements, using Monte Carlo sampling. The central weight vector of enoxaparin without preference information is (0.7, 0.3), confidence factor 0.89, whereas that for heparin is (0.28, 0.72), confidence factor 0.84. In this case, there is a large difference between the first-rank acceptability of enoxaparin (0.54) and the confidence factor of its central weight vector (0.89). This indicates that the ranking of the alternatives is sensitive to the preference information, and that the low first-rank acceptability of enoxaparin can not be attributed to poor performance of enoxaparin, nor to uncertainty in the measurements. Thus, precise preference information is clearly needed to make a decision.

As an initial step in analyzing the decision maker's preferences, we apply ordinal swing weighting to obtain a ranking of the criteria. We confront the decision maker with a hypothetical 'worst case' scenario:

- 25% proximal DVT, 10% major bleeding

and offer the choice to improve one of the two criteria to the best possible value:

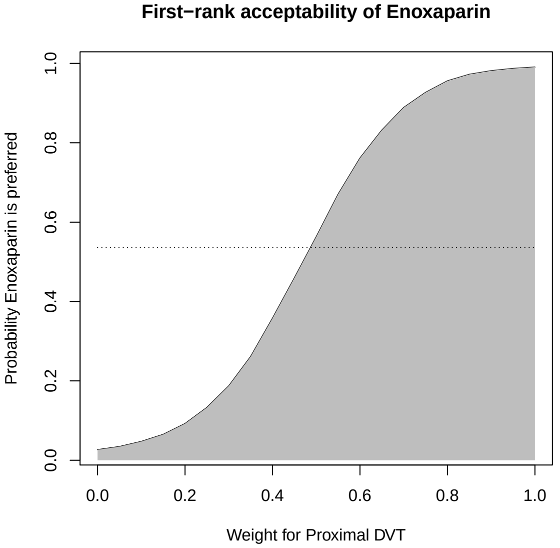


Figure 7.2: Plot of rank probability versus weight given to the benefit criterion (prevention of proximal DVT). The dotted line indicates the rank probability integrated over all feasible weights. For the analysis without preference information, the w_p varies freely (shaded area).

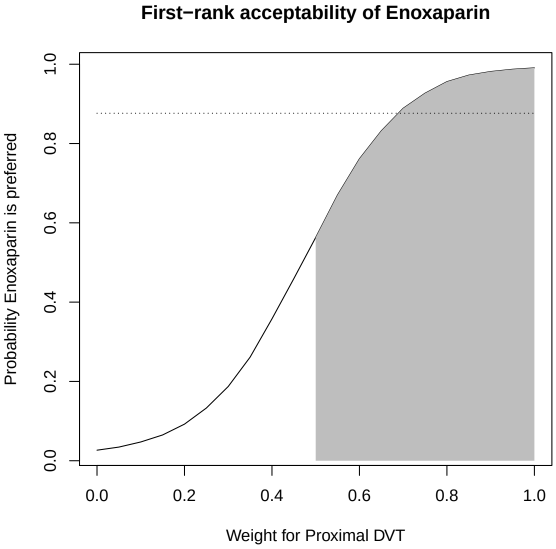


Figure 7.3: Plot of rank probability versus weight given to the benefit criterion (prevention of proximal DVT). The dotted line indicates the rank probability integrated over all feasible weights. For the analysis with ordinal preferences, $w_p > w_b$ (shaded area).

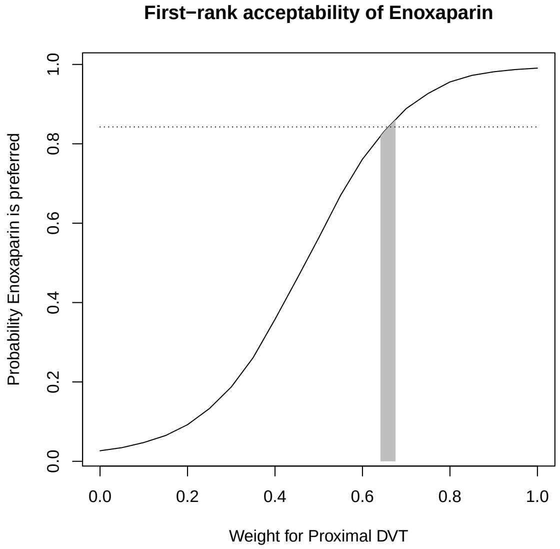


Figure 7.4: Plot of rank probability versus weight given to the benefit criterion (prevention of proximal DVT). The dotted line indicates the rank probability integrated over all feasible weights. For the analysis with ratio bound preferences, $1.79 < w_p/w_b < 2.08$.

- 0% proximal DVT, 10% major bleeding
- 25% proximal DVT, 0% major bleeding

The decision maker indicates that proximal DVT is more important, so we add the weight constraint $w_p > w_b$, which reduces the range of possible weights to the shaded area in Figure 7.3. With ordinal preferences, the first-rank acceptability of enoxaparin is 0.88. The central weight vector of enoxaparin with ordinal preferences is (0.77, 0.23), confidence factor 0.94. The confidence factor indicates that more precise preferences could potentially increase the confidence with which we are able to choose enoxaparin. In this two-dimensional case, this is obvious if we look at Figure 7.3.

To further refine the preferences, we must establish the *trade-off ratio* w_p/w_b. To do this, we apply swing weighting. In swing weighting, we compare two hypothetical scenarios:

- DVT has the worst possible value, and major bleeding has the best possible value:
$$(0.25, 0.0)$$
- DVT has an unknown value, and major bleeding has the worst possible value:
$$(x, 0.1)$$

We manipulate the value x (by asking whether the first alternative is better or worse than the second) until the decision maker is indifferent between the two. In normal swing weighting, we would be looking for an exact value of x. However, it is well known that precise weights elicited from decision makers will vary between preference elicitation sessions. Therefore, we explicitly attach uncertainty to the ratio. This can be done by decreasing and increasing x from the initially found indifference value, until one of the alternatives is likely to be preferred. In this case, x could vary between 0.11 and 0.13. This implies that $1.79 < w_p/w_b < 2.08$.

With ratio bound preferences, the first-rank acceptability of enoxaparin in 0.84, shown as a dotted line in Figure 7.4. The central weight for enoxaparin is (0.66, 0.34), confidence factor 0.84. Based on the negligible difference between the first-rank acceptability of enoxaparin and the confidence factor of its central weights, no important changes to the rank acceptability should be expected if we refine the preferences further. It is interesting to note that although using ratio bound preferences allowed us to more precisely determine the first-rank acceptability of enoxaparin, the estimated first-rank acceptability for enoxaparin was reduced. This underlines the importance of assessing stability of the first-rank acceptability over the feasible weight space, which can be achieved by inspecting the central weights and their confidence factors.

7.3 Higher-dimensional problems

The two-dimensional analysis presented above, and the original analysis [Lynd and O'Brien, 2004], take the prevention of proximal DVT to be the relevant benefit, and ignore the impact on distal DVT. However, both drugs do have an impact on distal DVT. The solution proposed in the original analysis [Lynd and O'Brien, 2004] is to combine both proximal and distal DVT into a single criterion. However, since proximal and distal DVT have different clinical implications, this makes the decision less concrete for the decision maker. Therefore, the decision problem should properly be viewed as a three-criterion problem, with proximal and distal DVT included separately.

To include distal DVT in the SMAA analysis, we set its measurement scale as 0.15 (best value) to 0.4 (worst value), by inspection of Table 7.1. With the addition of distal DVT, a third weight w_d is added to the weight vector: $\mathbf{w} = (w_p, w_d, w_b)$. In that case, the weight space is a 2-simplex in three-dimensional space, spanned by $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, as shown in Figure 7.5.

In a preference free SMAA analysis the first-rank acceptability for enoxaparin is 0.65, shown as a dotted line in Figure 7.2. The central weight vector of enoxaparin without preference information is (0.41, 0.36, 0.23), confidence factor 0.89, whereas that for heparin is (0.19, 0.28, 0.53), confidence factor 0.66. From this, it is clear that if the decision maker chooses enoxaparin, he should find the DVT criteria more important than major bleeding.

Again, we start by eliciting ordinal preferences. In this case, the criteria scale swings are ranked from most to least important as: proximal DVT $\succ$ major bleeding

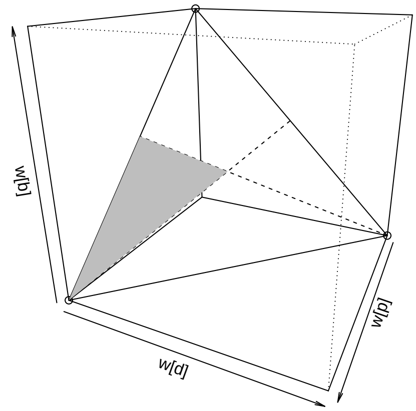


Figure 7.5: The triangle represents the full weight space. The grey polygon shows the feasible weight space given the ordinal constraints $w_1 > w_2 > w_3$ (shown as dashed lines).

$\succ$ distal DVT. The shaded region of Figure 7.5 shows the feasible weight space given these constraints. With ordinal preferences, the first-rank acceptability of enoxaparin is 0.86. The central weight vector of enoxaparin with ordinal preferences is (0.63, 0.11, 0.27), confidence factor 0.91. The confidence factor indicates that more precise preferences could potentially increase the confidence with which we are able to choose enoxaparin. On the other hand, the central weight vector of heparin is (0.52, 0.12, 0.35), confidence factor 0.19, which shows that heparin is unlikely to be preferred.

The ratio bounds for proximal DVT and major bleeding are the same as in the two-dimensional example: $1.79 < w_p/w_b < 2.08$. In addition, we elicited ratio bounds for major bleeding and distal DVT, giving $1.67 < w_b/w_d < 2.00$. These preferences are visualized in Figure 7.6. Using ratio bound preferences increases the first-rank acceptability of enoxaparin to 0.89.

7.4 Discussion

We took a published analysis of the benefit-risk profile of two anti-thrombolytics and argued that this should properly be considered a problem with three criteria: proximal DVT, distal DVT and major bleeding. The original analysis considered only proximal DVT and major bleeding because the method used is not applicable to problems with more than two criteria. Our analysis using SMAA showed that the analysis including all three criteria leads to similar results as the analysis including only two.

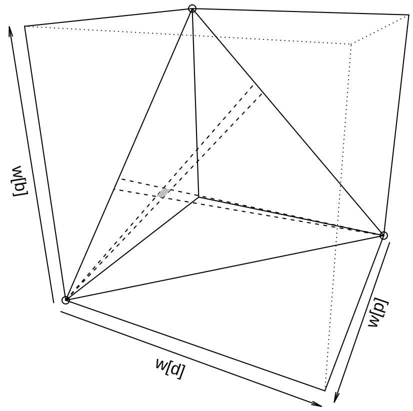


Figure 7.6: The triangle represents the full weight space. The grey polygon shows the feasible weight space given the ratio bound constraints $1.79 < w_p/w_b < 2.08$ and $1.67 < w_b/w_d < 2.00$ (shown as dashed lines).

This is due to the preferene information, which identified distal DVT as the least important criterion by a relatively large factor. Thus, this is not an a priori fact, but emerges only after we take into account the decision maker's preferences. To the authors' credit, this is a clinical judgment that they appear to have made correctly.

The case study has shown how a SMAA generalizes to multiple criteria, and how more precise preference information enabled by hit-and-run sampling can enhance the confidence with which a decision can be made. The SMAA analysis can also easily be generalized to take into account additional alternatives. In many-dimensional problems involving uncertainty, SMAA has two important advantages. First, the flexible approach to preference information allows one to incrementally refine the preferences. For example, if a ranking of the criteria from most to least important, with the addition of bounds on the trade-off ratio between the two most important criteria provides sufficient certainty on the best alternative, no further preference elicitation is needed. Secondly, by sampling from the full set of weights compatible with the preference information, a multi-dimensional sensitivity analysis is automatically carried out. The central weights and confidence factors of lower-ranked alternatives provide important information of the robustness of the alternative with highest first-rank acceptability.

In conclusion, SMAA can conceptually be seen as an extension of simple stochastic simulation techniques to many-dimensional decision problems. The strong theoretical underpinnings of MCDA ensure founded and rational decision support, while the stochastic approach of SMAA enables a natural way of assessing uncertainty and robustness. Finally, using hit-and-run weight sampling enables a flexible and iterative approach to weight elicitation.

CHAPTER 8

Multi-criteria benefit-risk assessment using network meta-analysis

G. van Valkenhoef, T. Tervonen, J. Zhao, B. de Brock, H. L. Hillege, and D. Postmus. Multi-criteria benefit-risk assessment using network meta-analysis. *Journal of Clinical Epidemiology*, 65(4):394–403, 2012e. doi: 10.1016/j.jclinepi.2011.09.005

Abstract

Objective: To enable multi-criteria benefit-risk assessment of any number of alternative treatments using all available evidence from a network of clinical trials.
Study design and setting: We design a general method for Multiple Criteria Decision Analysis (MCDA) with criteria measurements from Mixed Treatment Comparison (MTC) analyses. To evaluate the method, we apply it to benefit-risk assessment of four second-generation anti-depressants and placebo in the setting of a published peer reviewed systematic review.
Results: The analysis without preference information shows that placebo is supported by a wide range of possible preferences. Preference information provided by a clinical expert showed that while treatment with anti-depressants is warranted for severely depressed patients, for mildly depressed patients placebo is likely to be the best option. It is difficult to choose between the four anti-depressants, and the results of the model indicate a high degree of uncertainty.
Conclusions: The designed method enables quantitative benefit-risk analysis of alternative treatments using all available evidence from a network of clinical trials. The preference-free analysis can be useful in presenting the results of an MTC considering multiple outcomes.

8.1 Introduction

The pharmaceutical regulatory authorities and pharmaceutical health care decision makers increasingly request an explicit Benefit-Risk (BR) analysis of drugs as it can provide a basis for rational decisions when choosing a particular therapy [Holden, 2003]. Drug BR analysis can be used to identify trade-offs between benefit and risk, where benefit is the efficacy of a drug and risk relates to its safety [Victor and Hasford, 1987]. If there is only one measure of efficacy and one measure of safety, the BR analysis can be conducted by plotting the joint density of the benefit and risk criteria on a plane [Lynd and O'Brien, 2004]. However, there is a growing need for evidence-based pharmacotherapy to consider more than two criteria, such as multiple safety criteria, the patient's quality of life, and costs. In these cases, the two-dimensional visualization technique cannot be applied.

Multiple Criteria Decision Analysis (MCDA) methods can help by structuring the decision problem and making the underlying value trade-offs explicit. Specifically, Tervonen et al. [2011] proposed a Stochastic Multicriteria Acceptability Analysis (SMAA) model for analyzing BR. Their model allows taking into account the probability distributions of the criteria measurements and is able to quantify the uncertainty surrounding a decision. Moreover, measurements and value judgments (preferences) are clearly separated. However, the model relies on a single trial to evaluate the comparative BR profiles of the alternatives. In most cases, a BR assessment will need to be based on evidence synthesized from multiple trials or possibly a complex network of trials.

Although evidence synthesis is most often done through pair-wise meta-analyses, they are ill-suited as a basis for a computational BR method for a number of reasons. First, relative effects have to be assessed against a common comparator, and not all evidence structures have a single treatment against which all others are compared [Salanti et al., 2008b]. Second, choosing a common comparator introduces a selection bias by excluding studies that do not include the comparator. Sensitivity analyses would have to be carried out for every possible choice of comparator and even then some studies might be excluded. Finally, when a large number of treatments is available, the majority of evidence may be indirect regardless of the chosen common comparator. Traditional meta-analysis does not allow these indirect comparisons to be taken into account.

The recently proposed Mixed Treatment Comparison (MTC) method (also known as network meta-analysis) synthesizes all the available evidence through application of a Bayesian evidence network [Salanti et al., 2008a, Lu and Ades, 2004]. The relative effects of all included treatments are estimated using both direct and indirect evidence. In this way, the results are consistent regardless of the chosen comparator, and it is not necessary that one of the treatments has been compared with all others. Graphical summaries of MTC results have been proposed as an informal decision aid in trading effectiveness against other factors [Salanti et al., 2011]. To enable the formal BR analysis of a number of alternative treatments taking into account all relevant studies, this paper proposes to apply MTC for evidence synthesis in SMAA-based multi-criteria drug BR analysis. We call this method MTC/SMAA, and for illustration, we constructed a model to evaluate the comparative BR profiles of four second-generation anti-depressants and placebo using 25 studies from the literature, selected on the basis of an existing systematic review [Hansen et al., 2005].

8.2 Stochastic Multicriteria Acceptability Analysis

SMAA-2 [Lahdelma and Salminen, 2001] considers a discrete, multi-criteria decision problem consisting of a set of m alternatives that are evaluated in terms of n criteria. The vector of criteria measurements corresponding to alternative i is denoted by $\xi^i = (\xi^i_1, \ldots, \xi^i_n)$, where ξ^i_k is a random variable representing the performance of alternative i on criterion k, modeled using some density function. For each criterion, a partial value function $v_k(\xi^i_k)$ is defined to normalize the criteria measurements, so that they are represented by values between zero (the worst value) and one (the best value). The overall value function is then defined as a weighted additive combination of the partial value functions:

$$v(\xi^i, \mathbf{w}) = \sum_{k=1}^{n} w_k \cdot v_k(\xi^i_k) \ ,$$

where $v(\xi^i, \mathbf{w}) > v(\xi^j, \mathbf{w})$ implies that alternative i is preferred to alternative j given the weight vector $\mathbf{w}$. The weights define relative importances of the scale swings (changes from the worst to the best criterion values), and $w_k > w_l$ implies that if the Decision Maker (DM) would have to choose between improving either criterion k or criterion l from the worst to the best value, he or she would increase the performance on criterion k.

The DM's preferences may be unknown or partially known, and therefore the weights $\mathbf{w}$ are also represented by a probability density. Total lack of preference information is represented by a uniform distribution in the feasible weight space. Partial information, such as importance ranking of the criteria, can easily be included by restricting the feasible weight space accordingly [Lahdelma and Salminen, 2001].

For given (exact) values of ξ and $\mathbf{w}$, the rank of each alternative is defined as an integer from the best rank ($= 1$) to the worst rank ($= m$) by means of a ranking function $\mathrm{rank}(i, \xi, \mathbf{w})$. The main decision aiding measure is the *rank acceptability index*, denoted by b^r_i. It describes the share of all possible values of the weight vector $\mathbf{w}$ and criteria measurements ξ for which $\mathrm{rank}(i, \xi, \mathbf{w}) = r$. For example, $b^5_2 = 0.3$ means alternative 2 has 5th-rank acceptability 0.3. The preferred (best) alternatives are those with high acceptabilities for the best ranks.

Instead of using the value function to rank the alternatives for an elicited weight vector $\mathbf{w}$, which is the traditional approach in multi-attribute value theory, the SMAA methods allow computing the weights a 'typical' DM supporting each alternative might have. This so-called *central weight vector* w^c_i can be presented to the DM to help him or her understand what kind of weights would favor a certain alternative i. The *confidence factor* p^c_i is the probability for alternative i to obtain the first rank when its central weight vector is chosen. The confidence factors indicate whether the criteria measurements are sufficiently accurate to discern the efficient alternatives.

Low confidence factors (< 0.50) should be interpreted with care, as then even if a DM finds the central weight vector corresponding to his or her preferences, there might be another alternative that achieves a higher first rank acceptability with those weights.

8.3 Mixed treatment comparison

The MTC method (also called network meta-analysis) synthesizes all available clinical evidence through application of a Bayesian hierarchical model [Lu and Ades, 2004, Salanti et al., 2008a]. It enables the detection of heterogeneity (differences in studies comparing the same treatments) and inconsistency (differences between direct and indirect comparisons) in the evidence [Salanti et al., 2008a, Lu and Ades, 2006, Dias et al., 2010]. In this section, we briefly introduce the structure of a random effects MTC model for dichotomous data, as this type of model will be used in the case study (Section 8.5). For other model types and the handling of multi-arm trials, we refer to [Salanti et al., 2008a, Lu and Ades, 2006].

Let i be a clinical trial. For each included treatment x we are given the sample size $n_{i,x}$ and the number of events $r_{i,x}$, modelled as a binomial process:

$$r_{i,x} \sim \mathrm{Bin}(p_{i,x}, n_{i,x}) \ ,$$

where $p_{i,x}$ is the success probability (i.e. the *absolute risk* of an event). The risk $p_{i,x}$ of an effect observed in the individual studies is transformed to log odds $\theta_{i,x}$ through:

$$\theta = \mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) \ .$$

The inverse transformation is given by:

$$p = \mathrm{logit}^{-1}(\theta) = \frac{1}{1+e^{-\theta}} \ .$$

The advantage of this transformation, also used in logistic regression, is that $\theta_{i,x}$ can be assumed to be normally distributed. Moreover, if $\theta_{i,x}$ and $\theta_{i,y}$ are the log odds for x and y, then $\theta_{i,x} - \theta_{i,y}$ is the log odds ratio of y compared to x in trial i (and $e^{\theta_{i,y}-\theta_{i,x}}$ is the odds ratio).

Synthesis in MTC models is done in terms of treatment contrasts (relative effects) and not the absolute effects, as this leads to a more robust model that preserves the randomization in the trials [Lu and Ades, 2004]. To do this, we choose a baseline treatment $b(i)$ for every trial i, and express the effect of $b(i)$ as:

$$\theta_{i,b(i)} = \mathrm{logit}(p_{i,b(i)}) = \mu_i \ ,$$

and for every other treatment $y \neq b(i)$ the effect is:

$$\theta_{i,y} = \mathrm{logit}(p_{i,y}) = \mu_i + \delta_{i,b(i),y} \ ,$$

where $\delta_{i,b(i),y}$ is the random effect of y relative to $b(i)$ in trial i. The random effects are related to the *relative effect* as follows:

$$\delta_{i,x,y} \sim \mathcal{N}(d_{x,y}, \sigma^2_{x,y}) \ ,$$

where $d_{x,y}$ is the relative effect of y compared to x, the parameter of interest, and $\sigma^2_{x,y}$ is the *random effects variance*. If we set $\sigma^2_{x,y}$ to be identical for all x and y, $\sigma^2_{x,y} = \sigma^2$, the model is a homogeneous variance model. Otherwise it is a heterogeneous variance model.

The model discussed so far is just a Bayesian formulation of pair-wise random effects meta-analysis. MTC enables the simultaneous synthesis of a network of trials through the additional assumption of *consistency*. Suppose we have three treatments, say A, B, and C, and studies comparing AB, AC, and BC. The consistency assumption then defines the relation between the relative treatment effects as

$$d_{AC} = d_{AB} + d_{BC} \quad .$$

A model that includes this assumption between all relative effects is a consistency model. Conclusions based on an MTC model are always derived using the consistency model. The model is estimated through stochastic simulation, e.g. using the BUGS [Spiegelhalter et al., 2003] or JAGS [Plummer, 2009] software. This enables the derivation of a point estimate and 95% credibility interval (CrI, the Bayesian analog to a confidence interval) for each of the relative effects, as well as the derivation of any other statistics of interest.

The assumption of consistency may be violated by the data at hand, in which case there exists *inconsistency*. As with pair-wise meta-analyses, the first step in dealing with inconsistency should be assessing whether the included studies are sufficiently similar to be combined. Statistical means of detecting inconsistency provide an additional safeguard against drawing conclusions from inconsistent datasets, though the lack of demonstrable inconsistency does not prove that the results are free of bias and diversity.

There are two competing methods for detecting inconsistency: inconsistency models [Lu and Ades, 2006] and node splitting models [Dias et al., 2010]. Inconsistency models assess inconsistency by adding inconsistency factors to closed loops in the evidence graph, whereas in node splitting models a single comparison is chosen for which the direct and indirect evidence are contrasted. Inconsistency models have the advantage that only a single model needs to be run, but the results are often difficult to interpret. Node splitting models are easier to interpret, but require a different model to be run for each of the potentially inconsistent comparisons. Which method should be preferred is not yet clear and, in this paper, we will present the results of the node-splitting analysis because they are easier to interpret and verify them with an inconsistency model.

Inconsistency within an evidence network could reflect genuine diversity, bias or a combination of both [Salanti et al., 2008a]. If there is inconsistency, the reason for the inconsistency must be determined, and a clinically sound explanation must be given. If the explanation is sufficient, the offending studies are removed [Lu and Ades, 2006], a new inconsistency model is constructed and inconsistency evaluation is repeated until no relevant inconsistency remains. If there is considerable inconsistency that cannot be eliminated, the consistency model cannot be used. It is difficult to judge whether a certain amount of inconsistency should be considered relevant, and the debate on how to do this is ongoing [Salanti et al., 2008a, Lu and Ades, 2006, Dias et al., 2010].

8.4 MTC/SMAA for BR analysis

The process of performing an MTC/SMAA analysis is shown in Figure 8.1. Analyzing BR based on clinical studies starts with a systematic review of the available studies relevant to the clinical domain for which BR should be assessed. In this step, which should be carried out with experts in the clinical domain, the relevant studies and important issues are identified. In the ideal case, a relevant high-quality systematic review can be found in the literature. Based on the review, the criteria to be considered are agreed upon and operationalized. Then, for each criterion, the relevant outcomes are extracted from the individual studies and inconsistency is evaluated. If there is no relevant inconsistency, a consistency model is subsequently constructed and used to create the measurements for the SMAA model. If no reasonable explanation of inconsistency is found, the whole process has to be terminated.

8.4.1 Measurement scales

For reasons of statistical robustness, evidence synthesis methods estimate only relative effects, while absolute measures are more suitable for applying evidence to concrete decisions [Egger et al., 1997]. In a multi-criteria model, the use of absolute measures is desirable since explicit trade-offs must be made between unit increases in the scaled criteria. The problem is especially salient for dichotomous criteria, as the result in these cases is expressed as an odds ratio, which is difficult to interpret when assessing the relative importance of the scale swings between criteria. To solve this, the log odds ratio can be converted to (absolute) risk by assuming a distribution for the log odds of a baseline treatment 1:

$$\theta_1 \sim \mathcal{N}(\mu, \sigma^2) \ .$$

Note that θ_1 is an overall estimate for treatment 1, and should not be confused with the trial-level log odds $\theta_{i,1}$. It does not matter which of the m included treatments is selected as the baseline. For every non-baseline treatment $j \neq 1$, the MTC analysis gives us the log odds ratio:

$$\begin{pmatrix} d_{1,2} \\ \vdots \\ d_{1,m} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \nu_2 \\ \vdots \\ \nu_m \end{pmatrix}, \Sigma \right) \ ,$$

which can be used to obtain the distribution of the non-baseline treatments' log odds conditional on θ_1:

$$\begin{pmatrix} \theta_2 \\ \vdots \\ \theta_m \end{pmatrix} \mid \theta_1 \sim \mathcal{N}\left(\begin{pmatrix} \theta_1 + \nu_2 \\ \vdots \\ \theta_1 + \nu_m \end{pmatrix}, \Sigma \right) \ .$$

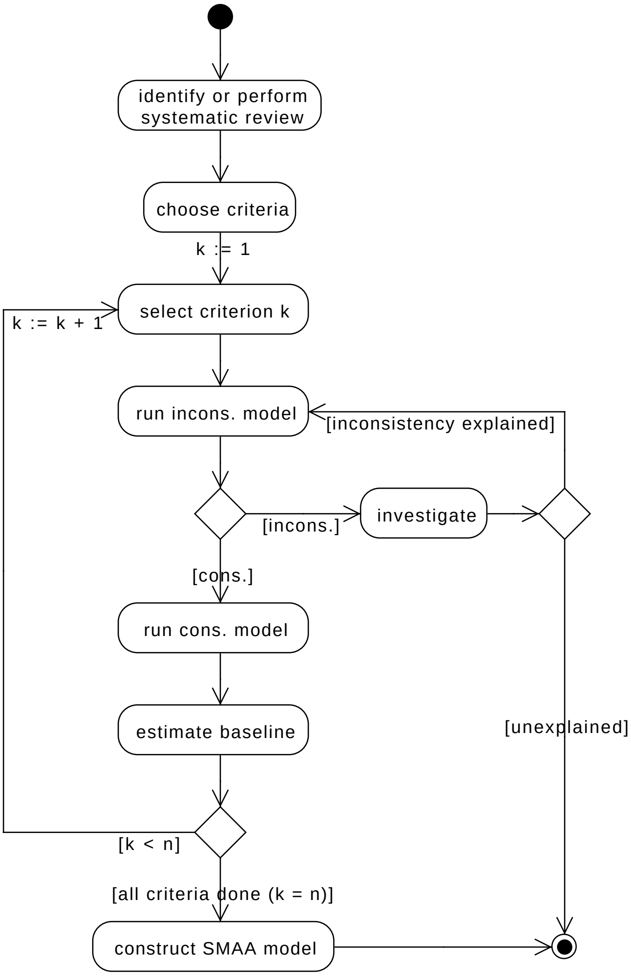

identify or perform systematic review
choose criteria
k := 1
select criterion k
k := k + 1
run incons. model
[inconsistency explained]
[incons.]
investigate
[cons.]
run cons. model
estimate baseline
[unexplained]
[k < n]
[all criteria done (k = n)]
construct SMAA model

Figure 8.1: The process of performing an MTC/SMAA analysis (UML activity diagram notation). n is the number of criteria

Then, for any treatment i the risk is

$$p_i = \text{logit}^{-1}(\theta_i) \ ,$$

as discussed in Section 8.3. The p_i are the measurements used in the SMAA analysis (thus $\xi_k^i = p_i$, where p_i is obtained for criterion k). In the SMAA simulation, to obtain samples of the p_i's, we first sample the baseline log odds θ_1 and then sample the log odds θ_i for all other alternatives based on θ_1, and transform them to risk, as given above. Note that ranking the treatments based on the p_i is equivalent to ranking them based on the $d_{1,i}$ (with $d_{1,1} = 0$), and will thus result in the same rank probabilities as from the MTC analysis if the $d_{1,i}$ accurately reflect the posterior distribution. The rank probabilities in MTC [Salanti et al., 2011] are calculated for a single criterion and are therefore distinct from the rank acceptabilities discussed in Section 8.2, which incorporate trade-offs between multiple criteria.

Different methods can be used to arrive at a sensible assumption for the baseline log odds θ_1. One could use an observational effectiveness study with a suitable population, let a clinical expert provide estimates, or attempt to derive them from the included trials. In this paper, we will apply arm-based pooling of the placebo arms. This is supplemented by a visual assessment (through a forest plot) of the effects found in the individual studies.

Since the risk scale is bound to $[0, 1]$, either $v_k(\xi_k^i) = \xi_k^i$ or $v_k(\xi_k^i) = 1 - \xi_k^i$ can be used as the partial value function v_k for any dichotomous criterion k, respectively when more or less events are preferred (see Section 8.2). We will return to the advantages and disadvantages of this approach in the discussion.

8.5 Application to anti-depressants

To illustrate the use of MTC/SMAA, we used an existing systematic review [Hansen et al., 2005] to create a model for evaluating the comparative BR profiles of four second-generation anti-depressants (fluoxetine, paroxetine, sertraline and venlafaxine) and placebo. The application is meant as an example, and the results should be interpreted with care. A full BR analysis of anti-depressants should ideally be based on a more recent systematic review that explicitly includes placebo-controlled studies. This is even more important in the light of recent doubt on the efficacy of anti-depressants [Pigott et al., 2010]. Even if we consider the efficacy of anti-depressants to be proven, in the context of a multi-criteria decision model consideration of other factors may imply that placebo is the best option, as the placebo response in depression trials is considerable [Storosum et al., 2004].

8.5.1 Previous work

The review included 46 studies comparing 10 second-generation anti-depressants on the Hamilton Rating Scale for Depression (HAM-D) or Montgomery-Asberg Depression Rating Scale (MADRS). In total, 20 comparisons were made in the included studies (out of 45 possible comparisons). Meta-analysis was applied for just 3 comparisons using 16 studies in total. All meta-analyses assessed efficacy (50% or greater

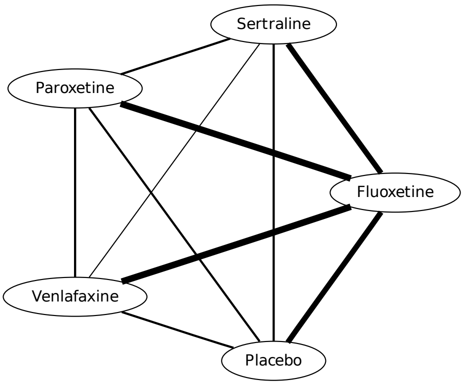


Figure 8.2: Evidence network of studies comparing the four included second-generation anti-depressants and placebo. The width of the lines indicates the number of studies that include that comparison (the minimum is 1 and the maximum 6)

improvement from baseline on the HAM-D or MADRS scale) relative to fluoxetine, and studies between the other drugs (paroxetine, sertraline and venlafaxine) were not considered. Meta-analysis yielded risk ratios relative to fluoxetine, with a significant but small additional effect for sertraline and venlafaxine. The authors concluded that the four anti-depressants did not differ substantially for treatment of major depressive disorder. A more recent review [Cipriani et al., 2009] used an MTC analysis to show that there are differences among second-generation anti-depressants in terms of efficacy and the proportion of patients completing the study.

8.5.2 Methods

An MTC/SMAA analysis was performed to compare fluoxetine, paroxetine, sertraline, venlafaxine, and placebo on one benefit criterion (efficacy) and five risk criteria. Efficacy was assessed by means of treatment response, defined as a 50% or greater improvement on the HAM-D rating scale for depression. The five risk criteria corresponded to the most common Adverse Drug Reactions (ADRs): diarrhea, dizziness, headache, insomnia, and nausea. All of the criteria were measured in terms of absolute risk, based on dichotomous data from the included trials.

As Hansen et al. [2005] did not include sufficient information to construct the MTC models, we did not take the measurements directly from the review, but used the included individual studies to perform our own analysis. Although the review did not consider placebo, sufficient studies with a placebo arm were present to include it in the analysis. The papers included in the review were retrieved and the data extracted. We used the drugis.org MTC software (`http://drugis.org/mtc`) [van Valkenhoef et al., 2012d] to generate MTC models for the 25 studies (see Ta-

Criterion	Placebo	Fluoxetine	Paroxetine	Sertraline	Venlafaxine	Total
HAM-D	8	18	9	8	9	24
Diarrhea	5	11	7	8	5	17
Dizziness	3	9	4	5	6	12
Headache	5	12	8	8	6	19
Insomnia	7	12	8	6	6	18
Nausea	6	15	9	8	8	22
Total	8	18	10	9	9	25

Table 8.1: The number of studies included in the network meta-analysis for each criterion

ble 8.1 and Figure 8.2) comparing fluoxetine, paroxetine, sertraline, venlafaxine, and placebo. We used the homogeneous variance assumption and specified a uniform prior $\sigma \sim \mathcal{U}(0, 4)$ for the random effects variance. For the trial baseline effects μ_i and random effects $\delta_{i,b(i),y}$ we specified a $\mathcal{N}(0, 10^3)$ prior. Markov Chain Monte Carlo simulation with 4 parallel chains of 30,000 tuning and 20,000 simulation iterations each was used to estimate each MTC model, and the computations were done using JAGS [Plummer, 2009] and R [R Development Core Team, 2008]. Inconsistency was primarily assessed using node-splitting models [Dias et al., 2010] and inconsistency models [Lu and Ades, 2006] were run as a secondary analysis. Convergence was assessed using the Brooks-Gelman-Rubin diagnostic [Brooks and Gelman, 1998], where a potential scale reduction factor of 1.05 or lower was considered sufficient if visual inspection of the convergence plots and time-series also indicated convergence.

We constructed a SMAA model with the measurements derived from the consistency models, and baseline estimates derived from the trials and discussed with an expert. The SMAA model was computed using R with 10,000 Monte Carlo iterations giving sufficient accuracy for the indices [Tervonen and Lahdelma, 2007]. The SMAA analyses were performed for three scenarios: one with missing preference information and two with a criteria ranking elicited from the expert: mild and severe depression. The data files are available online at `http://drugis.org/network-br`. There we also provide a JSMAA [Tervonen, 2010] v0.8.4 model that allows the reader to explore the trade offs in an interactive graphical user interface.

8.5.3 Results

Inconsistency analysis The node-splitting analysis of inconsistency revealed two potential problems at the $\alpha = 0.05$ significance level, though given that there were 56 comparisons, it is to be expected that some are significant due to chance. However, we chose not to correct the threshold a priori, but rather to investigate these two cases. One occurred in the headache network, where one split node was significant, and the other in the nausea network, where two directly related split nodes were significant. In neither of these cases could we identify any systematic differences between the studies, and as the number of significant findings is compatible with chance, we decided to continue on the basis of consistency models including all studies. The studies

Criterion	Parameters	Risk (95% CrI)
HAM-D	-0.17 ± 0.11	0.46 (0.40, 0.51)
Diarrhea	-2.19 ± 0.21	0.10 (0.07, 0.14)
Dizziness	-2.23 ± 0.61	0.10 (0.03, 0.26)
Headache	-1.20 ± 0.29	0.23 (0.15, 0.35)
Insomnia	-2.61 ± 0.19	0.07 (0.05, 0.10)
Nausea	-2.02 ± 0.19	0.11 (0.08, 0.16)

Table 8.2: Baseline measurements derived from the placebo trials, given as mean ± standard error for the log-odds and the corresponding median and 95% CrI of the resulting logit-normal distribution for the absolute risk

involved in these comparisons did not lead to inconsistencies in the other evaluated networks, and the secondary analysis using inconsistency models did not indicate any inconsistencies.

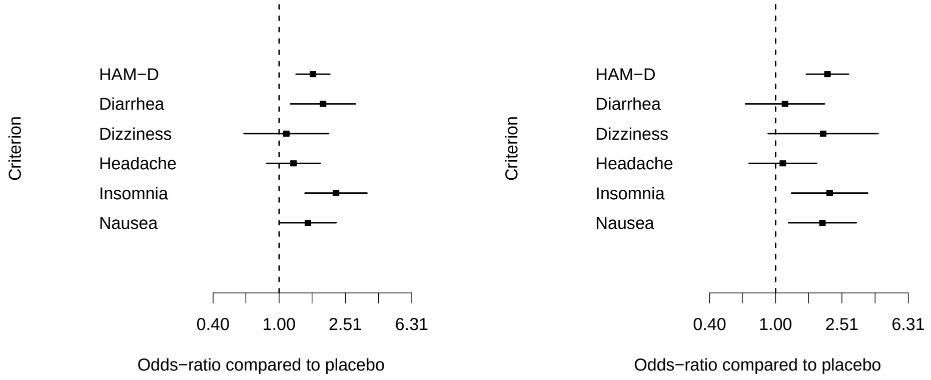
Consistency analysis The results of the consistency analysis are visualized as forest plots for the odds ratio relative to placebo in Figure 8.3. Including indirect evidence leads to somewhat smaller 95% credibility intervals for treatment response than pairwise meta-analysis. Therefore the evidence from the studies additionally included in the MTC model discriminate the drugs better with respect to efficacy.

Preference-free model Baseline estimates were derived by random-effects pooling of the placebo arms (Table 8.2) and discussed with an expert, who compared them to sources known to him and did not contest the values or the method used to derive them. He did note that these values are expected to vary greatly between trials, and that this fact is reflected in the width of the confidence intervals.

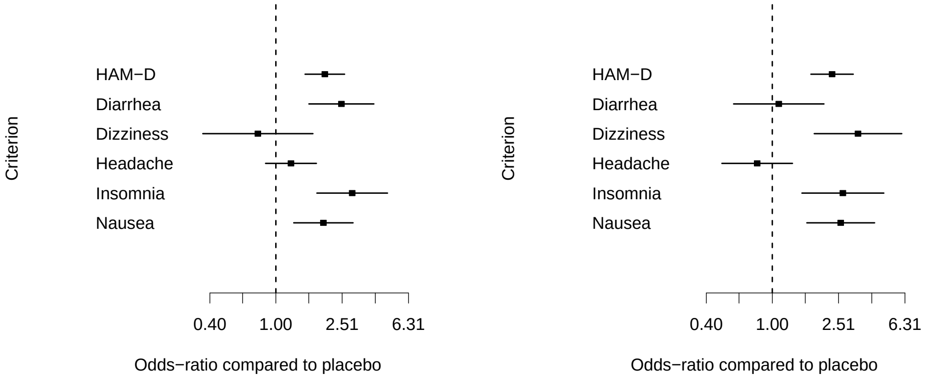
The rank-acceptabilities with missing preferences are shown in Figure 8.4. There is a large share of the possible preferences for which placebo attains rank 1. From the central weights (Figure 8.5), it is estimated that the preference scenarios that are favorable to placebo have a low weight for efficacy, and that a 'typical' DM that would choose placebo implictly finds each of the ADRs to be about twice as important as efficacy. Placebo is also the only alternative to attain a confidence factor close to 1 (Table 8.5).

Fluoxetine has a low confidence factor (0.12) for its central weights, and in fact given its central weights, other alternatives have a higher first-rank acceptability. Thus, fluoxetine is likely to be dominated by the other alternatives. In general, if efficacy is highly valued, placebo is unlikely to be the best option, but it is difficult to choose a drug based on the data.

Mild depression Preferences for the mild depression scenario were elicited from the expert using ordinal swing weighting. This resulted in the following ranking of the criteria: Insomnia $\succ$ HAM-D $\succ$ Dizziness $\succ$ Nausea $\succ$ Diarrhea $\succ$ Headache. The rank acceptabilities for this scenario are shown in Figure 8.6. Placebo obtains the highest first-rank acceptability (0.56), followed by paroxetine (0.28), while venlafaxine has



(a) Fluoxetine (b) Paroxetine



(c) Sertraline (d) Venlafaxine

Figure 8.3: Network meta-analysis results: odds ratios relative to placebo, with 95% CrI. Results to the right of the no-effect line indicate a higher incidence for the active treatment

Alternative	CF	HAM-D	Diarrhea	Dizziness	Headache	Insomnia	Nausea
Fluoxetine	0.12	0.21	0.09	0.28	0.12	0.09	0.20
Paroxetine	0.57	0.30	0.18	0.12	0.12	0.14	0.14
Placebo	0.99	0.09	0.18	0.18	0.16	0.20	0.20
Sertraline	0.55	0.28	0.08	0.30	0.13	0.10	0.12
Venlafaxine	0.63	0.29	0.18	0.08	0.25	0.12	0.09

Table 8.3: Central weights and confidence factors (CFs) for the preference-free model

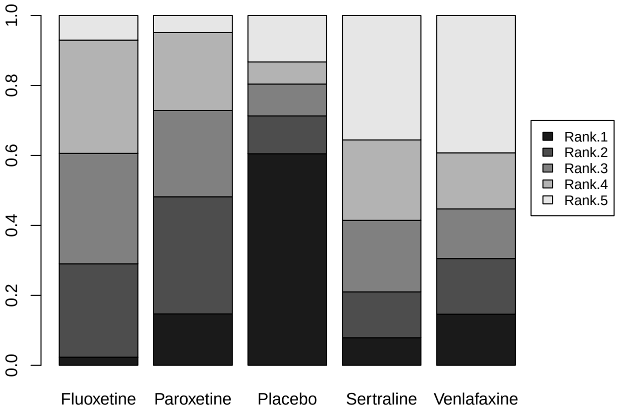


Figure 8.4: Rank acceptabilities for the preference-free model

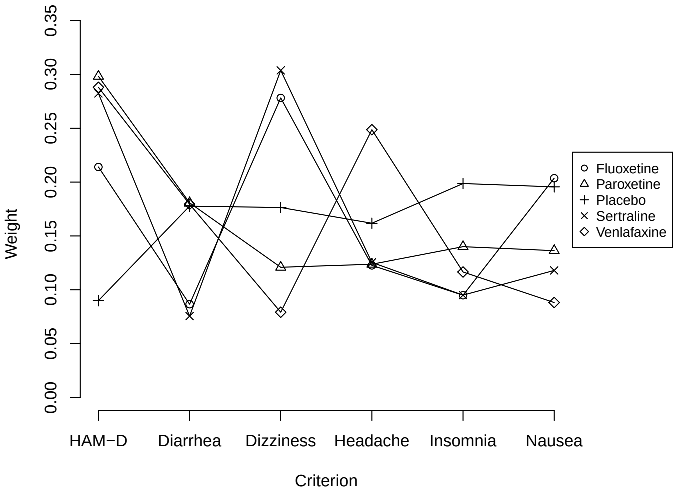


Figure 8.5: Central weights for the preference-free model

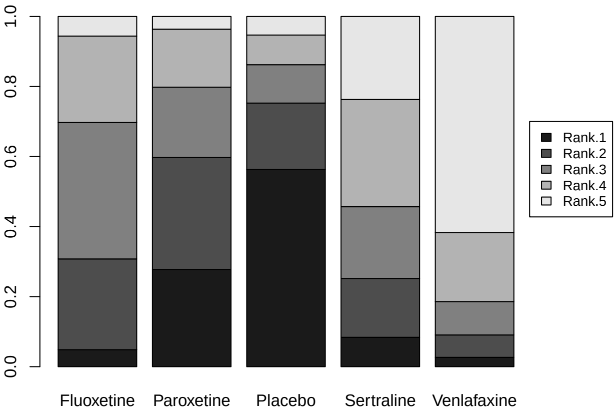


Figure 8.6: Rank acceptabilities for the mild depression scenario

the highest last-rank acceptability (0.62), followed by sertraline (0.24). Clearly, the high incidence of both insomnia and dizziness are unfavorable to venlafaxine given the preferences. Only placebo, fluoxetine, and paroxetine have > 0.5 probability of being among the best 3, and only placebo and paroxetine have < 0.5 probability of being among the worst 3.

Severe depression The preferences elicited for this scenario differed only in that the Insomnia and HAM-D criteria were swapped. The rank acceptabilities for severe depression are shown in Figure 8.7. As would be expected based on the central weights analysis with missing preferences, ranking HAM-D as the most important criterion reverses the situation for placebo, which now has only 0.09 first-rank acceptability, and 0.56 last-rank acceptability. Placebo is also the only alternative to have < 0.5 probability of being among the best 3. Paroxetine has the highest first-rank acceptability (0.47) and paroxetine and sertraline are the only alternatives that have < 0.5 probability of being among the worst 3.

8.6 Discussion

Pharmacological decision making is a complex domain in which decisions regarding multiple criteria are informed by complex evidence networks consisting of heterogeneous clinical studies. This paper introduced MTC/SMAA, which uses the MTC evidence synthesis method together with SMAA to assess multi-criteria BR trade-offs while taking into account all available evidence from clinical trials.

The MTC/SMAA method has four main advantages. First, MTC/SMAA allows taking into account all the available evidence no matter whether the treatments are

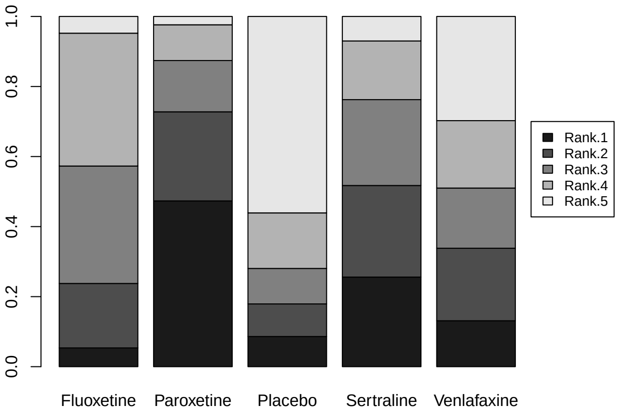


Figure 8.7: Rank acceptabilities for the severe depression scenario

directly or indirectly related. Second, a group of treatments without a common comparator can be analyzed, which is impossible with pair-wise evidence synthesis methods. Third, inconsistencies in the evidence structure due to incompatible study design can be detected early in the analysis and systematically removed if the inconsistency is judged to be clinically relevant. Fourth, application of SMAA enables explicit assessment of trade-offs that exist between the criteria and provides valuable insights even if the DMs are not willing or able to provide exact preferences.

8.6.1 Case study

We illustrated the MTC/SMAA method with a case study on second-generation anti-depressants. Although the case study is indicative of the method's feasibility, further work should evaluate the model in other therapeutic groups.

As we demonstrated in the case study, a preference-free analysis of the central weight vectors can provide substantial insight into trade-offs between the treatments under consideration. As such, a SMAA central weights analysis of the most important outcomes could be a valuable addition to any (network) meta-analysis. It allows drawing firmer conclusions on which treatments are likely to be most suited to specific situations, and which treatments are unlikely to be the best in any situation. The mild and severe preference scenarios showed that for severe depression, treatment with an anti-depressant is warranted, but for mild depression this is not clear. Recent research suggests that placebo may be effective even without deception [Kaptchuk et al., 2010] (in irritable bowel syndrome), so it may be worthwhile to explore this option for mildly depressed patients. The analyses also suggest that fluoxetine is unlikely to be the best among the five alternatives.

However, the data do not conclusively distinguish the alternatives, especially the active treatments, and given the amount of data it is likely that much of this uncertainty is inherent to the field, especially when distinguishing the active treatment options. Some improvement may be possible by eliciting more precise weights. However, except for placebo in the mild depression scenario, making the weights more precise within the constraints imposed by the ordinal preferences elicited from the expert would not allow much more conclusive results as the data have a high degree of uncertainty.

Compared to the systematic review on which we based the case study, the MTC/-SMAA analysis explicitly takes into account the ADRs in addition to efficacy and gives a clearer picture of the strengths and weaknesses of the alternatives. Including placebo in the analysis provides further insight into the trade-offs. Moreover, the model can quantitatively support the statement, also made in the original review, that it is difficult to choose among the four considered anti-depressants.

8.6.2 Limitations and future work

The main challenges in applying MTC/SMAA are the evaluation of inconsistency and estimation of baseline effects. Assessing inconsistency is especially difficult in cases where many potential inconsistencies have to be considered (large evidence networks or many different criteria) since significant results may also arise by chance. Different methods to assess inconsistency have been proposed [Lu and Ades, 2006, Salanti et al., 2008a, Dias et al., 2010], and general consensus on the best method has not yet been reached. The second concern is the scale employed for the criteria measurements. We developed a procedure for converting the relative scales from evidence synthesis to absolute ones to be used in decision making using minimal information. However, baseline effects have to be estimated, and further work is necessary to identify the best way to do this.

Another consideration is the scale on which criteria are evaluated in preference elicitation. In contrast to the previous work on SMAA for BR analysis [Tervonen et al., 2011], we choose to use the full $[0, 1]$ scale instead of the hull of the 95% confidence intervals. This has the advantage that trade offs are easier to evaluate, and that introducing additional alternatives does not require re-eliciting the preferences. The disadvantage of this approach is that a stronger linearity assumption on the partial value functions is required [see Tervonen et al., 2011]. This limitation is especially important when the observed frequencies differ greatly, e.g. when a trade off between high efficacy and rare but serious adverse events needs to be made. In those cases the scales should be assessed using the confidence interval hull. Of course, for scales that do not have natural bounds (e.g. weight gain in kg) the confidence interval hull approach is the only viable option.

In the current work we applied a SMAA decision model based on additive value functions. Although the additive model is widely applied and reasonably easy to understand, we acknowledge that other approaches are possible. For example, Data Envelopment Analysis (DEA) models have been commonly applied in cost-benefit analyses outside the area of healthcare, and there is also a SMAA variant for DEA, the SMAA-D [Lahdelma and Salminen, 2006b]. Future work should assess whether other simulation-based decision models are applicable in the context of drug BR analysis.

Finally, our model is based only on criteria that are measured in clinical trials, which is appropriate in the context of health policy decision making. However, other criteria may need to be considered, such as cost in reimbursement decisions, or the route of administration in prescription decisions. While we did not consider such criteria, they would not be difficult to include in an MTC/SMAA analysis.

CHAPTER 9

ADDIS: a decision support system for evidence-based medicine

G. van Valkenhoef, T. Tervonen, T. Zwinkels, B. de Brock, and H. Hillege. ADDIS: a decision support system for evidence-based medicine. *Decision Support Systems*, 2012f. doi: 10.1016/j.dss.2012.10.005. (in press)

Abstract

Clinical trials are the main source of information for the efficacy and safety evaluation of medical treatments. Although they are of pivotal importance in evidence-based medicine, there is a lack of usable information systems providing data-analysis and decision support capabilities for aggregate clinical trial results. This is partly caused by unavailability (i) of trial data in a structured format suitable for re-analysis, and (ii) of a complete data model for aggregate level results. In this paper, we develop a unifying data model that enables the development of evidence-based decision support in the absence of a complete data model. We describe the supported decision processes and show how these are implemented in the open source ADDIS software. ADDIS enables semi-automated construction of meta-analyses, network meta-analyses and benefit-risk decision models, and provides visualization of all results.

9.1 Introduction

Two kinds of decision support systems for evidence-based medicine can be distinguished: rule-based systems for supporting operational decisions of practising physicians and strategic decision support systems. The rule-based systems represent clinical knowledge and include inference rules for aiding professional decision making in clinical practice. They have been in existance since the 1970s [Shortliffe and Buchanan, 1975]. The most common of these are Computerized Physician Order Entry (CPOE) systems which contain evidence-based rules that enable issuing warnings when an inappropriate combination of medicines is prescribed. To the best of our knowledge, there are no established systems that inform strategic (rather than operational) decisions such as identifying the best treatment practices based on the consideration of benefit-risk trade-offs.

Strategic health care decision making, with or without a supporting system, depends heavily on the availability of unbiased evidence from controlled clinical trials [Evidence-Based Medicine Working Group, 1992]. One of the core activities and sources of information in evidence-based medicine is the systematic review [Sutton et al., 2009], a literature review that attempts to identify and synthesize all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question [Higgins and Green, 2009]. Currently the process of systematic review is extremely labor intensive and error prone due to the lack of a comprehensive source of clinical trials, the inaccuracy of literature searches, interpretation issues, tedious manual data extraction and, importantly, the duplication of effort that is necessary for every review [Sim et al., 2000]. The emergence of clinical trial registries [Zarin et al., 2007] and the move towards a more open clinical research community [Sim et al., 2006, Ghersi et al., 2008], as well as the initiatives of the Cochrane foundation [Grimshaw et al., 2006] to share and update meta-analysis data sets offer opportunities for more efficient approaches to evidence synthesis. Still, to date there is no single complete collection of performed clinical trials and outcome data, and importantly none of the available sources store results in a format that is suited for re-analysis [Wood, 2009, Zarin et al., 2007].

Thus, although suitable methods for evidence-based strategy decision support exist [Tervonen et al., 2011, Coplan et al., 2011, Mussen et al., 2007, van Valkenhoef et al., 2012e], evidence-based decision making is difficult to implement because of the substantial effort required to systematically review the literature for relevant studies and to manually extract the data from these studies, which has to be done on a case by case basis. Even when a relevant published systematic review exists, evidence-based decision making including multiple (possibly conflicting) objectives is difficult and in practice often done ad hoc due to a lack of supporting information technology. In addition, sometimes it will be necessary to incorporate additional studies to the body of evidence present in the systematic review, e.g. in the regulatory context where the manufacturer sponsors studies to prove the efficacy and safety of a newly developed drug. Moreover, the analyses reported in the published systematic review may not be valid for the decision at hand, so re-analysis of the included clinical trials may be needed. Text-based reports of systematic reviews do not support such use cases.

There do exist methods for automated extraction of trial design and results from the literature, but although the field is rapidly evolving [see e.g. Kiritchenko et al., 2010], their accuracy is not yet sufficient to be directly used in systems supporting strategic decisions.

In this paper, we present ADDIS (Aggregate Data Drug Information System, `http://drugis.org/addis`), an open source evidence-based drug oriented strategy decision support system. It is an integrated software application that provides decision support for strategic decisions such as guideline formulation, marketing authorization, and reimbursement. ADDIS stores aggregate clinical trial results with a unifying data model, and implements semi-automated evidence synthesis and benefit-risk modeling. These use cases were derived from direct discussion with experts from pharmaceutical industry, regulatory authorities, and academia, and from their feedback to early prototypes of the system. Before the models can be applied, trial results must be available in the system; for this, we present an assisted procedure for importing study designs from an existing database. The evidence synthesis and decision models of ADDIS allow decision makers to visualize and understand the available evidence and the trade-offs between different treatment options, thus addressing information overload and reducing the complexity of strategy decisions informed by clinical evidence. We stress that ADDIS does not aim at operational decision support, but aids in strategic decision making and provides a platform for computational methods in clinical trial informatics. In addition, the generation of the models cannot be completely automated: some steps require decisions from a domain expert, but can be supported by ADDIS as will be shown in this paper. To the best of our knowledge, ADDIS is the first system to allow on demand generation and use of the evidence synthesis and decision support models in a suitable way for strategic decision making.

We start by discussing existing systems and standards for clinical trial design and results in Section 9.2. The unifying data model is presented in Section 9.3. After that, in Section 9.4, we present ADDIS and the assisted procedures of study import and generation of evidence synthesis and benefit-risk models. In Section 9.5 we summarize our principal findings and propose directions for future research.

9.2 Background

Several systems and standards dealing with clinical trial information exist. We provide an overview of these systems and standards in Sections 9.2.1 and 9.2.2, respectively. Subsequently, in Section 9.2.3, we briefly describe the current state of methods for extraction of information from predominantly text-based sources of clinical trial designs and results. Finally, Sections 9.2.4 and 9.2.5 give an overview of the most relevant evidence synthesis and decision modeling approaches for strategic decision making.

9.2.1 Clinical trial information systems

In this section we briefly summarize the information systems that deal with clinical trials information, first those in operational management of trials and the regulatory environment, then the dissemination to the scientific community through publication in journals and registration, and finally how the results are summarized in systematic reviews.

Operational management and regulatory submission

Operational management refers to the administrative and data-gathering activities for a single trial. The operational management of clinical trials can be automated by using a Clinical Trial Management System (CTMS). Until circa 2000, the management and data collection of the vast majority of clinical trials were paper-based activities [CDISC, 2005], but the use of a CTMS has quickly become the norm [El Emam et al., 2009, Tufts, CSDD and CDISC, 2007]. The automation of operational management is now a mature field, and increasingly standardized (see also Section 9.2.2). However, CTMSs are data-centric single study systems that are focused on enabling the efficient operation of the trial and, often, submission of data to the US Food and Drug Administration (FDA). As of yet these systems do not enable cross-study analyses, data integration and data sharing.

After drug development, the pharmaceutical company compiles the evidence collected from clinical trials (and other research) into an electronic dossier that is submitted to the regulators who decide upon its market authorization. The dossier, especially the clinical trial results, forms the basis on which regulators assess the benefit-risk profile of a new drug. Submissions to the European Medicines Agency (EMA) and most other regulatory agencies worldwide are mainly text-based, containing aggregate-level results of clinical trials based on the applicant's statistical analyses. The FDA, on the other hand, requires an electronic submission of individual patient data to be able to perform independent analyses [FDA, 2009], and is currently building JANUS, a standards-based clinical data repository specifically designed for the integration of data [CDISC and FDA, 2005].

Results dissemination

Pharmaceutical companies and clinical research organizations may choose to publish the results of clinical trials in peer-reviewed scientific articles that do not include the underlying data set. Abstracts of publications are indexed in databases such as PubMed (`http://pubmed.com/`), which includes over 20 million citations from over 5,000 journals, of which more than 600,000 were published in 2009 [PubMed, 2011-05-02]. Although large in size, PubMed contains only a selected subset of the biomedical literature [Mulrow, 1994]. Abstract databases include meta data that might be incomplete due to being provided by external parties; for example, to achieve high sensitivity in searching for clinical trials in PubMed, restricting the search to the 'clinical trial' publication type is too restrictive [Haynes et al., 2005], and a broader query is recommended [Higgins and Green, 2009]. The Cochrane CENTRAL database of clinical trials is dedicated to indexing reports of clinical trials only, and contains references to 645,086 publications of clinical trials, of which 286,418 have been published since 2000 [Cochrane Library, 2011-05-02].

Until recently journal publications were the only non-confidential source of trial designs and results. This led to insufficient or inaccurate trial reporting and publication bias [Dickersin and Rennie, 2003] as e.g. over half of the clinical trials supporting successful new drug submissions made to the FDA had still not been published 5 years after the medicines' market approval [Lee et al., 2008]. Publication bias is a serious problem that can lead to incorrect conclusions in a systematic review. As early as in 1986 the registration of trials in advance was proposed as a solution to publication bias [Simes, 1986]. In 1997 the US became the first country to make trial registration a legal requirement, leading to the development of the ClinicalTrials.gov registry [McCray and Ide, 2000]. In 2004, both the World Health Organization (WHO) and the International Committee of Medical Journal Editors (ICMJE) released statements in support of the prospective registration of clinical trials. This policy has been widely adopted [ICTRP, 2010] and now assures that the existence of in any case most (recent) trials is known [Zarin et al., 2007]. Registries primarily focus on providing a record of trials for enabling patient recruitment and investigator accountability. Various organizations, including the WHO, have called for a full disclosure of the trial protocol (including amendments) and results [Krleza-Jeric et al., 2005, Sim et al., 2006, Kaiser, 2008, Ghersi et al., 2008, Zarin and Tse, 2008, Chan, 2008, Sim et al., 2009], but only the US have adopted legislation that requires registering results in ClinicalTrials.gov [FDA, 2007, Wood, 2009]. Study protocols can be retrieved from ClinicalTrials.gov in a (semi-structured) XML format [ClinicalTrials.gov, 2009b], while the retrieval of results is only possible in a text-based format. Other registries provide protocol information as semi-structured text, and do not include results.

In order to unify trial registration worldwide the WHO Registry Network was established in 2007. Twelve national and international registries are now part of the network. The European Union clinical trials registry, EudraCT, was opened to the public only recently, on 22 March 2011 [European Medicines Agency, 2011b], and is not part of the WHOs Registry Network. Table 9.1 gives an overview of the WHO primary registries, ClinicalTrials.gov, and EudraCT. ClinicalTrials.gov is by far the largest registry, containing more than 8 times the number of trials recorded in the second largest registry (EudraCT).

Systematic review

Evidence-Based Medicine (EBM) tries to use the best available evidence in assessing the benefits and risks of a treatment [Evidence-Based Medicine Working Group, 1992]. The most frequently implemented methods to assess the available evidence are the systematic review and meta-analysis of published research results [Sutton et al., 2009]. Systematic reviews are usually presented in a textual format without the underlying dataset. Given the effort required to perform a systematic review, fragmented reports regarding an indication are common [Caldwell et al., 2010]. The rapidly growing number of systematic reviews published each year [Honig, 2010]

Register	**Studies**	**Results**
ClinicalTrials.gov (United States)	106,649	yes (3,441)
EudraCT, the *European Union* Clinical Trials Register	12,990	no
ISRCTN register (international)	9,645	no
Japan Primary Registries Network	6,193	no
Australian New Zealand Clinical Trials Registry	5,221	no
The Netherlands National Trial Register	2,728	no
Clinical Trials Registry - *India*	1,704	no
Chinese Clinical Trial Register	1,319	no
Iranian Registry of Clinical Trials	1,291	no
German Clinical Trials Register	482	no
South Korea Clinical Research Information Service	108	no
Cuban Public Registry of Clinical Trials	105	no
Sri Lanka Clinical Trials Registry	60	no
Pan African Clinical Trial Registry	48	no

Table 9.1: ClinicalTrials.gov, EudraCT, and the 12 WHO primary registries. The 'studies' column indicates the number of registered trials (per 2 May 2011) and the 'results' column whether the registry also enables results publication.

has led to the 'overview of reviews' or 'umbrella review' to summarize the results of the existing reviews for an indication [Ioannidis, 2009]. Umbrella reviews generally merely repeat the pooled summaries of treatment effects from the original reviews, but it has been argued that they may lead to misleading and inconsistent conclusions [Caldwell et al., 2010].

The flagship of systematic reviewing is the Cochrane Library, kept up-to-date by the non-commercial Cochrane Collaboration (`http://www.cochrane.org/`). It is composed of three main components: the CENTRAL literature database of trial publications [Dickersin et al., 2002], the Cochrane Database of Systematic Reviews [Starr and Chalmers, 2003], and software for conducting and reporting on meta-analyses (`http://ims.cochrane.org/revman`). The Cochrane Library provides reviews of effects of healthcare interventions generated and updated by medical researchers, which are on average regarded to be of better quality than the corresponding studies published in traditional journals [Jadad et al., 1998]. Compared with the traditional journal publications that usually provide data in tables or figures, the Cochrane Reviews incorporate descriptions and results of the original studies, while the software enables making odds-ratio diagrams that can also include the newest studies. However, the available datasets are not complete and they are structured according to the reviews rather than the included studies. Moreover, due to the inaccessibility of clinical trials information, systematic reviews are static entities that only reflect the state of knowledge at the time of the literature search.

9.2.2 Standards and data models

The information systems discussed above, especially those in operational management, are enabled by standards and data models that have been developed over the last two decades. Two main standardization bodies in the field are the Clinical Data Interchange Standards Consortium (CDISC) and Health Level 7 (HL7). The CDISC develops vendor-neutral and freely available standards that enable information system interoperability in the operational management and regulatory submission of clinical trials. HL7 develops standards that apply broadly to clinical and administrative data in health care, and thus do not focus on any specific clinical domain. The foundation of HL7 standards development work is the Reference Information Model (RIM), a high level object model of the health care domain. Several standards are derived from the RIM, such as V3 Messages for the meaningful interchange of data between health care systems, GELLO for rule-based decision support, and the Clinical Document Architecture for semantically structured documents. HL7 also maintains the Arden Syntax that enables rule-based expert systems that support operational decision making in health care.

The Biomedical Research Integrated Domain Group (BRIDG) project aims at bringing together the common elements of their various standards to a shared view of semantics of the domain of protocol-driven research and its associated regulatory artifacts [Biomedical Research Integrated Domain Group (BRIDG), 2010]. The model is intended to be implementation independent in the sense that it models the problem domain, and not any specific solution. For example, unlike some other CDISC standards it does not specify the format in which to submit data to the FDA. The BRIDG model is subdivided into several sub-domain views: the protocol representation, study conduct, adverse event and regulatory perspectives. While the operational aspects of clinical trials are well covered by these perspectives, a data analysis perspective is currently missing as there is no adequate standard for statistical analysis.

The ClinicalTrials.gov registry has developed their own model, the Data Element Definitions (DED) [ClinicalTrials.gov, 2009a,c]. They allow the reporting of aggregated outcome data and statistical analyses to some extent, but the semantic depth of the information is limited as most fields are free text. For example, since eligibility criteria are free text fields, searching for a trial relevant to a specific patient condition is inaccurate [Tu et al., 2011].

The Human Studies Database (HSDB) project aims to share fully machine understandable representations of study design information between institutions [Sim et al., 2010]. To enable this, they develop the Ontology of Clinical Research (OCRe), which defines the concepts that should be queried across the individual institutions' databases. The creators of OCRe have argued that while the BRIDG model accurately captures the operational semantics of clinical trials, its modeling of many aspects relevant to cross-study analyses is weak [Sim et al., 2010]. The main contributions of OCRe at this time are a study design topology [Carini et al., 2009], the ERGO formal machine readable representation of eligibility criteria [Tu et al., 2011], and a model of study outcomes that separates the phenomena of interest from the variables that code them [Sim et al., 2010]. It also contains a study design representation derived from BRIDG [Sim et al., 2010]. While OCRe is a promising effort, it is still far from

Model	Study design	Aggregate results	Semantic depth	Completeness
BRIDG	++	-	+/-	+
DED	+/-	+/-	-	+
OCRe	+	-	+	-
OBX	+	+/-	+/-	+

Table 9.2: Approximate scoring of data models on several dimensions relevant to automated processing of aggregate clinical trials results. 'Study design' is the extent to which complex study designs can be accurately represented, 'aggregate results' refers to the inclusion of aggregate results and description of the means by which they were derived, 'semantic depth' refers to the level of semantic structure achieved (e.g. contrast the text-based eligibility criteria in the DED to the ERGO model used in OCRe), while 'completeness' refers to the extent to which the model in its current state achieves its stated goals. Note that these dimensions are difficult to assess and the assigned ratings are subjective.

comprehensively representing study design and lacks results completely.

The Ontology Based Extensible Conceptual Model (OBX) is another ontology for representing clinical trials [Kong et al., 2011, Scheuermann, 2010]. It is specifically aimed at making available the results of immunology studies for data re-use and re-analysis. The OBX also incorporates study design representation ideas from BRIDG and the ClinicalTrials.gov DED [Kong et al., 2011]. While it appears successful in developing a broadly applicable data model for biomedical studies, and also includes results, it would appear that the objections raised by HSDB researchers about the depth of modeling in BRIDG also apply to OBX, and the results are represented in a way similar to the ClinicalTrials.gov DED.

We rate the four major models in Table 9.2 on how well they represent study design and aggregate results, as well as their semantic depth and completeness. Table 9.3 gives a summary of the main goal as well as the strengths and weaknesses of each model. One common property of all models is that they rely on an external terminology for their clinical content. Controlled terminologies (synonymously: controlled vocabularies, coding systems) of clinical terms are an important first step in the application of information technology to medicine [Cimino, 1996]. Controlled terminologies predate information technology, e.g. the International Classification of Diseases (ICD) was already introduced in 1893. The ICD formally codes diseases and enables, for example, the assessment of disease incidence from medical records. Other terminologies fill other niches, for example the Medical Subject Headings (MeSH) [Nelson et al., 2004] is used to index the medical literature (e.g. PubMed meta data is coded in MeSH), and the Medical Dictionary for Regulatory Activities (MedDRA) is used for coding safety data (e.g. adverse events). Many of these specialized terminologies are organized into a strict hierarchy, which means that some specific terms may fit in multiple places [Cimino, 1996]. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) terminology is an important attempt to create a clinical terminology with comprehensive coverage [Schulz et al., 2006]. It currently contains around 311,000 concepts and 800,000 terms [The In-

Model	Purpose	Strengths	Weaknesses
BRIDG	Operational management, regulatory submission	Standardization process, practical applications	No aggregate data, limited depth of modeling (some aspects)
DED	Limit publication bias, enable disclosure of results	Completeness, working system, US trials required to register (data available)	Limited semantic structure
OCRe	'Computable' representation of human studies	Broad scope, semantic depth	Not finished, not implemented, results not represented
OBX	Make available data from immunology studies	Balance of the depth of modeling and the practical feasibility, working system	Limited depth of modeling (some aspects)

Table 9.3: The purpose of each of the data models as well as their strengths and weaknesses from the perspective of enabling automated evidence synthesis and decision support.

ternational Health Terminology Standards Development Organisation, 2011]. It also goes beyond a simple hierarchical structure and provides the logical relationships that hold between terms; over 1.3 million such relationships are currently modeled [Schulz et al., 2006, The International Health Terminology Standards Development Organisation, 2011]. The Unified Medical Language System (UMLS) [Lindberg et al., 1993, Bodenreider, 2004] 'Metathesaurus' brings together over 60 biomedical terminologies and their relationships. The ICD, SNOMED CT and MeSH are among the integrated terminologies.

9.2.3 Data extraction

The free-text nature of clinical trial publications is an important obstacle to the application of data mining and other automated knowledge discovery and decision aid uses [Sim et al., 2000]. There are many existing approaches to extract some of this data from abstracts or full texts of journal articles or health records, as reviewed in [Cohen and Hersh, 2005, Meystre et al., 2008]. However, the potential benefits are currently not fully realized due to lack of directly applicable tools [Cohen and Hersh, 2005] and text mining approaches for supporting research [Meystre et al., 2008].

Text mining of articles describing clinical trials could support researchers in performing a systematic review. Information extraction, on the one hand, attempts to create structured datasets from unstructured text by identifying entities and relationships between entities in the text. Most current approaches focus on the abstract rather than the full text as it provides a more controlled environment, and they tend to focus on only a few information elements [Kiritchenko et al., 2010]. The ExaCT system [Kiritchenko et al., 2010] assists systematic reviewers in extracting 21 key trial characteristics from full text articles. The system is accurate enough to save a considerable amount of time in extracting these elements, but systematic reviewers do have to verify the extracted information manually. Text analytics, on the other hand, identifies patterns in large collections of texts in order to classify documents and unlock relationships between documents. Text analytics can help systematic reviewers in structuring large sets of search results from abstract databases (e.g. PubMed) and increase the efficiency of finding the relevant clinical trials. However, to be able to reliably perform evidence synthesis and decision modeling based on the extracted clinical trials data, a higher level of accuracy and generality is needed than is currently offered by text mining methods. Thus, although automated methods can lower the workload, manual data extraction remains necessary.

9.2.4 Evidence synthesis

The most commonly applied evidence synthesis method is pair-wise meta-analysis, in which a number of studies comparing the same pair of treatments A and B are synthesized to assess their relative performance δ^{AB} on a specific outcome [Normand, 1999]. For example, do more depressed patients respond to treatment with paroxetine than with fluoxetine (both antidepressants)? Or, do more patients treated with paroxetine experience nausea during the studies than those treated with fluoxetine? The observed treatment differences $\hat{\delta}_i^{AB}$ in the individual studies i are used to estimate the overall difference δ^{AB}. Network meta-analysis, a recent extension of pair-wise meta-analysis, synthesizes evidence on the relative effects of a whole network of treatments simultaneously [Lumley, 2002, Lu and Ades, 2004, Salanti et al., 2008a]. It incorporates both direct and indirect evidence on the relative effects, and allows a statistical analysis of evidence consistency [Lu and Ades, 2006, Salanti et al., 2008a, Dias et al., 2010]. Except for the possible inconsistency between direct and indirect evidence, the assumptions underlying network meta-analysis are the same as those underlying pair-wise meta-analysis [Caldwell et al., 2005]. The method has gained acceptance, and applications are being published in top medical journals [e.g. Cipriani et al., 2009, Stettler et al., 2008]. However, application of the method has so far remained the work of a select few experts, as model specification is difficult and no automated tools are available. Many other evidence synthesis methods exist [Sutton and Higgins, 2008], but pair-wise and network meta-analysis are by far the most important ones for decision support.

9.2.5 Decision models

Although evidence synthesis is an important tool for evidence-based medicine as it helps to summarize the available evidence, it does not help the decision maker to take into account the trade-offs of the risks of a treatment and its related benefits. There is an increasing interest in evidence-based multi-criteria decision models [Haynes et al., 2002, Cooper et al., 2005] taking into account efficacy and safety of alternative treatments. The target domains of model-based decisions include marketing authorization for new drugs, development of guidelines concerning recommended treatments, and prescription decisions such as which anti-depressant to subscribe, for example, in a setting where besides efficacy specific safety issues are also of interest, e.g. dizziness could be life-threatening given the specific patient's occupation. Many such decisions have to take into account trade-offs between different decision criteria (e.g. efficacy and safety), and can be aided through multi-criteria decision models [Guo et al., 2010] or application-specific ways of mapping benefits and risks to a single scale [Ouellet, 2010]. Multi-criteria decision models can structure the decision problem and make trade-offs between the alternative medical treatments explicit. In general, Multiple Criteria Decision Analysis (MCDA) methods compare m alternatives on n criteria. The performance of each of the alternatives is measured in terms of the criteria, and explicit trade-offs (preferences) between the criteria may be specified by the decision maker. The decision is aided by finding the optimal alternative (choice problem), by ranking the alternatives from best to worst, or by classifying the alternatives into discrete classes, such as good, acceptable and bad alternatives [Roy, 1996]. An inverse approach, in which typical preferences that favor each of the alternatives are derived using the decision model, is also possible [Lahdelma et al., 1998].

There exists benefit-risk models based on point estimates of the criteria measurements [Mussen et al., 2007, Felli et al., 2009]. However, taking into account decision uncertainty is necessary in the medical context as the data might not distinguish the alternatives with sufficient certainty to make an informed decision. In that case, a decision has to be postponed until more or higher quality information becomes available. Therefore, we focus on stochastic methods, where the performances are measured using probability distributions rather than with point estimates. Stochastic methods based on single studies [Tervonen et al., 2011, Lynd and O'Brien, 2004] model the 'absolute' treatment effects and use those as performance measures (e.g. the binomial success probability of a treatment response can be modeled using a Beta distribution for each of the treatments). Using absolute measures has the advantage that the observed differences in performance have an immediate clinical implication, and thus eliciting preferences from the decision maker is relatively easy. For example, one could ask 'Would you consider improving the probability of treatment response from 0.73 to 0.80 to be more important than reducing the probability of the side effect dizziness from 0.12 to 0.09?'. However, the generalizability of a model using absolute measures is questionable, as the absolute treatment effect and the incidence of side effects depends heavily on the design and specific population of the study. Models based on evidence synthesis [van Valkenhoef et al., 2012e] are preferable from this perspective, as measurements would be based on *relative* effects estimated using all available studies. Thus, such a method is more robust and generalizable, but the relative scales make the interpretation of the clinical implications more difficult [Egger et al., 1997]. A hybrid approach, in which the (relative) measurements are derived using evidence synthesis, but framed in clinically meaningful (absolute) terms using (assumed or estimated) baseline risk for the population of interest may be the best one [van Valkenhoef et al., 2012e].

So far, only benefit-risk models based on Stochastic Multi-criteria Acceptability

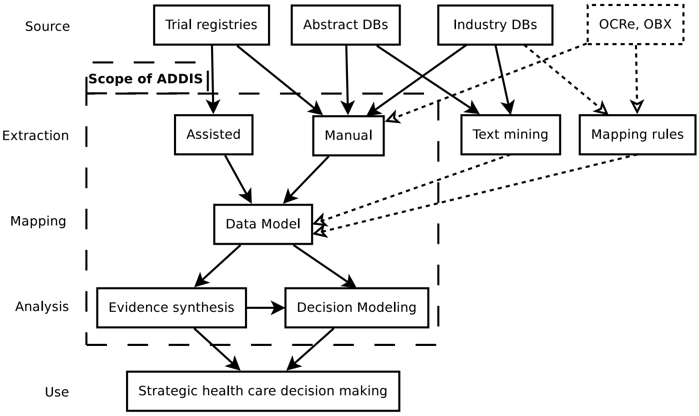

Source
Trial registries
Abstract DBs
Industry DBs
OCRe, OBX
Scope of ADDIS
Extraction
Assisted
Manual
Text mining
Mapping rules
Mapping
Data Model
Analysis
Evidence synthesis
Decision Modeling
Use
Strategic health care decision making

Figure 9.1: The current and future flow of information into the ADDIS system, and the role of the unifying data model in supporting evidence synthesis and decision support. The dashed rectangle indicates the scope of the functionality currently implemented by ADDIS. The solid arrows show the current situation, while dotted arrows indicate how future developments will benefit ADDIS.

Analysis (SMAA) [Lahdelma et al., 1998, Lahdelma and Salminen, 2001, Tervonen and Figueira, 2008] allow taking into account the full uncertainty surrounding the measurements from clinical trials as well as imprecise preferences, while enabling the comparison of $m \geq 2$ treatments on $n \geq 2$ outcomes through Monte Carlo simulation. A two-dimensional visual approach (also based on Monte Carlo simulation) may be preferable if $m = 2$ and $n = 2$ [Lynd and O'Brien, 2004]. This model is based on standard cost-effectiveness analysis techniques, and we shall refer to it as the "Lynd & O'Brien" model. Both methods enable the inverse approach, where the preferences supporting specific decisions are derived using the model.

9.3 The unifying data model

We developed a unifying data model to enable evidence-based decision support methods based on either individual studies or evidence synthesis. As discussed before, the most important methods are pair-wise meta-analysis, network meta-analysis, and stochastic multi-criteria benefit-risk assessment. The data model is aimed at supporting these use cases. As was shown in Section 9.2.2, several worthwhile data modeling efforts are underway. Unfortunately none of them have the needed level of modeling to be directly applicable to our use cases. It is clear that while very precise representations (such as are being created for OCRe) will be needed for the reuse of clinical trial data for general purposes, they will not translate directly to the application of evidence synthesis.

In addition, the data model is intended to be a well defined data extraction target to enable health strategy decision support. Data extraction is currently based

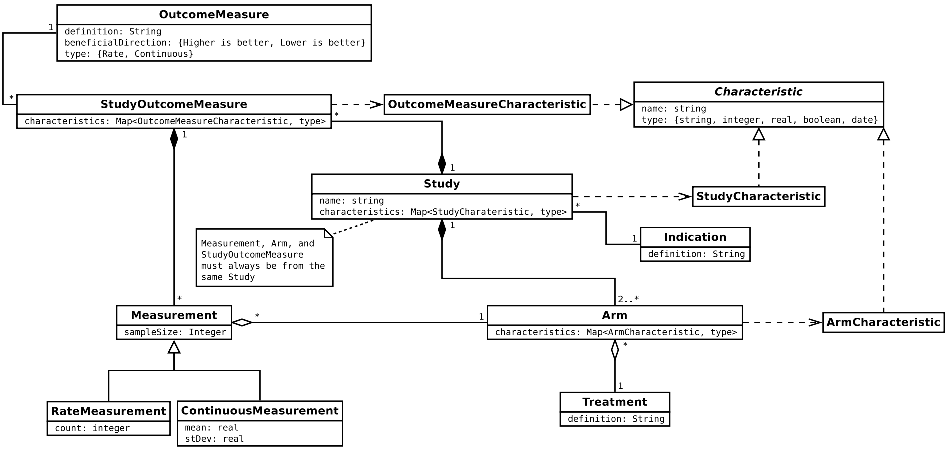

OutcomeMeasure
definition: String
beneficialDirection: {Higher is better, Lower is better}
type: {Rate, Continuous}
1
*
StudyOutcomeMeasure
characteristics: Map<OutcomeMeasureCharacteristic, type>
OutcomeMeasureCharacteristic
Characteristic
name: string
type: {string, integer, real, boolean, date}
*
1
1
Study
name: string
characteristics: Map<StudyCharateristic, type>
StudyCharacteristic
*
1
1
Indication
definition: String
Measurement, Arm, and StudyOutcomeMeasure must always be from the same Study
*
2..*
Measurement
sampleSize: Integer
*
1
Arm
characteristics: Map<ArmCharacteristic, type>
ArmCharacteristic
*
1
Treatment
definition: String
RateMeasurement
count: integer
ContinuousMeasurement
mean: real
stDev: real

Figure 9.2: The unifying data model for common types of aggregate analysis of clinical studies in UML2 class notation.

on manual and assisted import from clinical trial registries and journal publications, and in the future on automated rule-based import from structured databases or semi-automated data extraction from the medical literature. This vision is shown in Figure 9.1. The data model is not intended to fully model clinical trials, as we believe BRIDG and OCRe are better positioned to eventually fill this gap. Rather, there is a need for a unifying data model that captures the invariants of the domain from the perspective of evidence synthesis. Such a data model provides clear requirements for more fine-grained models, a target for text mining and (sub-)domain-specific rules for data conversion, and a basis on which to build decision support systems. Thus, our data model represents the structure of trials only to a limited extent and appropriate (domain-specific) mapping is required to enable its use. Mapping rules from more fine-grained data models such as OCRe can be developed once these models have matured. The unifying data model is described below and illustrated in Figure 9.2. In the text, we will refer to entities in the domain model using capitalized words (e.g., Study and OutcomeMeasure).

Clinical trials are represented by the class Study. The data model may also apply to other studies with human populations, such as observational studies, but it was primarily designed to represent clinical trials. Each Study is identified by a name (e.g., “Coleman et al. 2001” or “NCT00296517”). A Study considers a single (therapeutic) Indication. Each Indication is identified by a definition (e.g., “Depression” or “Type 2 Diabetes”). A Study consists of (two or more) Arms. An Arm within the context of a clinical trial can be seen as a group of patients within a Study who all receive the same medical Treatment. Within a Study there can exist different Arms for the same medical Treatment (e.g., receiving different dosages). Each Treatment is identified by a definition (e.g., “Placebo”, a simulated medical intervention, “Fluoxetine”, an anti-depressant, or “Rosiglitazone”, an anti-diabetic).

An OutcomeMeasure is identified by a definition, referring to an endpoint (e.g., "Responders on the HAM-D rating scale" or "Change from baseline triglycerid levels (mg/dL)") to be measured in studies, or an adverse event (e.g., "Headache", "Nausea" or "Chest pain") that can occur in studies. An OutcomeMeasure has a beneficial direction (higher is better or lower is better). There are two Types of OutcomeMeasures in terms of how they are measured: rate or continuous (see below). A Study can have (zero or more) OutcomeMeasures and an OutcomeMeasure can apply to (zero or more) Studies. Such a combination is called a StudyOutcomeMeasure (identified by the OutcomeMeasure and Study). Note that the BRIDG model discussed before also uses the term StudyOutcomeMeasure in this context.

A Measurement refers to a combination of a StudyOutcomeMeasure and an Arm within the same Study. Each such combination can have at most one Measurement. A Measurement has a sample size (e.g. 98 patients). The sample size is associated with the Measurement and not with the Arm, as the relevant sample size depends on the way the outcome measure is analyzed, and may change over time due to patients dropping out of the study. Each Measurement is either a RateMeasurement or a ContinuousMeasurement, depending on the type of OutcomeMeasure the measurement refers to. A RateMeasurement describes the number of individuals in the Arm for whom the OutcomeMeasure occurred. A ContinuousMeasurement describes the result by a mean and a standard deviation (two real numbers).

The entities described above form the core of the unifying data model. Generation of evidence synthesis and decision models is based on Studies, Arms, Treatments, OutcomeMeasures, and Measurements. In addition, the data model includes Characteristics for more descriptive information. The Characteristics are identified by a name (e.g. for StudyCharacteristics "Study size", "Group allocation", "Treatment blinding" or "Patient eligibility criteria", for ArmCharacteristics "Arm size", "Dosing" or "Gender distribution", for OutcomeCharacteristics "Is primary outcome" or "Assessment time") and include the type of the characteristic. The type is used for validating the values input for the actual characteristic values, and is useful in generating Graphical User Interface (GUI) components for input of characteristic values. The object diagram in Figure 9.3 includes an example instantiation of the data model.

If the characteristics are left out, the data model contains the minimal information for generation of the evidence synthesis and decision models described in the previous section. The minimal representation makes it easier to import data to a system implementing it, and increases applicability of the model from being specific to a certain subfield (e.g. cancer treatments) to being general for all. However, the minimality causes the data model to be specific for the chosen types of analysis models. If, for example, meta-regression techniques should be applied, the data model would need to be extended accordingly. We allow descriptive extensions by including the characteristics. The characteristics also serve for storing the information that is unnecessary for analysis model generation, but necessary for expert judgment on which studies should be included in the analysis (e.g. based on the type of dosing). They also serve for specialization of the data model in that, if the need arises, new ones can be added without breaking the functionality of analysis model generation.

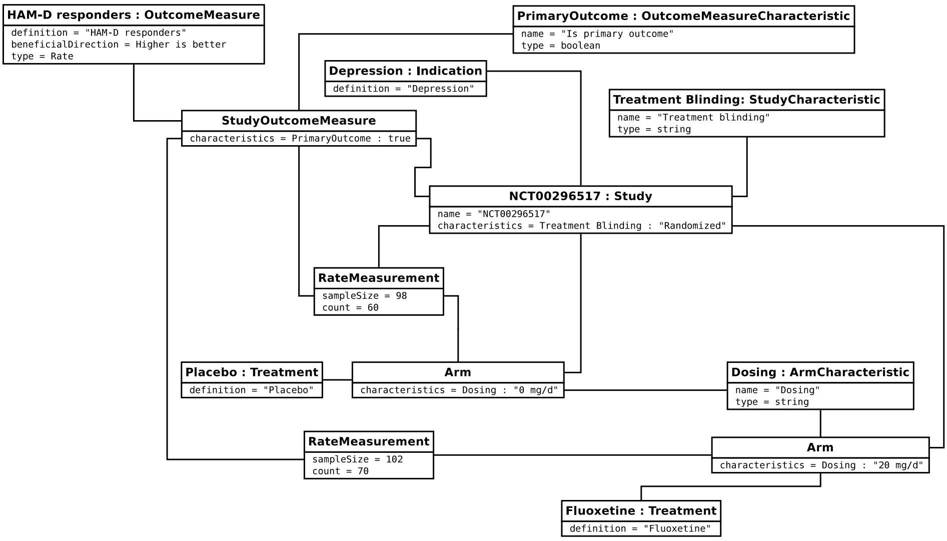

HAM-D responders : OutcomeMeasure
definition = "HAM-D responders"
beneficialDirection = Higher is better
type = Rate
PrimaryOutcome : OutcomeMeasureCharacteristic
name = "Is primary outcome"
type = boolean
Depression : Indication
definition = "Depression"
Treatment Blinding: StudyCharacteristic
name = "Treatment blinding"
type = string
StudyOutcomeMeasure
characteristics = PrimaryOutcome : true
NCT00296517 : Study
name = "NCT00296517"
characteristics = Treatment Blinding : "Randomized"
RateMeasurement
sampleSize = 98
count = 60
Placebo : Treatment
definition = "Placebo"
Arm
characteristics = Dosing : "0 mg/d"
Dosing : ArmCharacteristic
name = "Dosing"
type = string
RateMeasurement
sampleSize = 102
count = 70
Arm
characteristics = Dosing : "20 mg/d"
Fluoxetine : Treatment
definition = "Fluoxetine"

Figure 9.3: An example instantiation of the unifying data model as UML2 object diagram. The example instantiation depicts one Study including two Arms with two different Treatments. For both of the arms, a measurement on one OutcomeMeasure is shown. For the Arm, StudyOutcomeMeasure, and Study, there is each one Characteristic presented together with the associated value.

9.4 ADDIS decision support system

The unifying data model together with a semi-automated analysis generation system are implemented in the open source decision support software ADDIS[1]. It provides an easy interface to enter, import and manage study design and outcome information from clinical trials, and is specifically aimed at supporting the user in creating (network) meta-analyses and (multi-criteria) benefit-risk models. The main components of the software are:

- Implementation of the unifying clinical trial data model,
- GUI for managing trials and analyses,
- Semi-automated import of studies from ClinicalTrials.gov,
- GUI 'wizards' for semi-automated generation of analyses,
- External packages for computing the analyses,
- GUI components for results visualization, and
- Links to external databases (PubMed, ATC database, drug compendium).

---

[1] `http://drugis.org/addis`. Screen casts are based on version 1.6, `http://drugis.org/addis1.6`.

ADDIS integrates an external network meta-analysis library[2] [van Valkenhoef et al., 2012d] and JSMAA [Tervonen, 2010] for computation of SMAA benefit-risk models. The ADDIS data format is represented by an XML schema[3] that instantiates the unifying data model. Evolution of the format is supported by versioned XML schemas that are forward-compatible through XSL transformations (XSLT). ADDIS supports the coding of drugs with their Anatomical Therapeutic Chemical Classification System (ATC) code and uses them to link to drug compendia. The ATC codes can be filled in automatically by ADDIS through integration with an online database, when given the compound name. A coding system for outcome measures and adverse events will be integrated in the future. A number of study characteristics are supported by default in ADDIS to enhance the user experience. For studies, these include the study title, randomization, treatment blinding, the study objective, the in- and exclusion criteria, the start and end date of the study, PubMed IDs of relevant publications and several others.

9.4.1 Study import from ClinicalTrials.gov

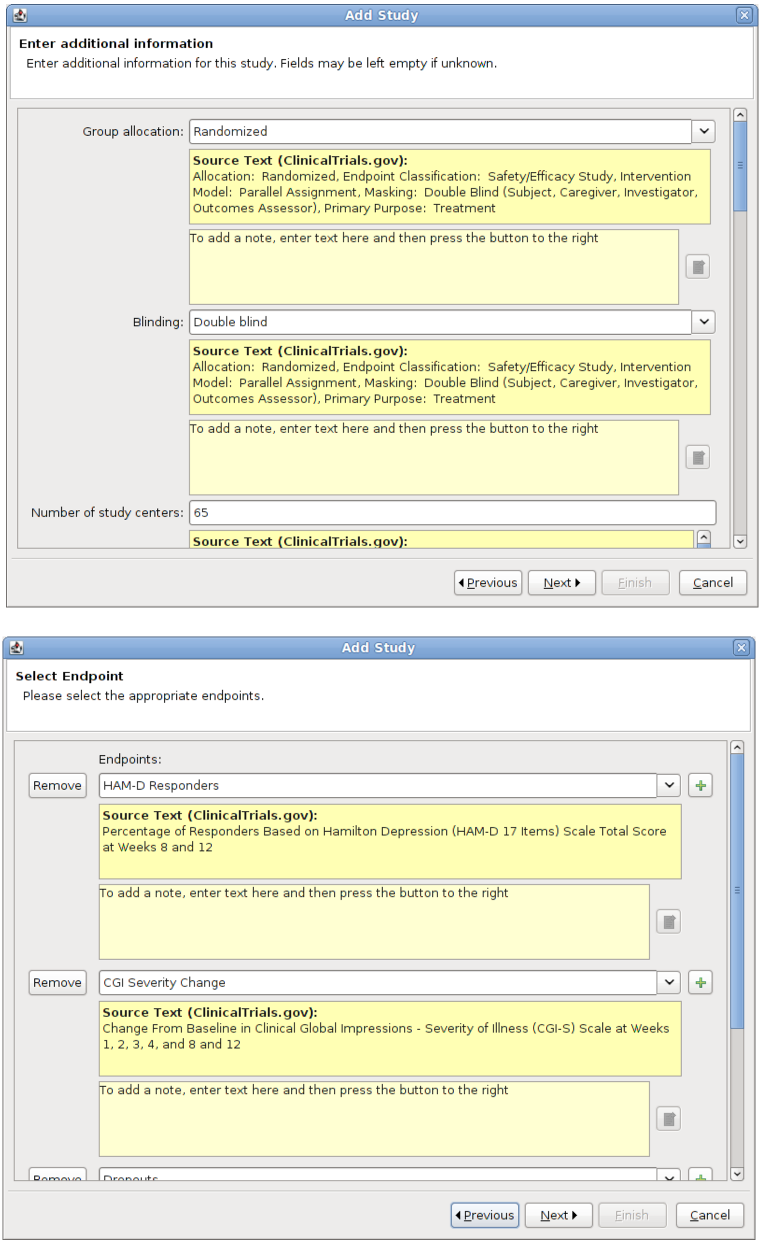
The ClinicalTrials.gov registry is by far the most comprehensive clinical trials registry in the world, currently containing information on over 100,000 trials (see Table 9.1). ClinicalTrials.gov has a simple and easy to use interface to programmatically search for trials and retrieve their protocols in XML format (according to their own DED) [ClinicalTrials.gov, 2009b]. Unfortunately, the results are currently not available as XML and it is unclear when this will be remedied.

In ADDIS, we use this XML interface to semi-automatically import studies from ClinicalTrials.gov. The user inputs the NCT-ID of the trial that should be imported, and the software will retrieve the XML, from which it automatically fills in fields. For example, many of the study characteristics, such as randomization and treatment blinding are matched from DED fields using simple rules. However, those fields that form the core of our data model, such as the indication, treatments and outcome measures, have to be manually mapped to entities in the database. This is so because accurately mapping the free-text descriptions given in the ClinicalTrials.gov records would require (1) deep semantic modelling of the entities in our database, and (2) natural language processing of incredibly high accuracy. While both fields are rapidly evolving, neither of these problems have a fully satisfactory solution at the moment. This mapping step is critically important to the correctness of subsequent analyses and thus inaccurate automatic mapping could degrade the decision makers' trust in the system. Hence, for the time being, the mapping is deliberately left to the user. Figure 9.4 shows examples of the user interface for study import. The original source text is preserved as a note that is kept with the relevant field, and the user can also enter additional notes. Due to the lack of an XML interface for study results, those have to be entered manually, and can not be linked to the source text.

---

[2] `http://drugis.org/mtc`

[3] `http://drugis.org/files/addis-1.xsd`



Add Study

Enter additional information

Enter additional information for this study. Fields may be left empty if unknown.

Group allocation: Randomized

Source Text (ClinicalTrials.gov):
Allocation: Randomized, Endpoint Classification: Safety/Efficacy Study, Intervention Model: Parallel Assignment, Masking: Double Blind (Subject, Caregiver, Investigator, Outcomes Assessor), Primary Purpose: Treatment

To add a note, enter text here and then press the button to the right

Blinding: Double blind

Source Text (ClinicalTrials.gov):
Allocation: Randomized, Endpoint Classification: Safety/Efficacy Study, Intervention Model: Parallel Assignment, Masking: Double Blind (Subject, Caregiver, Investigator, Outcomes Assessor), Primary Purpose: Treatment

To add a note, enter text here and then press the button to the right

Number of study centers: 65

Source Text (ClinicalTrials.gov):

Add Study

Select Endpoint

Please select the appropriate endpoints.

Endpoints:

Remove HAM-D Responders

Source Text (ClinicalTrials.gov):
Percentage of Responders Based on Hamilton Depression (HAM-D 17 Items) Scale Total Score at Weeks 8 and 12

To add a note, enter text here and then press the button to the right

Remove CGI Severity Change

Source Text (ClinicalTrials.gov):
Change From Baseline in Clinical Global Impressions - Severity of Illness (CGI-S) Scale at Weeks 1, 2, 3, 4, and 8 and 12

To add a note, enter text here and then press the button to the right

Remove Dropouts

Figure 9.4: Example screens from the study input/import wizard. The top screen shows how study characteristics are input. Most of these are matched automatically from the source text. The bottom screen shows that endpoints must be mapped to entities in the database by the user.

9.4.2 Evidence synthesis

ADDIS assists generation of pair-wise and network meta-analyses in a step-wise fashion; the process is presented in Figure 9.5. To start, the user needs to select an indication. Based on the selected indication, the system selects and presents all outcome measures included in the different available studies in the system considering the indication. After the user selects the desired outcome measure for analysis, the system selects the studies and their included treatments based on the selected (indication, outcome measure) tuple, and constructs the evidence graph. The graph is presented visually and has the vertices labelled with treatment definitions and the edges labelled with the number of studies including that comparison (see Figure 9.6). The user can pick the treatments to be compared. For a pair-wise analysis, exactly two treatments have to be selected, and for network meta-analysis two or more treatments can be selected. The software will not allow the user to continue unless the selected treatments form a connected graph. Following the treatment selection, the system presents the set of studies together with their characteristics, and non-desired studies can be easily removed by the user on a case by case basis. The chosen treatments and studies must form a connected evidence graph. Finally, if studies include a specific treatment in more than one arm (e.g. in various doses), the user must choose which arm to use in the analysis (see Figure 9.7).

Visualization of results is of crucial importance for applicability of methods used in evidence-based medicine. ADDIS provides visualization of the odds ratios, mean differences, risk ratios, and risk differences of standard meta-analyses in terms of forest plots (Figure 9.8). The network meta-analysis rank probabilities are presented as bar charts as shown in Figure 9.9.

9.4.3 Benefit-Risk models

The creation of benefit-risk models in ADDIS can be based on either an individual study or (previously created) meta-analyses. The user first selects an indication, and chooses whether to base the analysis on a single study or evidence synthesis. If the analysis is based on a single study, the system selects studies belonging to the selected indication and allows the user to choose one. Then, the user is presented with the available criteria (outcome measures in the selected study) and alternatives (arms in the selected study), and may select two or more of each to include in the benefit-risk model. If the analysis is based on evidence synthesis, the available criteria are the outcome measures for which a (network) meta-analysis exists within the selected indication. If multiple analyses exist for an outcome measure, one must be chosen. The available alternatives are the intersection of the sets of treatments included in the selected analyses. Two or more criteria and alternatives can be selected to include in the benefit-risk model. The final step in the creation of a benefit-risk model based on evidence synthesis is shown in Figure 9.10.

Benefit-risk decision models were already broadly discussed in Section 9.2.5. ADDIS supports decision makers using several different methods (see Table 9.4). These methods are organized along three axes: the number of alternatives, the number of criteria and the number of clinical trials in the evidence base. For a single-criterion

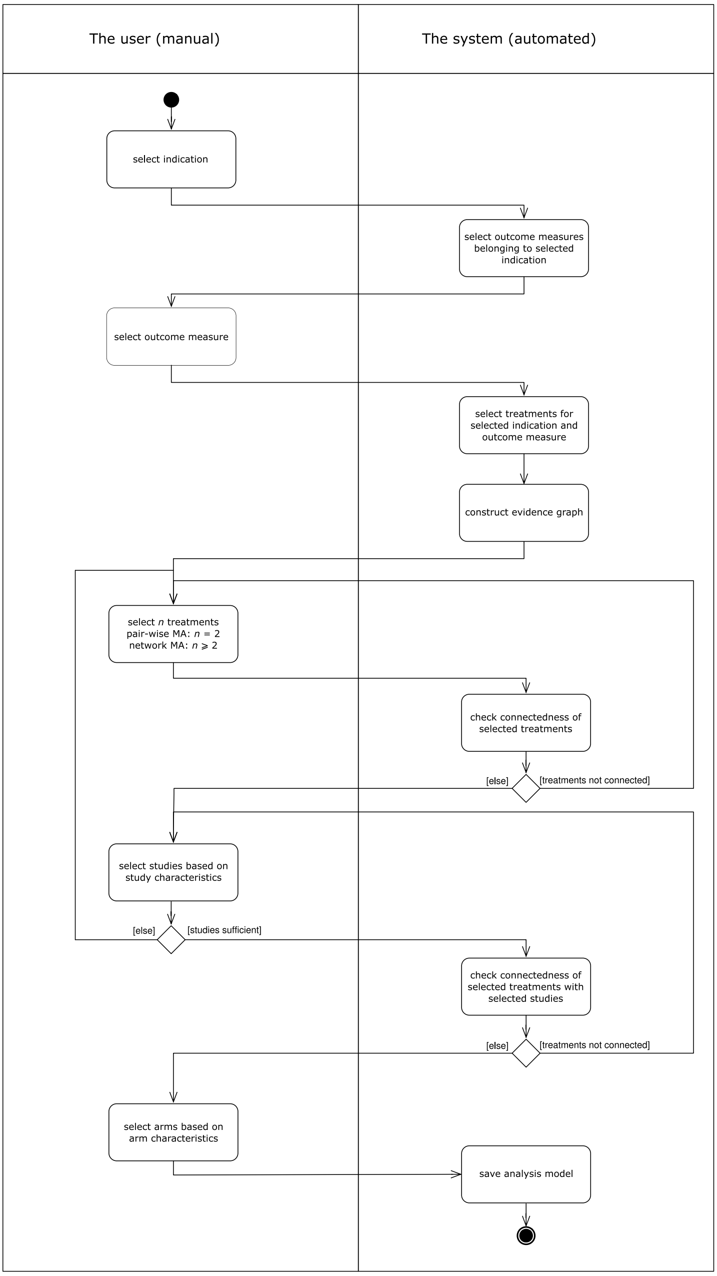
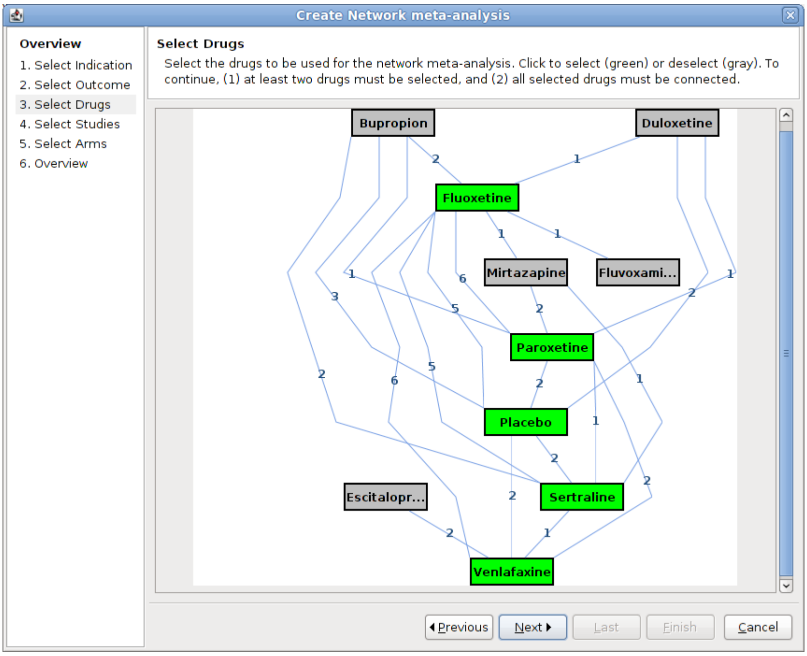

The user (manual)
The system (automated)
select indication
select outcome measures belonging to selected indication
select outcome measure
select treatments for selected indication and outcome measure
construct evidence graph
select n treatments pair-wise MA: $n = 2$ network MA: $n \geqslant 2$
check connectedness of selected treatments
[else]
[treatments not connected]
select studies based on study characteristics
[else]
[studies sufficient]
check connectedness of selected treatments with selected studies
[else]
[treatments not connected]
select arms based on arm characteristics
save analysis model

Figure 9.5: The process of meta-analysis creation as an activity diagram. The activities on the right-hand side are automated in the system, and the steps on the left require input and conscious decisions from the user. The process is identical for pair-wise and network meta-analysis, except that for pair-wise meta-analysis the number of treatments is restricted to exactly two.



Create Network meta-analysis

Overview

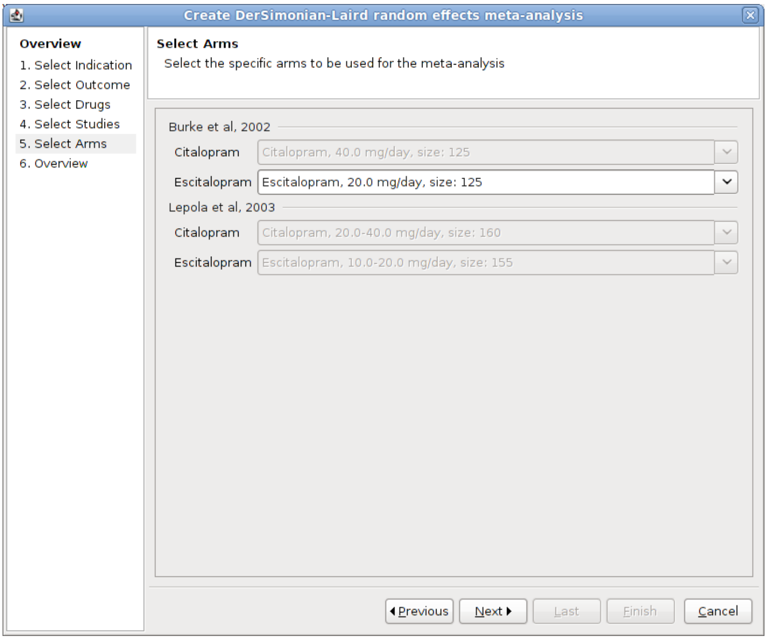
1. Select Indication
2. Select Outcome
3. Select Drugs
4. Select Studies
5. Select Arms
6. Overview

Select Drugs

Select the drugs to be used for the network meta-analysis. Click to select (green) or deselect (gray). To continue, (1) at least two drugs must be selected, and (2) all selected drugs must be connected.

Bupropion
Duloxetine
Fluoxetine
Mirtazapine
Fluvoxami...
Paroxetine
Placebo
Escitalopr...
Sertraline
Venlafaxine
Previous
Next
Last
Finish
Cancel

Figure 9.6: An evidence graph for a network meta-analysis. The treatments are the vertices, and the number of studies for each comparison label the edges (e.g., six studies compare fluoxetine and paroxetine). The green treatments are included in the analysis, the grey ones excluded.



Create DerSimonian-Laird random effects meta-analysis

Select Arms

Select the specific arms to be used for the meta-analysis

Burke et al, 2002

Citalopram: Citalopram, 40.0 mg/day, size: 125

Escitalopram: Escitalopram, 20.0 mg/day, size: 125

Lepola et al, 2003

Citalopram: Citalopram, 20.0-40.0 mg/day, size: 160

Escitalopram: Escitalopram, 10.0-20.0 mg/day, size: 155

Overview
1. Select Indication
2. Select Outcome
3. Select Drugs
4. Select Studies
5. Select Arms
6. Overview
Previous
Next
Last
Finish
Cancel

Figure 9.7: If several matching arms are available, the user must select an appropriate one based on the arm's characteristics. If the available arms are not appropriate, the user can go back to exclude the study.

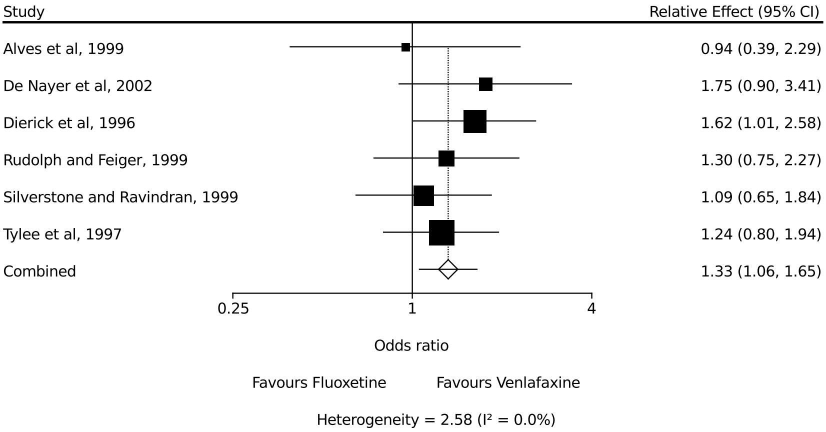
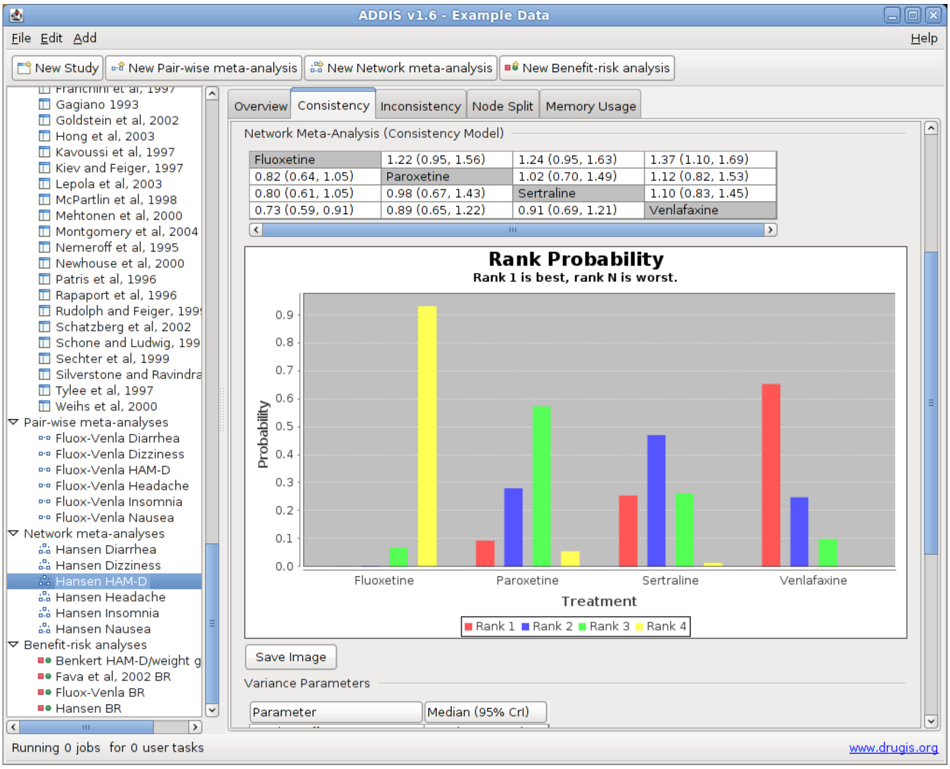
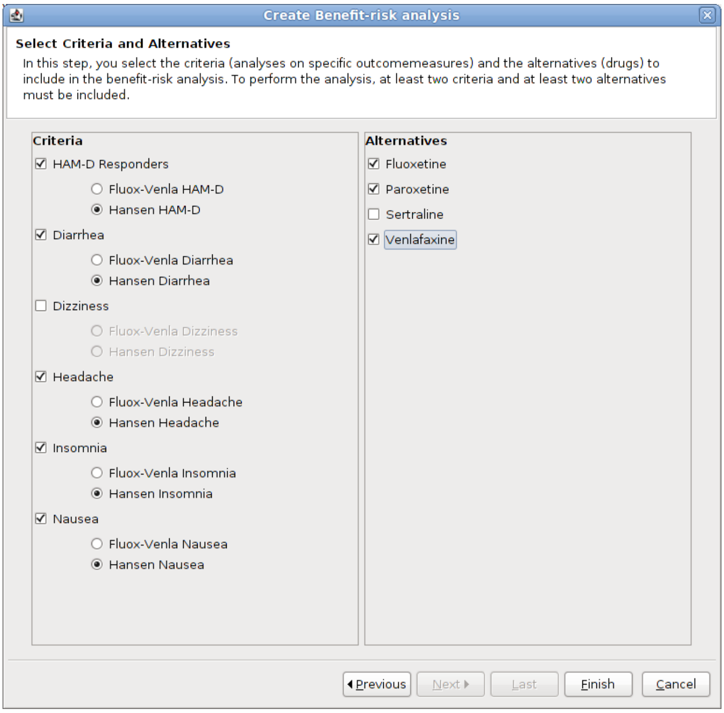


Figure 9.8: Visualization of standard meta-analysis results as a forest plot [Lewis and Clarke, 2001]. Here, odds-ratios (95% confidence intervals) are plotted on a logarithmic scale, with the pooled estimate shown last.


ADDIS v1.6 - Example Data
File
Edit
Add
Help
New Study
New Pair-wise meta-analysis
New Network meta-analysis
New Benefit-risk analysis
Franchini et al, 1997
Gagiano 1993
Goldstein et al, 2002
Hong et al, 2003
Kavoussi et al, 1997
Kiev and Feiger, 1997
Lepola et al, 2003
McPartlin et al, 1998
Mehtonen et al, 2000
Montgomery et al, 2004
Nemeroff et al, 1995
Newhouse et al, 2000
Patris et al, 1996
Rapaport et al, 1996
Rudolph and Feiger, 199
Schatzberg et al, 2002
Schone and Ludwig, 199
Sechter et al, 1999
Silverstone and Ravindra
Tylee et al, 1997
Weihs et al, 2000
Pair-wise meta-analyses
Fluox-Venla Diarrhea
Fluox-Venla Dizziness
Fluox-Venla HAM-D
Fluox-Venla Headache
Fluox-Venla Insomnia
Fluox-Venla Nausea
Network meta-analyses
Hansen Diarrhea
Hansen Dizziness
Hansen HAM-D
Hansen Headache
Hansen Insomnia
Hansen Nausea
Benefit-risk analyses
Benkert HAM-D/weight g
Fava et al, 2002 BR
Fluox-Venla BR
Hansen BR
Overview
Consistency
Inconsistency
Node Split
Memory Usage
Network Meta-Analysis (Consistency Model)
Rank Probability
Rank 1 is best, rank N is worst.
Probability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Fluoxetine
Paroxetine
Sertraline
Venlafaxine
Treatment
Rank 1
Rank 2
Rank 3
Rank 4
Save Image
Variance Parameters
Parameter
Median (95% CrI)
Running 0 jobs for 0 user tasks
www.drugis.org

Fluoxetine	1.22 (0.95, 1.56)	1.24 (0.95, 1.63)	1.37 (1.10, 1.69)
0.82 (0.64, 1.05)	Paroxetine	1.02 (0.70, 1.49)	1.12 (0.82, 1.53)
0.80 (0.61, 1.05)	0.98 (0.67, 1.43)	Sertraline	1.10 (0.83, 1.45)
0.73 (0.59, 0.91)	0.89 (0.65, 1.22)	0.91 (0.69, 1.21)	Venlafaxine

Figure 9.9: Network meta-analysis results. The table gives (posterior) odds-ratios (95% credibility interval) for all treatments relative to each other. The bar chart visualizes the (posterior) probability for each treatment to be best, second-best, etc. given the analysis model and the data.



Create Benefit-risk analysis

Select Criteria and Alternatives

In this step, you select the criteria (analyses on specific outcomemeasures) and the alternatives (drugs) to include in the benefit-risk analysis. To perform the analysis, at least two criteria and at least two alternatives must be included.

Criteria

- ☑ HAM-D Responders
  - ○ Fluox-Venla HAM-D
  - ● Hansen HAM-D
- ☑ Diarrhea
  - ○ Fluox-Venla Diarrhea
  - ● Hansen Diarrhea
- ☐ Dizziness
  - ○ Fluox-Venla Dizziness
  - ○ Hansen Dizziness
- ☑ Headache
  - ○ Fluox-Venla Headache
  - ● Hansen Headache
- ☑ Insomnia
  - ○ Fluox-Venla Insomnia
  - ● Hansen Insomnia
- ☑ Nausea
  - ○ Fluox-Venla Nausea
  - ● Hansen Nausea

Alternatives

- ☑ Fluoxetine
- ☑ Paroxetine
- ☐ Sertraline
- ☑ Venlafaxine

Figure 9.10: Criteria selection screen for construction of a benefit-risk model with synthesized evidence.

criteria **treatments**	1	2	≥ 2
2	PMA	L&O (S/PMA/NMA)	SMAA (PMA/NMA)
≥ 2	NMA	SMAA (S/NMA)	SMAA (S/NMA)

Table 9.4: Supported methods. Abbreviations: S = Single-study, PMA = Pair-wise meta-analysis, NMA = Network meta-analysis, L&O = Lynd & O'Brien benefit-risk, SMAA = SMAA-based benefit-risk.

decision between two alternatives based on a single study, standard statistical methods are sufficient. When there are several studies, pair-wise meta-analysis [Normand, 1999] can be used to pool the evidence, and for more than two alternatives network meta-analysis is needed [Lu and Ades, 2006]. When two criteria (e.g. one benefit and one risk) and two alternatives are to be considered, the "Lynd & O'Brien" model [Lynd and O'Brien, 2004] based on either a single study or two meta-analyses (one for each criterion) can be used. For more than two alternatives or criteria Stochastic Multicriteria Acceptability Analysis (SMAA) based models are available [Tervonen et al., 2011, van Valkenhoef et al., 2012e]. The SMAA methods used are described in [Lahdelma et al., 1998, Lahdelma and Salminen, 2001] and their computational details in [Tervonen and Lahdelma, 2007]. These can be based on either a single study, pair-wise meta-analyses (limited to 2 alternatives), or network meta-analyses (for $\geq$ 2 alternatives).

The results of a "Lynd & O'Brien" benefit-risk analysis are visualized both through plotting points from the probability distributions of incremental benefit and risk on the benefit-risk plane, and through the benefit-risk acceptability curve [Lynd and O'Brien, 2004]. The SMAA models are visualized using the JSMAA visualization components and tables, showing the rank acceptabilities (Figure 9.11) to indicate how likely the alternatives are to obtain a certain rank (from best to worst) and the central weights to indicate what preferences typically support specific alternatives.

9.5 Discussion

In this paper we introduced ADDIS, a decision support system for evidence-based medicine. ADDIS was developed in the context of a scientific project aimed to enable better use of information technology in the transfer and analysis of clinical trials design and results. The long term vision was developed in collaboration with a steering group composed of experts from the pharmaceutical industry, academia and the regulatory environment. Short term plans were developed with our 'customer', a regulatory assessor who oversaw the development. The design was further informed by several (completed and ongoing) case studies, such as a study of the benefit-risk profiles of second generation anti-depressants. Although the software has been presented to and used by experts in the field, no formal validation or usability studies have been conducted so far.

We presented a unifying data model for aggregate trial results, which is at the core of ADDIS. The model enables semi-automated generation of evidence synthesis

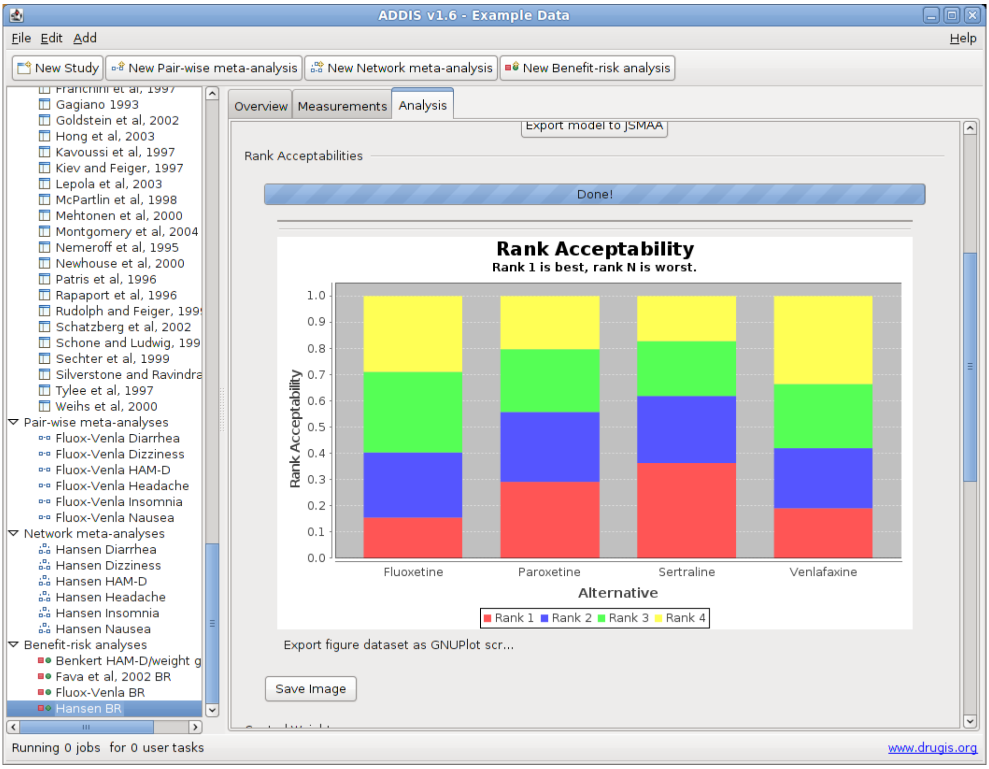

ADDIS v1.6 - Example Data
File Edit Add
Help
New Study
New Pair-wise meta-analysis
New Network meta-analysis
New Benefit-risk analysis
Franchini et al, 1997
Gagiano 1993
Goldstein et al, 2002
Hong et al, 2003
Kavoussi et al, 1997
Kiev and Feiger, 1997
Lepola et al, 2003
McPartlin et al, 1998
Mehtonen et al, 2000
Montgomery et al, 2004
Nemeroff et al, 1995
Newhouse et al, 2000
Patris et al, 1996
Rapaport et al, 1996
Rudolph and Feiger, 199
Schatzberg et al, 2002
Schone and Ludwig, 199
Sechter et al, 1999
Silverstone and Ravindra
Tylee et al, 1997
Weihs et al, 2000
Pair-wise meta-analyses
Fluox-Venla Diarrhea
Fluox-Venla Dizziness
Fluox-Venla HAM-D
Fluox-Venla Headache
Fluox-Venla Insomnia
Fluox-Venla Nausea
Network meta-analyses
Hansen Diarrhea
Hansen Dizziness
Hansen HAM-D
Hansen Headache
Hansen Insomnia
Hansen Nausea
Benefit-risk analyses
Benkert HAM-D/weight g
Fava et al, 2002 BR
Fluox-Venla BR
Hansen BR
Overview
Measurements
Analysis
Export model to JSMAA
Rank Acceptabilities
Done!
Rank Acceptability
Rank 1 is best, rank N is worst.
Rank Acceptability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Fluoxetine
Paroxetine
Sertraline
Venlafaxine
Alternative
Rank 1
Rank 2
Rank 3
Rank 4
Export figure dataset as GNUPlot scr...
Save Image
Running 0 jobs for 0 user tasks
www.drugis.org

Figure 9.11: SMAA benefit-risk analysis results. The bars indicate the probability that each treatment is the best, second best, etc., given the preferences specified by the decision maker. In this case, the results indicate that there is a lot of uncertainty regarding which alternative is the best, but sertraline and paroxetine are somewhat more likely to be than venlafaxine and fluoxetine.

and benefit-risk models implemented in ADDIS. All these components together allow for re-usable, re-analyzable repositories of trials and analyses to be maintained and shared among users. The value of the unifying data model is not to model the domain in detail, but to provide a uniform basis for automated evidence synthesis and decision modeling. As such, specific decision support systems may use domain specific information to further assist the decision maker. ADDIS makes use of some domain knowledge to support its primary goal: to enable the direct and indirect assessment of the comparative benefits and risks of different drugs based on all available evidence from clinical trials. For example, Arms always have a Dosing characteristic, and studies have a fixed list of characteristics that are relevant for clinical trials comparing the efficacy and safety of drugs.

Multiple data models have been proposed for comprehensively storing information on the design and outcomes of clinical trials, e.g. the ClinicalTrial.gov DED and the CDISC standards. The minimal unifying data model implemented in ADDIS is not competing with these, but rather provides a target for conversion from them in order to enable semi-automated generation of evidence synthesis and decision models operating on the trial results. Traditionally the systematic reviewing process to perform a (network) meta-analysis takes a considerable amount of time and effort. While ADDIS does not address this problem directly, it does provide a uniform platform for analysis and data sharing that obviates the need for repeated data extraction.

To the best of our knowledge, ADDIS is the first system to implement decision models that are directly and explicitly based on the (synthesis of) clinical trials results. By making the involved trade-offs and the link between trial results and decision model recommendations visible, ADDIS can enable more transparent strategic health care decision making. ADDIS can also help in improving the reporting of systematic reviews since the included trials are represented explicitly, rather than only in data tables pre-processed for the purpose of evidence synthesis. The decisions made in mapping the data and applying the evidence synthesis models are thus clearly represented.

9.5.1 Limitations and future work

The decision modeling in ADDIS is based on the assumption that a structured database of relevant clinical trials is available. However, to acquire such a database is a difficult and time consuming. The initial phase of development has focused on drug regulation – a use case for which it is reasonable to assume that the data will be provided in whatever format requested. For other use cases, such as guideline formulation, this assumption is not justified. If the data is not available in a suitable format, a systematic review will have to be performed and the data input into ADDIS mostly manually, although the ClinicalTrials.gov import functionality can reduce the required work. However, once the input is done, the data is more valuable than the same set of trials extracted for e.g. Cochrane RevMan, as they can be reused for different types of analyses. To make ADDIS a useful tool for a wider audience, functionality that further increases the efficiency of systematic reviewing should be added, possibly by implementing automated information extraction methods.

Until now, approximately 100 clinical trials were entered for the case studies. To assess the usefulness of ADDIS in various medical domains more trials should be entered. However, as their input is mostly manual, this is an expensive and time-consuming process. Also, as the trial database gets larger, the study selection step for evidence synthesis can get cumbersome with the current implementation. More intelligent study matching/filtering (e.g. with the different characteristics) should be explored for lowering the user's work load. This may require explicit modeling of some of the aspects that are currently stored as plain text, such as the patient eligibility criteria.

The scope of the unifying data model could be extended to support other types of evidence synthesis, such as meta-regression and stratified analyses. These possible extensions may introduce covariates at different levels, e.g. the time at which an outcome measure was assessed, the dosage level for a treatment, the baseline severity of illness in an arm, the length of the placebo washout phase of a study, or within-arm correlation of two or more outcome measures. As such, it will be a challenge to introduce these rather complex distinctions without making the generation of (network) meta-analyses impossible.

ADDIS enables generation of benefit-risk decision models that use aggregate level, possibly synthesized, clinical trial data as part of their input. However, health care decisions can include evaluation dimensions not reported in clinical trials (e.g. convenience of administration or storage), which consequently can not be included in ADDIS. Also, economical decision models applied in health technology assessment often do take into account the primary clinical endpoints of interest with high quality evidence, but seldom include high-quality adverse event sources [Cooper et al., 2005]. We acknowledge that adverse event reporting in general is inferior to clinical endpoint reporting due to various reasons. These include the rareness of some adverse events, the fact that most clinical trials are powered to show efficacy (which typically requires smaller sample size than detecting adverse events) and inconsistent reporting of adverse event data [Eichler et al., 2008]. Decision models based on evidence synthesis can help improve the included evidence on adverse events, but it may be necessary to include other evidence sources to consider the rarest events. To consider these and other use cases, future research should address semi-automated generation of a wider range of decision models and their implementation in ADDIS.

CHAPTER 10

Discussion

There is a clear gap between the evidence from clinical trials as published in various journals on the one hand, and the integrated view of the evidence required by decision makers on the other. Finding the evidence and applying it to the decision in hand is difficult and time-consuming. Aggregate Data Drug Information System (ADDIS) was developed to bridge this gap by automating the three steps of evidence-based decision making: data acquisition, evidence synthesis, and decision modeling. By starting from a structured database of aggregated clinical trial results, decision support can be provided in an on-demand setting. Once structured data are more widely available, ADDIS will be a powerful knowledge system for health care policy decision makers.

This thesis discusses the methods underlying the ADDIS decision support system for evidence-based drug benefit-risk decision making, as well as the ADDIS system itself. This chapter evaluates the developed methods by assessing how they answer the research questions posed in Chapter 1, and identifies their limitations. Then, the work that is currently ongoing is briefly described and directions for future work are given. Finally, some overall conclusions and recommendations are provided.

10.1 Answers to research questions

The central question of this thesis is:

> How can a network of clinical trials be used to inform benefit-risk assessment in a formal decision modeling framework, and how can information systems support such decision modeling?

The proposed ADDIS decision support system gives the answer to this question (Chapter 9). Several novel algorithms, methods, and data models were developed to make this possible (Chapters 3–7). These will now be briefly reviewed according to the sub-questions posed in Chapter 1.

10.1.1 Automating network meta-analysis

Comprehensive decision making regarding alternative treatment options must take into account all available evidence. In many cases more than two treatment options need to be considered, and the available clinical trials form a complex *network* of evidence. Network meta-analysis provides a framework for the *coherent* or *consistent* assessment of the (relative) effects of the treatment options based on a network of clinical trials. In the past conducting a network meta-analysis required the manual specification of a complex statistical model which is an obstacle when using network meta-analysis in decision making, and especially for an automated decision support system such as ADDIS.

Therefore, we set out to identify the problems that must be addressed to automate Bayesian (network) meta-analysis (Chapter 3). The problems are (1) to generate the model structure based on the nature of the data and the structure of the evidence network, (2) to specify adequate (vague) prior distributions in case no informative prior can be given, and (3) to generate over-dispersed starting values for independent Markov chains to enable the assessment of convergence. We provide solutions to these problems and automatically generate the basic random effects consistency model for both continuous and dichotomous data. The methods do not currently allow for other types of data such as time-to-event data, but could be easily extended to incorporate these. The data are also assumed to be available in absolute (per-arm) format, but in some cases only relative effect data is available. Again, the proposed methods could be easily extended to overcome this limitation. The limitation to consistency models is more fundamental, and while the solution to problems (2) and (3) are easily transferable to inconsistency and node-splitting models, problem (1) will require a different solution for each of these models.

Chapter 4 attacks the problem of generating an adequate model structure for inconsistency models. It formulates the problem of identifying the correct parameterization in a graph theoretic framework, and derives an algorithm from that formulation. While the algorithm gives appropriate parameterizations, there are several limitations. First, and most importantly, a certain definition of what constitutes 'potential inconsistency' was used, but other definitions are possible. The question which definition is the most appropriate still needs to be answered. The question can even be asked whether a solution to the parameterization problem under that definition always exists. Second, the algorithm has exponential worst-case complexity, and although it produces a model quickly on the evidence networks encountered thus far, sooner or later networks will be found on which it takes exceedingly long to come up with a result.

The work presented in this thesis represents the first method for automated model generation of (Bayesian) network meta-analysis. As such, it is an important innovation that enables a larger group of researchers to perform a reliable network meta-analysis. The developed software takes the tedious, time-consuming, and error-prone aspects of network meta-analysis model specification out of the hands of the researcher. At the same time, it provides an opportunity to draw the analyst's attention to important issues such as choice of prior distributions and problems with the data set that might otherwise not be noticed. Unfortunately, easy to use software also presents an opportunity for misuse of the methods by those unaware of its limitations, but hopefully this can be addressed by adequate documentation and safeguards against common mistakes.

10.1.2 Decision analysis for drug benefit-risk assessment

In decision analysis for drug benefit risk assessment we must distinguish the different alternatives (i.e. drugs), multiple criteria (i.e. assessments of efficacy and safety), the measurement of the performance of each alternative on the criteria, and the preferences of the decision maker. Criteria measurements are most often derived from clinical trials data and are therefore inherently uncertain. Any reliable decision making framework for drug benefit-risk assessment must take into account these different aspects. In Chapter 5, we show how the Stochastic Multicriteria Acceptability Analysis (SMAA) methodology can be applied to the assessment of benefit-risk. SMAA is an Multiple Criteria Decision Analysis (MCDA) method, and to apply it the MCDA approach to decision structuring has to be followed: (1) define the criteria (e.g. efficacy, adverse events), (2) define the alternatives (e.g. competing drugs), (3) measure the performance of each alternative on each criterion, (4) find a lower and an upper bound for each criterion, and (5) define partial value functions for each criterion. In step 3, we find probability distributions for the criteria values through statistical analysis of the underlying trial(s). We assumed these distributions to be independent, but they could be correlated across alternatives (Chapter 8), across criteria, or both. Because the probability distributions derived in step 3 could have infinite support, finding the scale bounds in step 4 is (strictly speaking) impossible. Instead, we use the upper and lower bounds of the 95% confidence interval to ensure that the decision metrics are at least valid for the 95% most likely values. As we argued in Chapter 5, this approach to defining the criteria scale bounds can be useful even if the support is not infinite, because this ensures that the scales are representative of the true range of values present in the data rather than the theoretically possible range of values. To define the partial value functions (step 5), we used simple linear scaling from the scale lower bound to the upper bound. However, this may not always be appropriate, and in those cases alternative partial value functions must be defined. This is more difficult in the probabilistic setting, as some values may fall outside the range defined by the scale upper and lower bound, and these cases must be handled in some way. Different solutions are possible, for example mapping values outside the defined range to the extreme values, but then it would be wise to use a wider confidence interval to define the scale upper and lower bound (e.g. the 99% confidence interval).

The above outlines the challenges of applying any MCDA method to drug benefit-risk assessment, and the answers to these challenges developed in this thesis. SMAA, however, has some unique features to support the decision maker. First, it enables an 'inverse approach' to decision making. In this approach, the decision maker does not provide any preference information, and SMAA derives the preferences that would 'typically' support each of the alternatives. This may help to exclude some alternatives, as it may turn out that they are always inferior to other alternatives. In addition, it helps the decision maker to map the alternatives to their stronger and weaker points, which may already be sufficient to make a decision. In that case, challenging and time-consuming preference elicitation can be avoided. Second, SMAA enables partial or imprecise preference information to be used by also defining a probability distribution over the weights that are compatible with the given preference information. The preference information can then be refined until a decision can be made, avoiding unnecessary effort in eliciting precise preferences. Moreover, imprecise preferences are more robust because they represent a range of compatible weights, thus reducing sensitivity to specific and overly precise weight information. In theory, any constraints on the weights are possible, but in practice efficient algorithms are required to sample from the probability distribution over the feasible weight space. So far, efficient algorithms had been given only for specific types of constraints, limiting the flexibility of the SMAA method in practice. This was addressed in Chapter 6, which developed a method to efficiently sample weights with arbitrary linear constraints. As was shown in Chapter 7, this enables more flexible decision support using SMAA.

Decision aiding using any MCDA method, and perhaps especially SMAA, is complex and requires some understanding of the underlying methods and steps to be taken in the analysis process. To implement a SMAA decision model would certainly be too complex and time consuming for most decision makers. However, implementation of these methods in ADDIS automates most of the difficult steps and allows the decision maker to focus on evaluating the decision problem. One important limitation of the decision modelling in ADDIS is that all criteria must have been measured in clinical trials. This may not always be the case, for example convenience of administration or cost of treatment are criteria that are externally derived. However, this is a shortcoming of the current implementation, and not of the SMAA method.

Notwithstanding these limitations, the work presented in this thesis provides the first decision modelling method for drug benefit-risk assessment that can take into account an arbitrary number of alternatives and criteria as well as the uncertainty inherent in criteria measurements taken from clinical trials *and* imprecise or partial preference information. All of these aspects are important for decision modelling in benefit-risk assessment, and the presented methods could pave the way for more transparent and well-documented decision making by regulatory authorities. This, in turn, will enable trust between regulatory authorities, pharmaceutical industry, academia, patients, and the public.

10.1.3 Using network meta-analysis in benefit-risk assessment

When making a benefit-risk decision involving multiple alternative drugs, the decision maker naturally wants to take into account all available evidence. The natural tool to do this is network meta-analysis. Network meta-analysis estimates probability distributions for the *differences* in treatment effects observed in clinical trials. Thus it answers, for all *pairs* of drugs in the analysis, the question 'how much better (or worse) is drug X than drug Y?'. As input for the decision analysis, we choose one drug as the comparator, and derive a joint distribution for the relative effects of the other drugs using network meta-analysis (Chapter 8). However, to make a clinically relevant assessment of the trade-offs, the scales must be defined in absolute terms. It would not make sense to ask the decision maker to choose between 'improving the odds-ratio for remission from 0.9 to 1.2' and 'improving the odds-ratio for stroke from 1.3 to 0.8', as these questions are meaningless without knowing the underlying incidence for the comparator drug. Although the odds-ratios appear to be similar, if the comparator drug has a very different incidence of stroke than of remission, the numbers of patients affected are also very different. Therefore, we show how a distribution for the absolute incidence for the comparator drug can be used together with the joint distribution of relative effects to derive a distribution for the absolute incidence for all included drugs. Although the problem was illustrated using the odds-ratio, using a relative measure of performance is problematic independent of the specific scale being used. A similar technique for converting relative measurements to absolute ones can be used for continuous outcomes.

There are two important challenges in applying the method presented in Chapter 8: the assessment of consistency in the evidence network and the estimation of baseline incidences. As was already pointed out above, the assumption of consistency of relative effects across all studies in the network is central to network meta-analysis. If this assumption is violated, the data can not be used to inform a decision. Thus, methods to detect inconsistency must be applied before the decision model is constructed. This was also illustrated in Chapter 8, and is facilitated by the automated network meta-analysis inconsistency models in ADDIS. To estimate baseline incidences, a variety of techniques can be applied, such as using expert judgment, using data from observational studies, or using the baseline arm from the trials in the network. ADDIS currently only implements the latter, and does this automatically. In the future, additional modes for baseline estimation should be implemented.

While the work on decision modelling based on a single trial (Chapter 5) already enables transparent evidence-based decision modelling, it did not allow taking into account all available evidence. Combining network meta-analysis and SMAA decision modelling makes this possible, and completes the methodological aspects of the thesis. This combination is a non-trivial exercise that requires a deep understanding of both methods to be performed correctly and in a generalizable way.

10.1.4 Storing aggregated clinical trial results

As demonstrated in Chapter 2, there are currently no systems that store aggregated clinical trial results in a way that is appropriate for on demand re-analysis of the results. Due to this, decision modelling as proposed in this thesis is difficult and time consuming to perform, because data must be identified and extracted manually, a process that can take weeks or months. This is mainly caused by the lack of an appropriate data model for storing aggregated clinical trial data in a reusable manner (Chapter 9). As the purpose of ADDIS is to enable re-analysis of the trials stored in the system, it must address this problem.

As outlined in Chapter 9, the approach is to define a minimal data model that allows the on demand application of evidence synthesis and decision modelling to the data. The ADDIS system itself implements a more elaborate version of this data model, that makes some additional assumptions about the included clinical trial data. The current ADDIS data model has severe limitations, including a lack of support for coding systems commonly used in medical research, not modelling some aspects of the trial such as patient eligibility criteria (instead representing them as text), and not modelling the deep structure of outcome variables (instead assuming them to be atomic). However, the distinction between the minimal data model for evidence synthesis and decision modelling on the one hand and more fine-grained representations of clinical trials on the other, is the key to enable the useful representation of data *now*.

10.2 Ongoing and future work

Work is currently ongoing with the Multi-Parameter Evidence Synthesis group at Bristol University to further automate models for checking consistency in network meta-analysis. This includes automated generation of node-splitting models Dias et al. [2010], as well as further investigation into the definition and interpretation of inconsistency that should lead to more insight into inconsistency models Lu and Ades [2006]. This should enable improving upon the work presented in Chapter 4. The development of ADDIS is continuing at the University of Groningen and the University Medical Center Groningen, and is currently focussed on improvements to the current functionality to best position the proof-of-concept. Additional funding is being sought to develop the proof-of-concept system into a marketable product or service. At the same time, a commissioned project funded by TI Pharma is underway to evaluate the usefulness of ADDIS for pharmaceutical companies, and to increase compatibility of the ADDIS data model with current industry standards and practices.

In future work, I want to automate more advanced evidence synthesis methods Sutton et al. [2008], such as network meta-regression to take into account effect-modifying covariates, or different types of data such as time-to-event data. This serves three main goals: it enables a broader audience of researchers to apply these advanced evidence synthesis models, it generalizes the decision modelling capabilities of ADDIS, and it extends the minimal data model for evidence synthesis and decision modelling, which in turn defines the requirements for the ADDIS data model. This would significantly broaden the audience for ADDIS from drug regulation to reimbursement, guideline formulation, early drug development, and prescription decisions. In addition, future work should address the data acquisition problem. This will likely require a combination of automated systems for abstract processing and data extraction, manual refinement and completion of automated extractions, and an infrastructure to share extracted datasets between researchers and institutions. Ideally, this would eventually lead to all clinical trials, systematic reviews, and decision models having their data sets published in a structured and reusable format.

10.3 Conclusions

The ADDIS decision support system developed in this thesis shows that a new and better way of evidence-based decision making becomes feasible when trial data are available in a more structured format. However, as was shown in Chapter 2, this situation is currently far from reality. ADDIS is not a direct solution to this problem, but it does bring a solution closer in two fundamental ways. First, as a proof-of-concept, ADDIS makes the benefits of a more comprehensive and structured repository of clinical trials evidence more concrete, thereby offering a strong motivation to work on this problem. Second, the work on ADDIS helps to identify the requirements that data models for aggregated clinical trial results must satisfy to be useful.

To enable this proof-of-concept, a number of innovative applications of existing methods, as well as the development of novel algorithms and data models were required. In this thesis, it was shown how a database of clinical trials can be combined with network meta-analysis and SMAA decision modelling to support drug benefit-risk assessment in a methodologically sound approach. That these methods can be applied in practice was shown through a series of case studies, and their generalizability is demonstrated by their implementation in ADDIS, which allows the same methods to be readily applied to other cases.

Because of its advanced features, the ADDIS decision support system is also useful independent of the ambitions that motivate its existence. For example, it could be used in systematic reviewing to manage the data from included clinical trials. In this role, ADDIS is currently the only such software that enables network meta-analysis of the clinical trials. ADDIS is also useful for decision makers that can request data to be delivered in the ADDIS data format. This could apply to marketing authorization decision makers, who ask the submitting pharmaceutical company to deliver the data, and for reimbursement authorities, who can commission a systematic review to be undertaken in order to support the decision.

APPENDIX A

Clinical trials information in drug development and regulation

Existing systems and standards

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Clinical trials evidence in drug development and regulation: a survey of existing systems and standards. SOM Research Report 12003-Other, School of Management, Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands, 2012b. URL `http://irs.ub.rug.nl/dbi/4fcf224db9977`

Abstract

Clinical trials provide pivotal evidence on drug efficacy and safety. The evidence, information from clinical trials, is currently used by regulatory decision makers in marketing authorization decisions, but only in an implicit manner. For clinical trials information to be used in a transparent and accountable way, it must be available in a format enabling efficient access and further processing, so that decisions can be explicitly linked to the underlying evidence. Thus, processing and management of clinical trials information plays a critical role in enabling transparent decisions.

With the aim of identifying bottlenecks that prevent transparent decision making based on clinical trials evidence, we review the information systems and data standards that process clinical trials data in drug development and regulation. We find that while systems and standards for the management of single clinical trials are relatively mature, the transfer of information to the public and to decision makers is still an ad-hoc and text-based affair, and the integration of data from multiple studies remains difficult.

A.1 Background

The drug lifecycle consists of the discovery, clinical development, market authorization and marketing phases [Rang, 2005]. In discovery, promising candidate drugs (leads) are identified (often using computational methods) and evaluated in several phases using different preclinical methods. If a lead is likely to be both efficacious (i.e. it likely has the capacity to produce a therapeutic effect) and safe in humans, it may enter the clinical phase. In clinical development, the drug is evaluated in humans through clinical trials, first primarily for safety (phase I), and then for efficacy (phase II and III). The phase III clinical trials are confirmatory trials that demonstrate the efficacy and adverse event profile of the drug in comparison to placebo, another drug, or both. In the market authorization phase, the evidence from discovery and development presented by the pharmaceutical company is assessed by a team of experts assembled by the regulatory authority (e.g. the Food and Drug Administration (FDA) in the US or the European Medicines Agency (EMA) in the EU). In the end, the regulators evaluate whether the drug has a favorable benefit-risk profile (i.e. the favorable effects are likely to outweigh the unfavorable effects). The pivotal evidence in this assessment is provided by the phase III trials. If the drug is approved by the regulatory authority, it enters the marketing phase. Clinical trials are also performed in the marketing phase, either by the pharmaceutical company or by others, often in the setting of a risk management plan. The outcome of these trials may trigger a re-evaluation by the regulatory authorities, that can lead to suspension of withdrawal of a drug from the market.

The development of new pharmaceuticals is often hampered by failure in their late-development phase, market authorization or subsequent clinical use. For example, in 2009 40% of the drugs submitted for approval in Europe received a negative opinion or were withdrawn before receiving an opinion [Eichler et al., 2010]. In many cases, failure may be prevented, or development may be stopped earlier, through better communication of the requirements for clinical evidence and the methods used to assess benefit-risk [Liberti et al., 2010, Coplan et al., 2011, Guo et al., 2010]. The accessibility of key information on past decisions and clinical trials could facilitate this process greatly.

However, presently there are clear information gaps with respect to the efficiency, transparency and reproducibility of the current drug development, approval, and regulation processes, including the secrecy surrounding data [Roberts et al., 1998], the lack of consistency, transparency, and reproducibility of the methods used to reach conclusions about benefit-risk [EMEA, 2007] and, finally, the insufficient communication of important information to patients and professionals [Irs et al., 2004].

Since the results of clinical trials serve as the main sources of information regarding new medicines, a comprehensive overview of the various information systems that store and process this information could identify the gaps in the information transfer that have led to the problems mentioned above. With this aim, we provide an overview of the existing systems and standards supporting the management and transfer of information from clinical trials. For a brief overview of drug information systems that considers the entire drug information life cycle, we refer to [Tervonen et al., 2010].

A.2 Clinical trial information systems

First, in Sections A.2.1 and A.2.2, we take the industry perspective and focus on the management of trial information within the pharmaceutical company. Section A.2.3 introduces trial registration and Section A.2.4 describes regulatory submission and assessment. In Section A.2.5 we concentrate on the systems that provide the product information for approved drugs, such as package inserts. Standards and data models for clinical trials information are discussed in Section A.2.6 and controlled terminologies in Section A.2.7.

A.2.1 Operational management

Operational management refers to the administrative and data-gathering activities surrounding a single trial. The operational management of clinical trials can be automated by using Electronic Data Capture (EDC), previously also called Remote Data Entry/Capture [Bleicher, 2003]. EDC can be defined as a computerized replacement for (paper-based) Case Report Forms (CRFs), in which information is entered into a database through a computer entry form [El Emam et al., 2009]. Normally data are validated before being accepted into a database to catch possible data entry errors early on (validation and cleaning), while an audit trail is provided for all data entries and modifications [El Emam et al., 2009]. Clinical Data Management Systems (CDMSs) offer a comprehensive solution to trial data management by including data from multiple sites into a single database. This may be done by means of EDC systems, but also by entering data from paper CRFs, usually through double data entry to prevent transcription errors. Clinical Trial Management Systems (CTMSs) are an extension of CDMSs and provide integrated solutions for the management of clinical trials, including advanced features such as subject recruitment tracking and randomization as well as medication inventory tracking. Real-world systems may overlap the distinctions between EDC system, CDMS and CTMS, and the terms are often used interchangeably. The term EDC, for example, may generally refer to all or most of the functionality described above [El Emam et al., 2009].

Until recently, the management and data collection of the vast majority of clinical trials were paper-based activities. In 2000, only 12% of trials were using an EDC system [CDISC, 2005]. However, this number has significantly increased over the last years, to 20% in 2004 [CDISC, 2005] and 41% in 2007 [El Emam et al., 2009]. It seems likely that this trend will continue in the foreseeable future, especially given the fact that in 2007 96% of the pharmaceutical companies and 71% of the Clinical Research Organizations (CROs) were to some extent using EDC technology [Tufts, CSDD and CDISC, 2007]. Because of the increasingly international character of many trials, it is not surprising that the advantages of web-based EDC are becoming widely recognized [Paul et al., 2005].

Therefore there is a clear need for interoperability of the different operational management systems. In the area of electronic source data (see e.g. [EMEA, 2007, Marks, 2004]), several technologies have gained momentum, such as Electronic Case Report Forms (eCRFs), Electronic Patient Reported Outcomes (ePROs), and Electronic Laboratory Data (eLab) [Tufts, CSDD and CDISC, 2007]. Now that data are being collected more and more in electronic form, there are efforts underway to automatically enter data from Electronic Health Record (EHR) systems in EDC systems in order to reduce the amount of information that has to be entered manually [Prokosch and Ganslandt, 2009, ClinPage, 2009]. It is believed that by reducing the necessity of frustrating time-consuming tasks such as double data entry, the reluctance of sites to participate in clinical trials will decrease [ClinPage, 2009].

Furthermore, while statistical analyses and report generation of clinical trials have been computer-aided for a long time now, they usually require programming of statistical routines. Analyses are performed using general-purpose statistics programs such as SAS, SPSS, R and SPlus, and are therefore not integrated with CTMSs. Thus, to extract data from the CTMS in a format suitable for statistics programs, intermediate processing is required. However, this approach leads to a loss of the traceability of the summary statistics back to the underlying source data.

Market share information about CTMSs is hard to come by, but a 2001 study indicated that Clinsoft Corporation (now acquired by Phase Forward) and Oracle Corporation (`http://www.oracle.com`) dominate the market. Oracle has recently acquired Phase Forward [Oracle Corp., 2010]. These commercial packages have been criticized for only being concerned with the delivery of valid and accurate data that conform to the Good Clinical Practice (GCP) guidelines, and neglecting end-to-end processes as well as the usability of the interface and interoperability with other systems [Oliveira and Salgado, 2006]. Moreover, CTMSs can be prohibitively expensive, require considerable expertise to set up, and need specialized in-house IT support to operate [Oliveira and Salgado, 2006]. Independent groups or organizations based in developing countries may not be able to afford CTMSs for these reasons, and there is little to no good information about the available commercial offerings [Fegan and Lang, 2008]. One solution that has been proposed is the development of an open source CTMS through a collaborative effort of research organizations and funders [Fegan and Lang, 2008]. OpenClinica (`http://www.openclinica.org`) provides an open source solution for web-based clinical data management and is compliant with the regulatory requirements. Similar functionality is provided by the caBIG Clinical Trials Suite (`http://cabig.nci.nih.gov/adopt/CTCF`) of the National Cancer Institute (NCI), although it is primarily focused on the cancer domain. There are several academic clinical data management systems, e.g., TrialDB [Nadkarni et al., 1998, Brandt et al., 2000, Nadkarni et al., 2010], COATI [Oliveira and Salgado, 2006], and OpenSDE [Los et al., 2005].

The field of CTMSs has gradually started with the transition from data-centric single study systems toward more interoperable, comprehensive and standards-based systems. This development was initially spurred by a push from the FDA for electronic submissions (see Section A.2.6), but has now taken on a life of its own, with the industry and CROs increasingly acknowledging the benefits of information technology [Tufts, CSDD and CDISC, 2007].

A.2.2 Data warehousing

A data warehouse for clinical research should enable analyses between trials and across compounds. Clinical trials data warehousing offers unique challenges owing to the diversity inherent in the domain. These challenges include establishing a shared vocabulary of the clinical domain to ensure consistent reporting and capturing the clinical domain in a data model that is sufficiently rich to enable data mining, synthesis, and the analysis of data across different trials. Several academic clinical data management systems have incorporated data warehousing features (e.g. TrialDB, COATI, OpenSDE). Phase Forward, the company behind ClinTrial, offers a data warehousing solution under the name ‘Clinical Data Repository’ [Phase Forward, Inc., 2010], which integrates data from different sources through a consistent audit trail, adopting a meta-data based approach to validation. Although Oracle Clinical facilitates the management of multiple studies in a single database, it is primarily aimed at reducing maintenance overhead rather than at enabling analysis across trials. Other data warehousing projects include the academic METABASE [Swertz et al., 2010], Organon D3W [Vervuren, 2005, Vervuren and Dietvorst, 2006], the Cancer Biomedical Informatics Grid (caBIG) CTODS/Cactus, and the Human Studies Database (HSDB) project [Sim, 2008]. ISO has recently published a standard for the deployment of a clinical data warehouse [ISO TC 215 (Health Informatics), 2011], but it focusses on health care rather than research.

A.2.3 Trial registration

When judging the merits of a treatment, it is critical that all relevant existing clinical trials can be efficiently identified. However, until recently, the scientific literature was the only public source of clinical trial results. This caused difficulty in finding relevant trials and led to insufficient or inaccurate trial reporting and publication bias [Dickersin and Rennie, 2003]. In the late 1990s and early 2000s, many countries world wide adopted legislation that requires the design (research plan) of clinical trials to be registered *before* participants are recruited. The prospective registration of clinical trials ensures that the existence of clinical trials is known, even if their results are not published in the peer-reviewed scientific literature. This enables publication bias to be detected more easily. In the US, investigators are additionally required to register the trial results in the ClinicalTrials.gov registry [Wood, 2009]. So far, other countries do not require the registration of results.

The registration of clinical trials is now a well-established practice and has become a key tool in addressing some of Evidence-Based Medicine (EBM)’s challenges [Zarin et al., 2007]. However, the current registries contain only text-based or semi-structured information and there is, for example, no common vocabulary for labeling interventions. The amount of protocol information registered is often insufficient to judge the validity of reported results and the problem of identifying all relevant studies has not yet been solved [Zarin et al., 2007]. In addition, the information publicly available may be incomplete or even “largely incomprehensible” [Wood, 2009].

A.2.4 Regulatory assessment

After drug development, the pharmaceutical company compiles the evidence collected from the discovery and development processes into a dossier that is submitted to the regulators who decide upon its market authorization. This is a critical step, which can lead to the disqualification of a compound due to rejection or withdrawal [Eichler et al., 2010]. Both the EMA and the FDA approve approximately 20 – 30 drugs per year [Eichler et al., 2010, Hughes, 2010, Mullard, 2011], and with a 30% failure rate, this amounts to 30 – 45 submissions per year. Submissions to the EMA and the national medicines boards in Europe are mainly text-based, containing aggregate-level results of clinical trials based on the applicant's statistical analysis. Since June 2009 the FDA has additionally required the electronic submission of individual patient data to be able to perform independent analyses [FDA, 2009]. The dossier, especially the clinical trial results, forms the basis on which regulators assess the benefit-risk profile of a new drug. In some cases, the regulatory authorities may give market approval on the condition that additional studies (phase IV trials) are conducted by the company. Such trials are most common in western Europe [Thiers et al., 2008].

In the context of the FDA Critical Path initiative, the FDA and the NCI initiated the JANUS project to build a standards-based clinical data repository specifically meant for the meaningful integration of data [CDISC and FDA, 2005]. It is argued that robust, machine-readable meta-data are required to achieve the full potential of such a repository. The construction of a suitable meta-data model is a monumental task, especially given the modeling of the numerous decisions made during a statistical analysis. Furthermore, this infrastructure would eventually need to be shared between the regulators and the industry, requiring a complex and sustained cooperation effort. However, the JANUS project is increasingly enabled by standards established by the Clinical Data Interchange Standards Consortium (CDISC) and Health Level 7 (HL7). It is built around an open source data model also called JANUS [Food and Drug Administration, 2010a]. The latest released version of the data model is from 2005, but the JANUS project is still ongoing, and was approved as an agency-wide initiative in 2008 [Oliva, 2009]. Eventually, JANUS should offer FDA reviewers easy access to both raw and derived data and facilitate re-analyses [CDISC and FDA, 2005]. The wider standards-based efforts should result in interoperable tools to be used by both the regulators and the industry [CDISC and FDA, 2005].

Since 2004, the EMA has established clinical trial registration in accordance with the EU Directive 2001/20/EC through the EudraCT system. EudraCT was opened to the public only recently, on 22 March 2011 [European Medicines Agency, 2011b], and the records are being released in a staggered fashion. Due to its limited functionality, EudraCT should be considered as a trials registry rather than a database supporting evidence-based regulatory assessment. However, the EMA does publish the European Public Assessment Reports (EPARs) of all centrally approved or refused medicines on its website. Note that this does not include all applications submitted to the EMA, as they can be withdrawn before a decision is reached [Eichler et al., 2010]. The EPAR contains information on all trials, but is completely textual without a semantic structure. Moreover, its information is directly derived from the submission by the applicant, while there is no standardization concerning what information should be provided, or in which format.

A.2.5 Medicinal product information

After the regulatory assessment, the information provided by the Summary of Product Characteristics (SmPC) is made available to professionals and patients via the drug label and package inserts. The SmPC is a text document containing important information on the approved medicinal product, such as recommended dosage, contra-indications, possible interactions with other medicines, and side effects. The information is initially stored in the annex for marketing approval as governed by the regulatory authorities [EMEA, 2005]. In Europe, the SmPC belongs to the EPAR. Although most of the data contained within the SmPC originate from Phase I-III trials, the label might be changed on the basis of new information obtained in the marketing phase. The results of pharmacovigilance processes, in the EU summarized in Periodic Safety Update Reports [EMEA, 1997], or the outcome of Phase IV clinical trials can lead to such changes. This especially applies to drug profiling, which may result in different safety instructions for patient subgroups, such as children. It must be noted that information does not automatically go from clinical trial reports to the SmPC, but the SmPC is the result of a dialogue between the pharmaceutical company and the regulators, that is mainly based on the results obtained in clinical trials.

Both the EMA and the FDA have proposed initiatives for a more structured SmPC. The EMA has introduced the Quality Review of Documents (QRD) and the Product Information Management (PIM) standards. The QRD annotated template [European Medicines Agency, 2010b] provides a loose verbal structure that should be followed by the SmPCs [European Commision, 2009]. PIM is a standard for submitting data in a structure defined by a Document Type Definition [European Medicines Agency, 2010a]. Both QRD and the more advanced PIM are designed for transferring information in a structured format that facilitates translating the product information into the official languages of all EU member states. The FDA has a Structured Product Labeling (SPL) standard similar to PIM [Food and Drug Administration, 2010b] and provides label information in SPL format; one can browse through the labels in a user-friendly format on the DailyMed site (`http://dailymed.nlm.nih.gov/dailymed/`) of the National Library of Medicine. However, QRD, PIM, and SPL do not impose semantics on pharmacokinetic and pharmacodynamic properties, the main quantitative clinical data visible in the SmPC. Some non-profit organizations provide condition-specific drug labeling and/or trial information, for example, the Saskatchewan Lung Association for lung diseases (`http://www.sk.lung.ca/drugs`) and the NCI for different types of cancer (`http://www.cancer.gov/drugdictionary`).

Although the efforts to realize publicly accessible SmPC information seem to have paid off and both the EMA and the FDA have created very good SmPC databases, they are not linked to any clinical trial results databases. Such functionality would be preferable as it would enable one to trace the scientific evidence from a drug on the market back to the original clinical trials. Moreover, the drug compendia merely replicate the SmPC information and have been shown to lack consistency in drug-to-drug interactions due to the insufficient standardization of the terminology used [Vitry, 2006].

A.2.6 Standards and data models

The realization that a common standard for clinical trial data would be beneficial for the semantic interoperability of information systems gave birth to the Clinical Data Interchange Standards Consortium (CDISC) in 1997. Currently, CDISC is a large global non-profit organization with representatives from industry, regulatory authorities, and academia, all dedicated to the development of vendor-neutral, platform-independent and freely available standards. One of the first standards established by CDISC was the Study Data Tabulation Model (SDTM) standard. SDTM is a content standard that describes the core variables and domains to be used when composing a clinical trial dataset to be submitted to the FDA. Other standards are the Trial Design Model (TDM) to represent trials' design and help interpret SDTM data sets, the Protocol Representation Model (PRM) to standardize the content of trial protocols, the Laboratory Data Model (LAB) for the exchange of clinical laboratory data, the Analysis Data Model (ADaM) for an efficient generation, replication, and review of statistical analysis results, the Clinical Data Acquisition Standards Harmonization (CDASH) for identifying a basic dataset of elements that should be captured in a CRF, and the Standards for Exchange of Nonclinical Data (SEND) for data collected from preclinical toxicology studies. Finally, the Operational Data Model (ODM) standard defines the content and structure of CRFs and clinical databases in XML. It facilitates the acquisition, exchange, and archiving of operational data from several sources during the course of a clinical trial. The importance of the CDISC standards partly lies in the fact that the FDA has indicated that clinical trial data should be presented to the agency in the SDTM and ADaM formats. The operational standards appear to be robust and well adopted, and the CDISC is expanding its activities to establishing interoperability with EHR systems, to developing models that support specific therapeutic areas, and to enabling cross-study analysis through the SHARE initiative. Figure A.1 illustrates the role of the CDISC standards in the operational management of clinical trials and in their regulatory submission. The CDISC website (`http://www.cdisc.org/`) offers extensive documentation on the standards.

CDISC has engaged in several collaborations, of which the one with HL7 is of special importance. HL7 has been developing standards for the electronic exchange of medical, financial, and administrative data between health care information systems since 1987. The foundation of HL7 standards development work is the Reference Information Model (RIM), a high level object model of the health care domain. Several standards are derived from the RIM, such as V3 Messages for the meaningful interchange of data between health care systems, GELLO for rule-based decision support, and the Clinical Document Architecture for semantically structured documents. CDISC has adopted the HL7 V3 Messaging standard for the exchange of clinical trial data. For the FDA, this standard will replace the antiquated SAS transport file submission format. By adopting the HL7 messaging standard, CDISC ensures that both the clinical trial data and their electronic exchange are standardized.

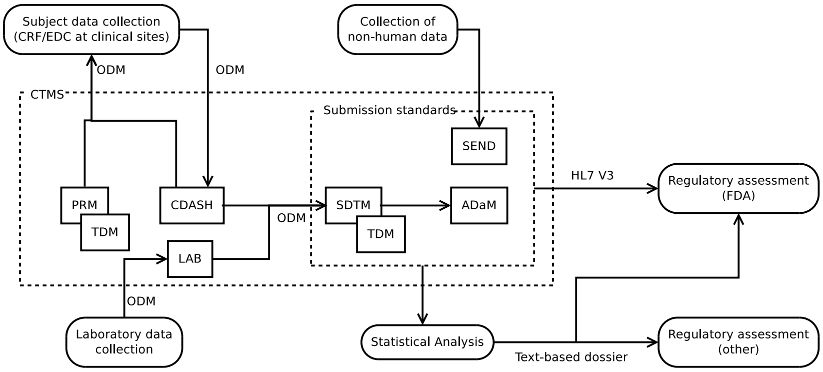

Subject data collection
(CRF/EDC at clinical sites)
Collection of
non-human data
ODM
ODM
CTMS
Submission standards
SEND
HL7 V3
Regulatory assessment
(FDA)
PRM
TDM
CDASH
ODM
SDTM
ADaM
TDM
LAB
ODM
Laboratory data
collection
Statistical Analysis
Text-based dossier
Regulatory assessment
(other)

Figure A.1: Standards in the operational management of clinical trials and their regulatory submission. Abbreviations: Analysis Data Model (ADaM), Clinical Data Acquisition Standards Harmonization (CDASH), Clinical Trial Management System (CTMS), Health Level 7 (HL7), Laboratory Data Model (LAB), Operational Data Model (ODM), Protocol Representation Model (PRM), Trial Design Model (TDM), Standards for Exchange of Nonclinical Data (SEND).

Moreover, through this collaboration, CDISC and other participating parties hope to align the current CDISC standards for clinical trial data with the HL7 RIM standards for healthcare. To achieve this, an overarching domain analysis model is being developed, called the Biomedical Research Integrated Domain Group (BRIDG) model, which is intended to bridge the gap between clinical research and healthcare [Fridsma et al., 2008]. Bridging this gap would offer new possibilities for the development of true translational medicine, which means that data obtained in healthcare could be more easily used in clinical research and vice versa.

The BRIDG project is a collaboration between the CDISC, HL7, the NCI and the FDA that aims at bringing together the common elements of their various standards to a shared view of semantics of the domain of protocol-driven research and its associated regulatory artifacts [Biomedical Research Integrated Domain Group (BRIDG), 2010]. The model is intended to be implementation independent in the sense that it models the problem domain, and not any specific solution. For example, unlike some other CDISC standards it does not specify the format in which to submit data to the FDA. BRIDG relies on external vocabularies and ontologies, but the specific terminology used is up to the implementer. Due to the increasing complexity of the BRIDG model, several sub-domain views are now delivered as part of the model. These are the protocol representation, study conduct, adverse event and regulatory perspectives. While the operational aspects of clinical trials are well covered by these perspectives, a data analysis perspective is currently missing as there is no adequate standard for statistical analysis. An 'ontological perspective' is planned in the form of a Web Ontology Language (OWL) representation of the BRIDG model, which would enable more formal validation, for example against the RIM. Sophisticated, highly formalized ontologies can be used in advanced applications, such as computer-aided reasoning [Rubin et al., 2008], and thus might enable even broader use of the BRIDG model.

A.2.7 Controlled terminologies

Controlled terminologies (synonymously: controlled vocabularies, coding systems) of clinical terms are an important first step in the application of information technology to medicine [Cimino, 1996]. Controlled terminologies predate information technology, e.g. the International Classification of Diseases (ICD) was already introduced in 1893. The ICD formally codes diseases and enables (for example) the assessment of disease incidence from medical records. Other terminologies fill other niches, for example the Medical Subject Headings (MeSH) [Nelson et al., 2004] is used to index the medical literature (e.g. PubMed meta data is coded in MeSH), and the Medical Dictionary for Regulatory Activities (MedDRA) is used for coding safety data (e.g. adverse events). Many of these specialized terminologies are organized into a strict hierarchy, which means that some specific terms may fit in multiple places [Cimino, 1996]. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) terminology is an important attempt to create a clinical terminology with comprehensive coverage [Schulz et al., 2006]. It currently contains around 311,000 concepts and 800,000 terms [The International Health Terminology Standards Development Organisation, 2011]. It also goes beyond a simple hierarchical structure and provides the logical relationships that hold between terms; over 1.3 million such relationships are currently modeled [Schulz et al., 2006, The International Health Terminology Standards Development Organisation, 2011]. Due to this complex logical structure, SNOMED CT could itself be viewed as an ontology.

The Unified Medical Language System (UMLS) [Lindberg et al., 1993, Bodenreider, 2004] is essentially a collection of over 60 biomedical terminologies and a coding of the relationships between them through the 'Metathesaurus'. The ICD, SNOMED CT and MeSH are among the terminologies integrated by the UMLS. Like SNOMED CT, concepts in the UMLS are linked through a complex system of relationships. Some of these relationships originate directly from the source terminologies, while others are generated specifically for the Metathesaurus [Bodenreider, 2004]. However, the mappings between terminologies in the UMLS are far from complete [Nadkarni and Darer, 2010] and mapping between terminologies, especially to SNOMED CT, is an active area of research, e.g. [Nadkarni and Darer, 2010, Vikstrom et al., 2007].

Thus, there are many controlled terminologies for medicine (in fact, most were not mentioned), but unfortunately there is as yet no standardization of which ones should be used, and mapping between them is an open problem. For example, in clinical research MedDRA is used to code Adverse Drug Events (ADEs), while the healthcare area prefers the SNOMED CT dictionary. This hinders the interoperability of the various information systems being used.

A.3 Discussion

Over the last decades, several standardization bodies (notably CDISC and HL7) and CTMS vendors have put great effort into automating the information-intensive aspects of drug development. In this area, the focus is shifting from core data management to electronic sourcing, such as linking to the EHR, and to using increasingly advanced and standardized information flows. However, these systems and standards are still largely oriented toward the operation of single studies, while the issue of storing the data of multiple studies in a structured and meaningful way remains largely unsolved. Although progress has been made, there are no known large, successful, and publicly available data warehouses, nor any standards that would enable cross-study analyses of aggregate level results.

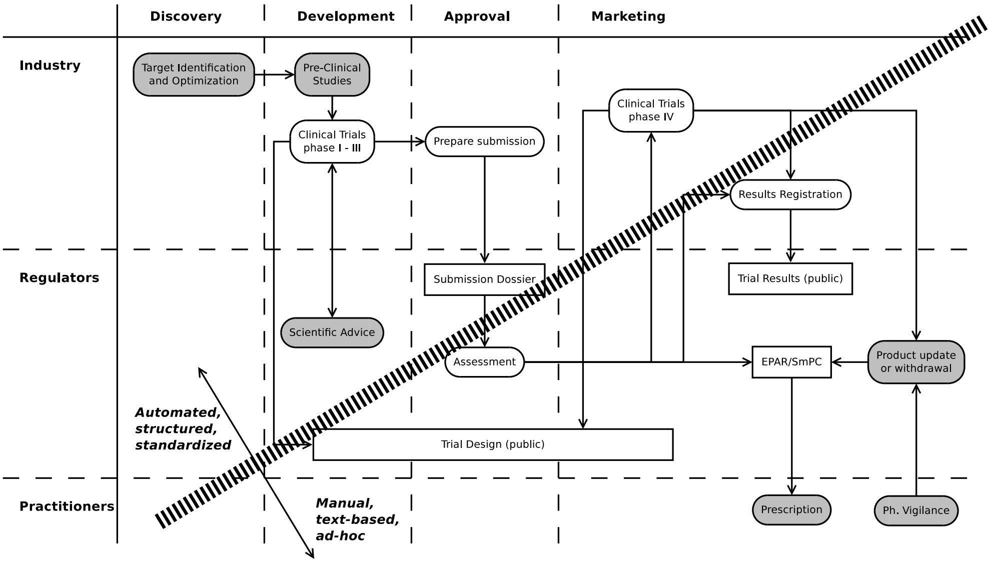

Discovery
Development
Approval
Marketing
Industry
Regulators
Practitioners
Target Identification and Optimization
Pre-Clinical Studies
Clinical Trials phase I - III
Prepare submission
Clinical Trials phase IV
Results Registration
Submission Dossier
Trial Results (public)
Scientific Advice
Assessment
EPAR/SmPC
Product update or withdrawal
Automated, structured, standardized
Trial Design (public)
Manual, text-based, ad-hoc
Prescription
Ph. Vigilance

Figure A.2: A process view on drug development and regulation. The boxes with round corners represent processes, and those with straight corners information products. The arrows indicate the transfer of evidence. The gray boxes are not discussed in detail, but complete the picture of drug development and regulation. Abbreviations: European Public Assessment Report (EPAR), Summary of Product Characteristics (SmPC).

A process view on drug development and regulation (Figure A.2) shows that in spite of the largely successful efforts to create an electronic infrastructure for information management during the execution and regulatory submission of clinical trials (top-left of the diagonal), both the dissemination and integration of the resulting evidence to the public and the scientific community remain inefficient and ad-hoc processes (bottom-right). The flow of information from the CTMS to the FDA in the

US is standards-based and largely automated, but in other countries it is usually a text-based transfer of aggregated results that does not support the independent verification or re-analysis of the submitted data.

Although the operational systems that help manage individual studies during the development phase are mature and standardized, the subsequent transfer of evidence to scientific journals, public registries, regulators, and (eventually) clinical practice is a largely ad hoc and text-based affair. Consequently, a lot of effort is put into making the results of trials public, but the current systems do not facilitate optimal use of that information. Whereas the current system is centered on single studies, most decision makers need a system that enables the integration of evidence across studies. We believe that this issue is one of the root causes of the lack of transparency in the processes of drug development and regulation.

Acknowledgements

We would like to thank Morris Swertz of the Groningen Bioinformatics Centre and Marcel Hekking and Hans van Leeuwen of Merck, Sharp & Dohme for their insightful comments.

APPENDIX B

Software deliverables

B.1 ADDIS

Full name Aggregate Data Drug Information System

Authors Gert van Valkenhoef, Tommi Tervonen, Tijs Zwinkels, Maarten Jacobs, Hanno Koeslag, Daniel Reid, Florin Schimbinschi, Ahmad Kamal, Joël Kuiper, Wouter Reckman

Available from `http://drugis.org/addis`

Source code `http://github.com/gertvv/addis`

Releases 0.2 (2009-06-30), 0.4, 0.4.1, 0.6, 0.6.1, 0.8, 0.8.1, 0.10, 1.0, 1.2, 1.2.1, 1.4, 1.6, 1.6.1, 1.6.2, 1.8, 1.10, 1.10.1, 1.12, 1.12.1, 1.12.2, 1.12.3, 1.12.4, 1.14, 1.14.1, 1.16 (2012-10-04)

Size 69,034 lines of code (Java, XSD, XSL), 3,088 commits (2012-10-04)

Aggregate Data Drug Information System (ADDIS) provides decision support for evidence-based benefit-risk decision making, and combines a structured database of aggregate clinical trial results with capabilities for automated network meta-analysis and multi-criteria decision modeling.

B.2 GeMTC

Full name Generate Mixed Treatment Comparisons

Authors Gert van Valkenhoef

Available from `http://drugis.org/gemtc`

Source code `http://github.com/gertvv/mtc`

Releases 0.2 (2010-05-07), 0.4, 0.6, 0.8, 0.8.1, 0.10, 0.10.1, 0.12, 0.12.1, 0.12.2, 0.12.3, 0.12.4, 0.14 (2012-10-01)

Size 22,184 lines of code (Java, XSD, R), 670 commits (2012-10-01)

GeMTC is a reusable software library (used by ADDIS), as well as a stand-alone application and an R package for network meta-analysis (mixed treatment comparison) model generation. Given a data set giving the results data for a single outcome over a network of clinical trials, the GeMTC software can generate analysis models that can be run in the BUGS or JAGS software. It also incorporates the YADAS library to run the analysis models directly, without the need to run BUGS or JAGS.

B.3 drugis.org common

Full name drugis.org common library

Authors Gert van Valkenhoef, Tommi Tervonen,
Tijs Zwinkels, Maarten Jacobs, Hanno Koeslag,
Daniel Reid, Florin Schimbinschi, Ahmad Kamal,
Joël Kuiper, Wouter Reckman

Available from –

Source code `http://github.com/gertvv/drugis-common`

Releases 0.1 (2010-09-16), 0.2, 0.4, 0.4.1, 0.4.3, 0.4.4, 0.4.5, 0.4.6, 0.4.7, 0.4.8, 0.4.9, 0.5, 0.5.1, 0.5.2, 0.5.3, 0.5.4 (2012-10-01)

Size 9,227 lines of code (Java), 223 commits (2012-10-01)

A shared library used by the ADDIS, GeMTC, and JSMAA applications. It implements some common GUI components, presentation and beans functionality and a threading model for running tasks with parallel sub-tasks, such as Markov Chain Monte Carlo (MCMC) simulations.

B.4 hitandrun

Full name "Hit and Run" for sampling uniformly from convex shapes

Authors Gert van Valkenhoef

Available from `http://cran.r-project.org/web/packages/hitandrun/`

Source code `http://github.com/gertvv/hitandrun`

Releases 0.2 (2011-01-10)

Size 1,481 lines of code (C, R), 44 commits (2012-02-16)

This R package implements the "Hit and Run" method for sampling from convex shapes defined by linear constraints. It also provides utilities to enable the easy generation of weights that can be used in simulation based multi-criteria decision analysis, especially Stochastic Multicriteria Acceptability Analysis (SMAA).

B.5 odcread

Full name Read ".odc" Oberon Compound Documents

Authors Gert van Valkenhoef

Available from –

Source code `http://github.com/gertvv/odcread`

Releases –

Size 2,106 lines of code (C++), 52 commits (2012-02-16)

Most published code for network meta-analysis is provided as WinBUGS models. These models are distributed in the binary ".odc" format, which is not easy to access on operating systems other than Microsoft Windows (e.g. Mac OS X or GNU/Linux). And, while the code is mostly compatible with JAGS, the binary distribution format prevents doing this easily. This program enables converting ".odc" files to plain text.

B.6 jags-jni

Full name Java Native Interface to JAGS

Authors Gert van Valkenhoef

Available from –

Source code `http://github.com/gertvv/jags-jni`

Releases –

Size 1,111 lines of code (C++, Java), 14 commits (2012-02-16)

ADDIS is implemented in Java, and uses the YADAS implementation of Markov chain Monte Carlo analysis. JAGS is a more generally accepted and advanced implementation of Markov chain Monte Carlo methods, but is implemented in C++. This library provides a bridge between Java and JAGS using the Java Native Interface (JNI), to enable the use of JAGS in ADDIS and other Java applications.

APPENDIX C

Product and Release Planning Practices for Extreme Programming

G. van Valkenhoef, T. Tervonen, E. O. de Brock, and D. Postmus. Product and release planning practices for extreme programming. In *Proceedings of the 11th International Conference on Agile Software Development (XP2010)*, Trondheim, Norway, 2010. doi: 10.1007/978-3-642-13054-0_25

Abstract

Extreme Programming (XP) is an agile software development methodology defined through a set of practices and values. Although the value of XP is well-established through various real-life case studies, it lacks practices for project management. In order to enable XP for larger projects, we provide the rolling forecast practice to support product planning, and an optimization model to assist in release planning. We briefly evaluate the new practices with a real-life case study.

C.1 Introduction

Extreme Programming (XP) is one of the most "agile" software development methodologies. Unlike plan-driven methodologies (e.g. waterfall) that define software development as a process, XP defines it through values and practices proven to work well together in real-life software development [Beck, 1999, 2005]. A good project management process and strong customer involvement are critical to project success in XP [Chow and Cao, 2008]. Although XP provides a consistent set of practices, it almost completely lacks practices for planning [Abrahamsson et al., 2003]. Therefore, although XP has been reported to be tailorable for large-scale projects [Cao et al., 2004], it is generally considered more suitable for small projects. Moreover, the 'on-site customer' practice [Beck, 1999] is often hard to implement due to organizational or time constraints [Rumpe and Schröder, 2002]. The XP customer is consistently under significantly more pressure than the developers or other participants in the project [Martin et al., 2004]. This causes the following problems (which become worse as projects get larger):

1. Lack of management context: XP does not address the larger context in which release planning takes place, or the long term project goals [Abrahamsson et al., 2003]. This means that the customer or the developers may loose track of the overall purpose of the system and consequently make sub-optimal planning decisions.

2. User story overload: the number of user stories to be considered in release planning can make the planning process too demanding for the customer.

3. Prioritization stress: the responsibility of prioritizing user stories may cause stress for the customer, even for a small number of stories. It is difficult to foresee the consequences and adequacy of the prioritization [Martin et al., 2004], and it is unclear whether the customer perceives business value in constantly managing the development priorities [Grisham and Perry, 2005].

To address these problems, this paper proposes two new planning practices for XP. First, we assist in product planning with the new practice of rolling forecasts (Section C.2). This practice helps to provide management context often lacking in XP (Problem 1 above). Second, we introduce an automated planning aid that can be used during release planning to reduce the customer workload by generating a suggested plan that satisfies simultaneously the constraints imposed by the customer and the limited development resources (Section C.3). This addresses issues 2 and 3 identified above. After introducing the practices, we demonstrate their use in a real-life study (Section C.4), before giving concluding remarks (Section C.5).

C.2 Rolling Forecast for Product Planning

Expectation management is often the key difference between failed and succesful software projects [Boehm and Turner, 2003]. XP originally proposes the 'system metaphor' practice for expectation management [Beck, 1999]. However, in practice, 'system metaphor' is difficult to apply and not useful, and is therefore often not implemented [Rumpe and Schröder, 2002]. The 'system metaphor' has since been removed from XP [Beck, 2005], and there is no replacement practice addressing expectation management. The lack of an expectation management practice that is coherent with the rest of the methodology can cause additional project risks, especially if the customer is not constantly available on-site, as is often the case (see [Rumpe and Schröder, 2002]).

Product planning should provide the context in which the release planning takes place [Cohn, 2005]. In each release, before stories are elicited, the customers should have a rough idea of the current state of the system and the direction of development. This is promoted in XP by having the customer test and accept implemented stories and by frequently giving system demonstrations. However, it is unclear how a shared vision of the direction of future development can be established, especially when the customer does not clearly know what (s)he wants. As a consequence, an upfront rigid planning of the whole product in concrete terms is often almost impossible and can also become counter productive ('analysis paralysis').

To support product planning, we introduce the practice *rolling forecast*. At project inception, an overview of the *product goals* is drawn up by the customer together with the project manager. The goals should be stated in a functional format but in such a way that they cannot readily be broken into themes without further analysis and elicitation. The goals serve to provide a shared vision of the system and to form a basis for user story elicitation, but they are not requirements *per se*. It is advisable to re-evaluate the overall goals periodically, e.g. after every fourth release.

After defining the product goals, a *theme forecast* is created by the customer, project manager and a development team representative (e.g., an analyst or technical manager). A theme forecast consists of a set of themes, their likely implementation order, and a prediction of which themes will be realized in the coming two or three releases. The theme forecast can be adapted in preparation of every release planning (before story elicitation). Thus, a *rolling forecast* manages the expectations about the software by iteratively developing theme forecasts based on overall product goals. Then, in release planning, the theme forecast is taken into account when deciding on the themes and stories for the next release, while iteration planning takes into account (and adjusts) the release plan in choosing the stories and identifying the tasks for the next iteration, i.e., the normal agile planning practices are applicable at the release and iteration levels [Cohn, 2005, Beck and Fowler, 2001].

C.3 Supporting Release Planning Model

Our planning model is aimed to support release planning. The developers elicit stories from the customer and ask him/her to evaluate them with respect to their business value on an interval scale, e.g. 1–5. Then the developers evaluate the stories' implementation complexity in story points. The model provides a planning aid by maximizing the implemented business value, taking into account constraints on im-

Model 1 The optimization model as a side-constrained knapsack problem.

$$
\begin{array}{lllll}
1. & \text{max} & b_1x_1 + \ldots + b_{n+m}x_{n+m} & & \\
2. & \text{s.t.} & c_1x_1 + \ldots + c_{n+m}x_{n+m} & \leq \; p & \\
3. & & x_j - x_i & \leq \; 0 & \text{for all } i, j \text{ where } x_i \succ x_j \\
4. & & \sum_{i=1}^{n} a_{ij}x_i - s_jx_{n+j} & \geq \; 0 & \text{for } j = n+1, \ldots, n+m \\
5. & & \sum_{i=1}^{n} a_{ij}x_i - x_{n+j} & \leq \; s_j - 1 & \text{for } j = n+1, \ldots, n+m \\
6. & & x_1, \ldots, x_{n+m} \in \{0, 1\}. & &
\end{array}
$$

plementation complexity and precedence relations. A precedence relation is interpreted as a story not having value unless another (preceding) story is implemented. Moreover, in XP, related stories are often grouped into themes that represent larger pieces of related user functionality, and synergy effects occur when all stories within a theme are implemented [Beck, 2005]. We model such effects by awarding extra value to a theme of stories if they are implemented together in a single release. Note that not all stories need to belong to a theme, and that one story can belong to more than one theme. We don't allow themes to span multiple releases in order to prevent the supporting planning model being used for making longer term plans, that might lower the overall agility of the XP development process. Longer-term product goals should instead be handled with the other proposed practice, rolling forecast.

Our model assumes adherence to the standard best practices regarding story and theme sizes. Stories should be small enough that they can easily be implemented in a single iteration, and themes in a single release. Moreover, a theme should consist of the *minimal set of stories* required to achieve the aforementioned synergy effect. Not adhering to these guidelines may lead to inappropriate results from the model.

The story selection can be formulated as a knapsack problem (the complete integer programming formulation is given in Model 1). Let us denote by n the number of uncompleted stories. Each story i has a business value of b_i and implementation complexity of c_i story points. The total amount of story points that can be implemented during a release is denoted by p. The decision problem is to select the most valuable subset of stories to implement in a release (Model 1: 1), subject to a budget constraint on the maximum implementation complexity (Model 1: 2). For each story $i \in \{1, \ldots, n\}$, let $x_i = 1$ if story i is selected and $x_i = 0$ otherwise (Model 1: 6). Precedence of story i to story j is denoted by $x_i \succ x_j$ and can be incorporated into the optimization model by adding the following constraint: $x_j - x_i \leq 0$ (Model 1: 3).

To model themes, let m be the number of themes and let s_j ($j \in \{1, \ldots, m\}$) be the number of stories within theme j. Theme j can be included in the model by introducing a dummy story $(n + j)$, such that $x_{n+j} = 1$ if and only if all stories within theme j are implemented (Model 1: 4–5). The business value b_{n+j} associated with story $(n + j)$ represents the additional value that is awarded when all stories within theme j are implemented; its implementation complexity c_{n+j} is set equal to zero.

We implemented the supporting release planning model using R[1] and lp_solve[2].

---

[1] `http://www.r-project.org`

[2] `http://lpsolve.sourceforge.net`

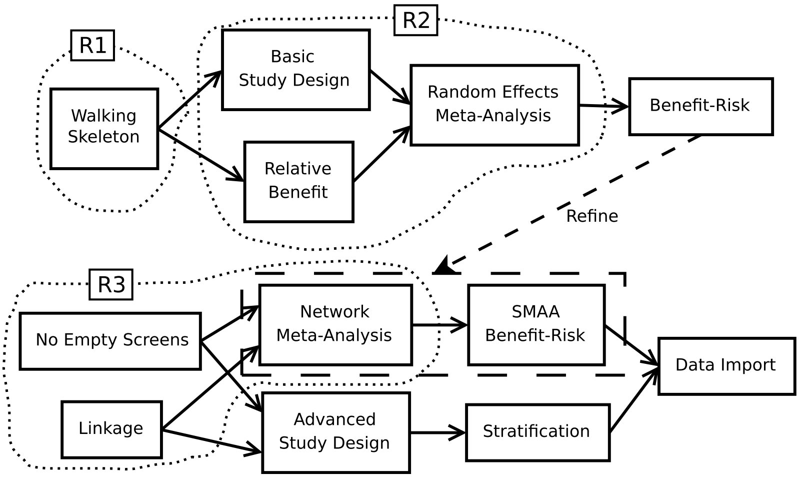

R1
Walking Skeleton
R2
Basic Study Design
Relative Benefit
Random Effects Meta-Analysis
Benefit-Risk
Refine
R3
No Empty Screens
Linkage
Network Meta-Analysis
Advanced Study Design
SMAA Benefit-Risk
Stratification
Data Import

Figure C.1: The theme forecast for the beginning of the project (top) and the updated forecast (bottom) before release 3 (R3). A solid arrow from A to B indicates A has priority over B. The actually implemented themes from release 1 (R1) and 2 (R2) are shown, as well as the expected themes for release 3 (dotted lines). The dashed arrow indicates a high-level theme being refined as more information became available.

Our implementation is freely available online[3].

C.4 Real-Life Example

We are involved in a research project with external customers that expect us to develop software artifacts for the application domain of pharmacological decision support. Our development environment consists of 2 teams working part-time. In the following, we detail how we used the rolling forecast practice and our planning model in the development of ADDIS[4].

Rolling Forecast. Although we didn't have clear requirements, we couldn't wait until the research results were present. In order to generate an overall view on the project and the first theme forecast, we interviewed the external customers of the research project. The initial forecast (Figure C.1, top) was constructed considering 16 goals, such as "the system should provide drug efficacy and safety information". The theme forecast consists of a detailed set of themes for the next release(s) and a more global set of (likely) themes for the more distant releases. The forecast helped us to elicit stories in release planning meetings with the main external customer, who also chose the stories to implement. Figure C.1 shows how our mutual understanding of the project evolved during the first half-year of development (only the most important themes are shown). We initially decided to focus on 'benefit risk' as a long-term goal. This defined our priorities for the first two releases. We knew that to actually implement 'benefit-risk', research input would be needed. As these results became available only during the second release, the forecast was refined. Simultaneously, we were able to identify additional themes that also support our long-term goals, as well as two themes ('no empty screens' and 'linkage') that generate interest for our software through usability.

---

[3] `http://github.com/gertvv/xpplan`

[4] `http://drugis.org`

Planning Model. We did the first release (ADDIS 0.2) as a burn-in for velocity estimation and to create an initial end-to-end working system[5]. Therefore we didn't estimate story business values while planning the first release. During the second release (ADDIS 0.4), we estimated story business values (scale: 1-5), story complexities (scale: 1,2,3,5,8), technical precedence relations (none were identified) and themes. In this release, we could identify 3 themes as being the most important. After the release was completed, we ran our supporting optimization model for release planning. We tested the sensitivity of the optimization model and differences between the model's solution and the stories we actually implemented by varying the theme value from 0 to 99. The results didn't differ much from our manually planned implementation order and the model showed to be robust with respect to changes in theme value: the only differences emerged when the theme value changed from 0 to 1 and from 10 to 11. When theme values varied between $1 - 10$ the same two out of three themes were included in the optimal solution whereas with theme value > 10 all three themes were included.

C.5 Conclusions

Lack of management context, user story overload, and prioritization stress cause high workload for the customer and hinder scalability of XP to larger projects. To overcome these limitations, we propose two new practices: rolling forecast for product planning and release planning support through an optimization model. We evaluated the applicability of our new practices in a software development project and found them useful. However, we do not have sufficient evidence to make claims about their suitability for projects with different customer profiles, numbers of developers, or levels of developer competency. Our ongoing development project cannot address these questions, and additional appropriate empirical studies should be initiated. Our future research will investigate how business value should be estimated for themes, and how uncertainty can be made explicit in the planning process.

---

[5] Also known as a 'walking skeleton', `http://alistair.cockburn.us/Walking+skeleton`

APPENDIX D

Quantitative release planning in Extreme Programming

G. van Valkenhoef, T. Tervonen, B. de Brock, and D. Postmus. Quantitative release planning in extreme programming. *Information and Software Technology*, 53(11):1227–1235, 2011. doi: 10.1016/j.infsof.2011.05.007

Abstract

Context: Extreme Programming (XP) is one of the most popular agile software development methodologies. XP is defined as a consistent set of values and practices designed to work well together, but lacks practices for project management and especially for supporting the customer role. The customer representative is constantly under pressure and may experience difficulties in foreseeing the adequacy of a release plan.
Objective: To assist release planning in XP by structuring the planning problem and providing an optimization model that suggests a suitable release plan.
Method: We develop an optimization model that generates a release plan taking into account story size, business value, possible precedence relations, themes, and uncertainty in velocity prediction. The running-time feasibility is established through computational tests. In addition, we provide a practical heuristic approach to velocity estimation.
Results: Computational tests show that problems with up to 6 themes and 50 stories can be solved exactly. An example provides insight into uncertainties affecting velocity, and indicates that the model can be applied in practice.
Conclusion: An optimization model can be used in practice to enable the customer representative to take more informed decisions faster. This can help adopting XP in projects where plan-driven approaches have traditionally been used.

D.1 Introduction

Extreme Programming (XP) is an agile software development methodology. XP defines software development through values and practices thought to work well together in practice [Beck, 1999, 2005]. In XP, good project management and strong customer involvement are critical for project success [Chow and Cao, 2008]. Yet, XP provides very little project management support [Abrahamsson et al., 2003] and the XP customer is consistently under significantly more pressure than the developers or other participants in the project [Martin et al., 2004]. Release planning in particular has been characterized as a difficult problem in which many variables have to be considered and judgments are primarily relative [Carlshamre, 2002], leading to user story overload: it is unpractical to consider everything, even with a moderate number of stories. In addition, the customer may suffer from prioritization stress: it is difficult to foresee the consequences and adequacy of prioritization [Martin et al., 2004], and it is unclear if the customer perceives business value in constantly managing the development priorities [Grisham and Perry, 2005]. Therefore, tool support for exploring the solution space and for generating potential solutions is desired [Carlshamre, 2002]. In the distinction between art and science in release planning [Ruhe and Saliu, 2005], planning in XP is traditionally all art [Beck and Fowler, 2001]. Easy to use low effort planning software could enable a hybrid approach, but would need to be tailored for the XP practices and values [Beck, 2005].

Quantitative models for supporting software release planning have been proposed previously, but a recent systematic review concludes that there is a lack of diversity among the existing models [Svahnberg et al., 2010]. Quantitative approaches for requirements prioritization based on value and effort in a plan-driven context were developed in [Karlsson and Ryan, 1997, Carlshamre, 2002], and a general mathematical formulation for incremental development was proposed in [Ruhe and Saliu, 2005]. More fine-grained models that take into account resource constraints due to developers with varying capabilities have been proposed for iterative development [Greer and Ruhe, 2004, van den Akker et al., 2008, Ngo-The and Ruhe, 2009]. Several approaches handle uncertainty in the parameters by generating multiple 'good' plans [Greer and Ruhe, 2004, Ruhe and Greer, 2003, Ruhe and Saliu, 2005, Saliu and Ruhe, 2007, Ngo-The and Ruhe, 2008] rather than a single optimal one. Another model [Li et al., 2010] can be used together with Scrum to deal with change, but it does not explicitly consider uncertainty during planning.

If a high degree of requirements change is expected, agile methods are better suited than plan-driven ones and consequently it may be necessary to adopt agile methods outside of their home ground [Boehm and Turner, 2003]. To enable this, our previous work introduced additional product and release planning practices for XP [van Valkenhoef et al., 2010]. One of the introduced practices uses an optimization model based on evaluating user stories not only on their sizes (i.e. implementation effort), but also based on business value (on an interval scale), possible precedence relations (i.e., story x needs to be completed before story y), and themes. In contrast to the models described previously, our model is tailored to XP as it embraces change by generating only a high-level global plan in terms of user stories, leaving room for

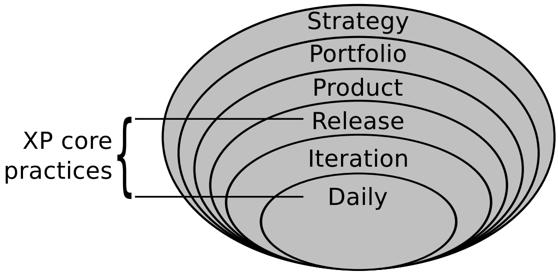


Figure D.1: A Planning Onion [Cohn, 2005] showing the levels of granularity in XP planning. XP core practices consider daily, iteration, and release planning.

agility in how these are realized, and which stories are left out when things do not go according to the plan. This paper extends the model proposed in [van Valkenhoef et al., 2010] to take into account uncertain project velocity.

There are few models that plan for uncertainty explicitly. In [Al-Emran et al., 2010], a simulation-based approach is introduced for evaluating the impact of uncertainty on the execution of release plans, but it does not enable generating a plan. Only one of the models in the literature assumes a probability distribution for the effort estimates and can generate plans with pre-specified completion probabilities [Ruhe and Greer, 2003]. However, reality can (and will) deviate from the plan depending on many other factors, and Bayesian models can aggregate these uncertainties in the probability density of the velocity [Hearty et al., 2009]. In our model parameter uncertainty is accounted for by assuming a probability distribution for the velocity.

The remainder of this paper is structured as follows. We start with a general overview of the model in Section D.2. Then, in Section D.3, we give the integer programming formulation of our quantitative planning model. Section D.4 addresses the problem of theme value elicitation and Section D.5 introduces a simple method to estimate a probability distribution for the velocity. We evaluate the velocity estimation heuristic with an example in Section D.6. Section D.7 evaluates the running time of the model. Finally, we conclude and give directions for future work in Section D.8.

D.2 Release planning model

We present a model to assist the XP development team, especially the customer, in the release planning process. First we briefly explain the context of planning in XP and clarify the used terminology. Then we provide a global overview of the model structure including the required inputs and produced outputs. Finally, we discuss how the model fits within the XP development process and the potential caveats in its application.

D.2.1 Planning in XP

The central idea of planning in XP is to plan features to implement rather than the development tasks necessary to implement these features. Planning features, represented by user stories, enables measuring progress through verifiable functionality. When the number of stories to implement is large, functionally related stories can be grouped into themes that form a consistent set of desirable features [Cohn, 2005]. This enables planning to take place at a higher level of abstraction before considering the merits of individual stories. To deal with uncertainty, the planning process occurs iteratively at release, iteration, and daily levels (see Figure D.1). In *release planning*, the user stories (themes) to be developed for the next release are chosen. The developers coarsely estimate (in story points) the implementation effort required for each story and the customer prioritizes the stories. Together they agree on a high level plan consisting of critical stories that are likely to be implemented during the release, and non-critical stories that may or may not be implemented. The release cycle is fixed at 3–6 months [Cohn, 2005] and is made up of 1–4 week iterations [Beck, 1999, 2005, Beck and Fowler, 2001, Cohn, 2005] that each result in a working system. In *iteration planning*, the development team breaks stories with high priority down into tasks and estimates the effort required to implement these tasks in order to decide which stories can be implemented during the iteration. This also results in an updated release plan. Work division and task scheduling takes place only at the *daily planning* level.

The amount of implementation effort (story points) available during a release depends on the length of the release cycle, and the *velocity* of the development team. The velocity is the number of story points that can be implemented during some unit of time (e.g. an iteration). Traditionally the velocity is predicted through simple methods [Beck and Fowler, 2001], and uncertainty assessed using rules of thumb [Cohn, 2005]. However, there also exists a dynamic Bayesian network model for project velocity monitoring and prediction in XP projects that explicitly quantifies uncertainty [Hearty et al., 2009].

D.2.2 Model overview

Our model aims at supporting release planning by generating a release plan that maximizes the implemented business value taking into account capacity constraints, precedence relations, themes, and uncertainty in the velocity. Applying the model results in a suggested release plan, which consists of sets of stories ordered according to decreasing completion probability. During release planning, the team may decide to accept the suggested plan as-is, or to amend it in any way they see fit. The release plan serves as the main input for iteration planning.

Model inputs

Our model requires explicit elicitation of a number of parameters (responsible team member in parentheses):

- story and theme values (customer)

- story sizes (developers)
- preference precedences (customer)
- technical precedences (developers)
- a velocity prediction (tracker)

Story size and value elicitation should be done by analogy, so that a story's size and value is defined relative to the size and value of other (past and present) stories. We propose using a 1-5 scale for story values as most customers are already familiar with it due to its similarity with the Likert scale that is widely applied in questionnaires. If large differences in story values (i.e. more than a factor 5) exist, a wider scale needs to be used. In some projects a metric of predicted monetary value [Hartmann and Dymond, 2006] may be more suitable. Synergy effects should occur when all stories within a theme are implemented. Similar to how [van den Akker et al., 2008] implemented revenue-based dependencies, we model such effects by awarding extra value to a theme of stories when all stories are implemented. How this extra value can be specified is discussed in Section D.4. Story size is estimated in story points, for which e.g. the Cohn scale [Cohn, 2005] can be used.

Two types of precedence relations exist: technical and preference. Technical precedence means that a story cannot technically be implemented before another one. Although stories in XP should be as independent as possible [Beck and Fowler, 2001], sometimes dependencies are unavoidable [Carlshamre, 2002]. Preference precedence allows the client to express preferences for stories, and is interpreted as a story not having value unless another (preceding) story is implemented.

Imprecision in the estimated story sizes, variability in programmer productivity, and uncertainty in other factors mean that release velocity is uncertain. This is accounted for in our model by assuming a probability distribution $f(v)$ for the total number of story points that can be implemented during the release. Our method can be used with any method that estimates $f(v)$. The construction of $f(v)$ conditional on knowledge about the underlying factors is discussed further in [Hearty et al., 2009]. We propose a simple heuristic estimation procedure for situations where this approach is too demanding (Section D.5).

Model outputs

Due to uncertainty in velocity, the optimal planning decision is a stochastic problem. In order to provide simple rules for release planning in practice, our model gives (disjoint) sets of stories with decreasing completion probabilities. Let $p_k \in (0, 1)$ denote the completion probability of story set ℓ_k. Then, the story sets are ordered in such a way that

$$p_i > p_j \ ; \ \forall_{i<j}$$

Normally it is sufficient to distinguish three story sets corresponding to the Dynamic Systems Development Method MoSCoW rules [Cohn, 2005, Stapleton, 1997]: 'must have' (ℓ_1), 'should have' (ℓ_2), and 'could have' (ℓ_3). For example, we could set $p_1 = 0.9$, $p_2 = 0.7$, and $p_3 = 0.3$ as the desired completion probabilities for (respectively)

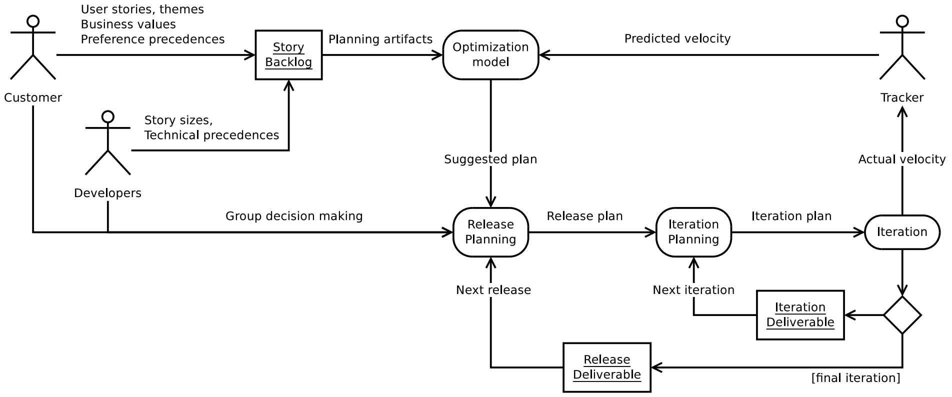

User stories, themes
Business values
Preference precedences
Story Backlog
Planning artifacts
Optimization model
Predicted velocity
Customer
Tracker
Story sizes,
Technical precedences
Developers
Suggested plan
Actual velocity
Group decision making
Release Planning
Release plan
Iteration Planning
Iteration plan
Iteration
Next release
Next iteration
Iteration Deliverable
Release Deliverable
[final iteration]

Figure D.2: The role of our proposed optimization model in the XP planning process. Release and iteration planning are shown and the other levels of planning are omitted.

the 'must have', 'should have', and 'could have' sets. Optionally a 'won't have' set can be used to store the rest of the story backlog. A similar approach is taken in [Miranda, 2002], where it is proposed to plan time-bound projects in several increments with decreasing completion probabilities. However, in our model, the story sets do not correspond directly to iterations as the scope of an iteration is decided in iteration planning.

D.2.3 Discussion

The model we propose fits well in the XP development process, as it assists in release planning while leaving the other XP practices intact. Moreover, we minimize the data elements that have to be specified. Compared to standard XP planning, we additionally elicit the value of stories and themes, which is already implicit in the current process because the customer prioritizes the stories. For velocity we use a probability distribution rather than just a point estimate, but the required calculations are based on historical data that is already available (see Section D.5). The role of the model within the XP planning cycle is illustrated in Figure D.2.

The model is not intended for planning multiple releases, as doing so might lower the overall agility of the XP development process. However, in some situations it may be crucial for the project success to plan multiple releases ahead, for example due to marketing reasons. Such long term plans can be handled otherwise, e.g. with the rolling forecast practice [van Valkenhoef et al., 2010]. In addition, our model assumes adherence to the standard best practices regarding story and theme sizes [Cohn, 2005]. Stories should be small enough so that they can easily be implemented in a single iteration. Themes should be small enough to be implemented in a single release. Moreover, the customer should understand that a theme should consist of the *minimal set of stories* required to achieve the aforementioned synergy effect, analogous to the concept of minimum marketable features [Denne and Cleland-Huang, 2003, 2004]. This ensures that the development team does not overcommit itself to a specific theme, thereby missing an opportunity to create more value elsewhere.

In iteration planning, the 'must have' stories are considered first. Since the completion probabilities $p_k < 1$, it is to be expected that not all planned stories will be implemented. Stories that are unlikely to be completed in the current release will be considered for inclusion in the next one, which is planned near the end of the current one.

D.3 Nested knapsack formulation

Our planning model is an instance of a nested knapsack problem [Dudziński and Walukiewicz, 1987]. It is defined by having a set of "knapsacks", each with a limited capacity, and a set of items, each with a size and value. The knapsacks themselves are nested, meaning that they are ordered in such a way that each knapsack contains the preceding one. The problem is then to maximize the value that fits into these knapsacks without exceeding their size limits. The problem is known to be NP-complete [Dudziński and Walukiewicz, 1987].

Let us define an index set of stories $S = \{1, \ldots, n\}$ and of themes $T = \{n + 1, \ldots, n + m\}$. All stories s_i and themes t_j have a business value, respectively u_i and u_j, and stories additionally have a size c_i:

$$\begin{aligned} u_i &\in \mathbb{N}\,; i \in S \cup T \\ c_i &\in \mathbb{N}\,; i \in S \end{aligned}$$

Let $L = \{1, \ldots, q\}$ denote the index set of story sets. Define q nested knapsacks, $\kappa_1, \ldots, \kappa_q$, such that

$$\kappa_k = \cup_{i \leq k} \ell_i \ ; \ k \in L$$

In contrast to the story sets, which are disjoint, each knapsack contains all preceding knapsacks. Associated with each knapsack κ_k is a budget $b_k \in \mathbb{N}$ that can be completed with a probability of at least p_k:

$$b_k = \lfloor F_c^{-1}(p_k) \rfloor \ ,$$

where F_c denotes the complementary cumulative distribution function of the estimated project velocity (see Figure D.3), derived from the estimated velocity distribution $f(v)$.

We define the decision variables for including story s_i in set ℓ_k and completing theme t_j by set ℓ_k as $x_{i,k}$ and $y_{j,k}$, respectively:

$$\begin{aligned} x_{i,k} &\in \{0, 1\}\,; i \in S, k \in L \\ y_{j,k} &\in \{0, 1\}\,; j \in T, k \in L \end{aligned}$$

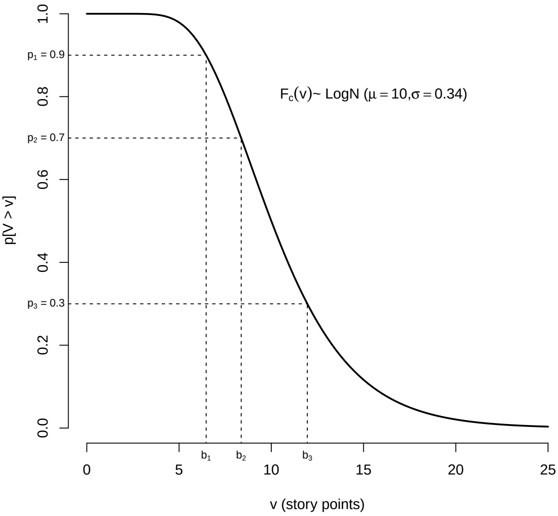


Figure D.3: Completion probabilities p_i and the complementary cumulative distribution $F_c(v)$ of the release velocity v define budgets for the 'must have' (b_1), 'should have' (b_2) and 'could have' (b_3) sets.

Now, we optimize the following objective function:

$$\begin{aligned} \max \quad & \sum_{k \in L} \sum_{i \in S} x_{i,k} p_k u_i + \sum_{k \in L} \sum_{j \in T} y_{j,k} p_k u_j \\ \text{s.t.} \quad & \sum_{i \in S} \sum_{h=1}^{k} c_i x_{i,h} \leq b_k \;\; \forall_{k \in L} && \text{(D.1)} \\ & \sum_{k \in L} x_{i,k} \leq 1 \;\; \forall_{i \in S} && \text{(D.2)} \end{aligned}$$

Constraint (D.2) ensures that the decision variable $x_{i,k}$ is set to 1 only for the set in which story s_i is included. The constraints in (D.1) are formulated in such a way that the story sizes included in sets prior to ℓ_k are included when evaluating the budget for knapsack κ_k. This is illustrated in Figure D.4.

The dependencies between completing themes and completing the individual stories within themes are accounted for by introducing a dummy decision variable

$$z_{j,k} \in \{0, 1\} \; ; j \in T, k \in L$$

such that $z_{j,k} = 1$ iff all stories in theme t_j are included in knapsack κ_k. To express

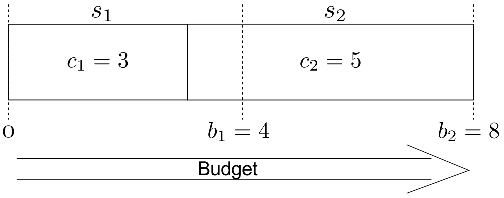


Figure D.4: A two-knapsack problem with budgets b_1 and b_2. The shown solution has a story s_1 in the 'must have' set and another story s_2 in the 'should have' set. Story s_1 is counted towards the budget of both knapsacks, and story s_2 uses the left-over capacity of the first knapsack to fit into the second.

this mathematically, we need the following constraints:

$$\left(\sum_{i\in S}\sum_{h=1}^{k} a_{i,j}x_{i,h}\right) - e_j z_{j,k} \geq 0 \quad ; \forall_{k\in L}\forall_{j\in T} \tag{D.3}$$

$$\left(\sum_{i\in S}\sum_{h=1}^{k} a_{i,j}x_{i,h}\right) - z_{j,k} \leq e_j - 1 \quad ; \forall_{k\in L}\forall_{j\in T} \tag{D.4}$$

Where $a_{i,j} = 1$ if story s_i is included in theme t_j and $a_{i,j} = 0$ otherwise, and e_j is the number of stories in theme t_j (i.e. $e_j = \sum_{i\in S} a_{i,j}$). If at least one of the stories belonging to theme j is not included in knapsack κ_k, $\sum_{i\in S}\sum_{h=1}^{k} a_{i,j}x_{i,h} < e_j$, in which case (D.3) ensures that $z_{j,k} = 0$. Similarly, if all stories belonging to theme t_j are completed in knapsack κ_k, $\sum_{i\in S}\sum_{h=1}^{k} a_{i,j}x_{i,h} = e_j$, in which case (D.4) ensures that $z_{j,k} = 1$. Finally, to make sure that $y_{j,k}$ is true iff $z_{j,k}$ is the first (in terms of k) for which $z_{j,k} = 1$, we add the following constraints:

$$y_{j,1} = z_{j,1} \qquad \forall_{j\in T} \tag{D.5}$$

$$y_{j,k} = z_{j,k} - z_{j,k-1} \quad \forall_{j\in T}\forall_{k\in L-\{1\}} \tag{D.6}$$

To complete the model, we note that if there are precedence relations, $i \prec j$ (i precedes j), they can be represented as

$$x_{j,k} - \sum_{h=1}^{k} x_{i,h} \leq 0 \quad \forall_{i\prec j}\forall_{k\in L} \tag{D.7}$$

Both technical and preference precedence relation are implemented in the same way, using constraint (D.7).

D.4 Theme valuation

The objective function of our optimization model combines theme and story values to obtain the total business value. This makes the theme values difficult to evaluate because on the one hand the value of completing a theme is an addition to completing the contained stories, whereas on the other hand all values should be on the same commensurable scale. We propose three theme elicitation approaches: constant theme value, ordinal evaluation and an indifference method.

A constant value c is appropriate if all themes have approximately the same business value to the customer. Ordinal evaluation of theme values is based on ranking the themes in an ascending order: the theme with the lowest business value is ranked at place 1 and the theme with the highest business value is ranked at place m. The theme value that is subsequently assigned to each of the themes should satisfy the ordering relation $\mathcal{R}$. A simple approach would be using a linear transformation function $h(\pi) = c\pi$, where π is the permutation of the theme index vector according to $\mathcal{R}$ (i.e. the j-th element of π denotes the rank of theme j). For example, if we have three themes, and our customer ranks them as $t_3 < t_1 < t_2$ (i.e. t_3 is least important), then $\pi = (2, 3, 1)$ and hence the value of t_1 would be $2c$.

The indifference method allows direct evaluation of the theme value. The idea is to consider a theme which (hypothetically) is one story away from being completed, and ask the customer whether he would like to complete the remaining story, or rather complete a set E of other (non-related) stories outside the theme. Such questions are asked until the customer states that he is indifferent between the two for some E. The theme value is then the difference of the sum of values of stories in E and the value of the remaining story in the theme. For example, if the last story to complete has value 3, and the customer is indifferent between completing the theme and completing three other stories with values 2, 3, and 5, then the value of the theme is $(2 + 3 + 5) - 3 = 7$. The indifference method is similar to the standard technique used in utility function elicitation (cf. [Keeney and Raiffa, 1976]).

D.5 Velocity estimation

A formal but simple velocity estimation method is extremely important as it has been shown that project management is often overly optimistic about the width of the 90% confidence interval for project duration [Jørgensen et al., 2004]. We propose a method that is consistent with software engineering literature from both the plan-driven and agile research communities.

We measure iteration velocity with a probability distribution on a scale of $[0, \infty)$ story points. The observed format of the distribution often corresponds to the log-normal one [Bourque et al., 2007]. A choice of a log-normal distribution is also consistent with the Bayesian model in [Hearty et al., 2009] as well as with the method of deriving confidence intervals for velocity proposed in [Cohn, 2005] and with NASA SEL guidelines [NASA, 1990]. The Bayesian network model [Hearty et al., 2009] can provide good velocity estimates, but is quite complex. We propose here a simple extension of the yesterday's weather model [Beck and Fowler, 2001].

Let us denote by $\mathbf{v}$ a vector of velocity observations from previous iterations. If we have a reasonable amount of observations, say at least five, we can estimate the log-normal velocity distribution through maximum likelihood with:

$$V_I \sim \log\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$$

where $\hat{\mu}$ is the mean of the log-transformed observations $\ln(\mathbf{v})$ and $\hat{\sigma}^2$ is the sample variance $\mathrm{sd}(\ln(\mathbf{v}))^2$. To estimate release velocity, we view a release as a collection of n_R independent iterations. Then V_R is the sum of n_R log-normal distributions, and can be estimated using the Fenton-Wilkinson 2-moment approximation [Fenton, 1960] simplified for equal mean and variance:

$$\begin{aligned}
V_R &\sim \log\mathcal{N}(\mu_R, \sigma_R^2) \\
\sigma_R^2 &\approx \ln(\exp(\hat{\sigma}^2) - 1 + n_R) - \ln n_R \\
\mu_R &\approx \hat{\mu} + \ln n_R + \frac{1}{2}(\hat{\sigma}^2 - \sigma_R^2)
\end{aligned}$$

The sample variance is overly precise in the beginning of a new project where there are only a small number (i.e. < 5) of observations, so instead of estimating it through maximum likelihood, we should take into account prior beliefs. A full Bayesian approach is possible but complex. We suggest using the following weighted sum (approximating an inverse-Gamma prior with prior degrees of freedom $\nu_0 = 2$):

$$\hat{\sigma} = \frac{\sigma_0 + n\mathrm{sd}(\ln(\mathbf{v}))}{1 + n} \tag{D.8}$$

where n is the number of observations and σ_0 is a prior belief of sample error that has a weight equal to one observation of true velocity. Note that (D.8) is applicable only if $n \geq 2$, otherwise the sample error is not defined. It should be expected that if σ_0 is reasonably large, $\hat{\sigma}$ will decrease as more observations become available.

The equation (D.8) requires a prior belief to be specified. In the complete absence of velocity observations, [Cohn, 2005] proposes using $\sigma_0 = 0.29$. This falls between the levels suggested by NASA SEL guidelines for plan-driven projects that have completed the requirements analysis ($\sigma_0 = 0.34$) and preliminary design ($\sigma_0 = 0.21$) phases, respectively. We suggest to err on the safe side and take $\sigma_0 = 0.34$. Table D.1 presents rules of thumb for specifying σ_0 for several levels of uncertainty, and gives confidence intervals at the 90% level conventional in software development [Jørgensen et al., 2004]. Note that [Cohn, 2005] suggests that the sample error should approach 0.06 as the number of iterations increases (see Table D.1).

D.6 Example

We are involved in a research project that includes a software deliverable. In a scientific setting, developers often have other tasks that overlap with the classical separation between the developer and the customer (see e.g. [Wood and Kleb, 2003]). Despite partly being customers for ourselves, we also have external customers that expect us to develop software artifacts that support a new way of working in the application domain of pharmacological decision making. The 'how' would be discovered only during the course of the project by exploring ways in which the software can support business processes.

Phase	Suggested CI	σ_0
Requirements Known *	$[\hat{\mu}/2.0, \hat{\mu} * 2.0]$	0.42
Requirements Analyzed *	$[\hat{\mu}/1.75, \hat{\mu} * 1.75]$	0.34
$<$ 2 Iterations Completed	$[\hat{\mu} * 0.60, \hat{\mu} * 1.60]$	0.29
Preliminary Design *	$[\hat{\mu}/1.40, \hat{\mu} * 1.40]$	0.21
Detailed Design *	$[\hat{\mu}/1.25, \hat{\mu} * 1.25]$	0.14
2 Iterations Completed	$[\hat{\mu} * 0.8, \hat{\mu} * 1.25]$	0.14
3 Iterations Completed	$[\hat{\mu} * 0.85, \hat{\mu} * 1.15]$	0.08
$>$ 3 Iterations Completed	$[\hat{\mu} * 0.90, \hat{\mu} * 1.10]$	0.06

Table D.1: Rules of thumb for uncertainty in velocity, from [NASA, 1990] (marked with *) and [Cohn, 2005]. σ_0 for a log-normal distribution to produce the CIs at the 90% level is given to two decimal places.

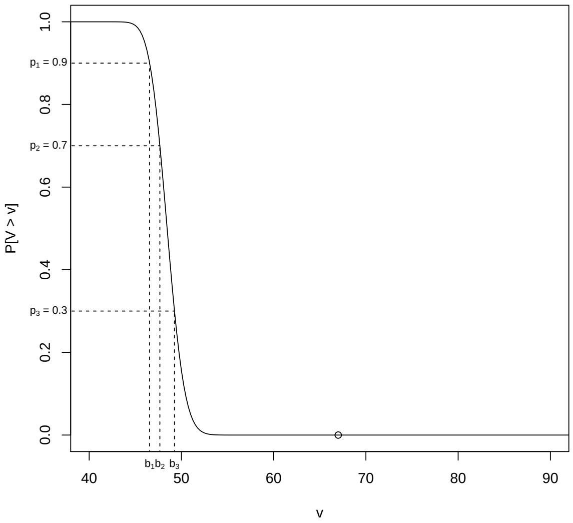
Our development environment consists of 2 teams working part-time. We programmed in Java SE 1.5 with Eclipse. All our code is available as open source. More information about the project and links to the code repository and to downloadable releases can be found at `http://www.drugis.org`. We retrospectively evaluated our velocity estimation techniques using data from this project.

We did the first release (ADDIS 0.2) as a burn-in for velocity estimation and for creating an initial end-to-end working system. Therefore we did not estimate story business values while planning the first release. During the second and third releases (ADDIS 0.4 and 0.6), we estimated story business values (scale: 1-5), story sizes (scale: 1, 2, 3, 5, 8), technical precedence relations and themes.

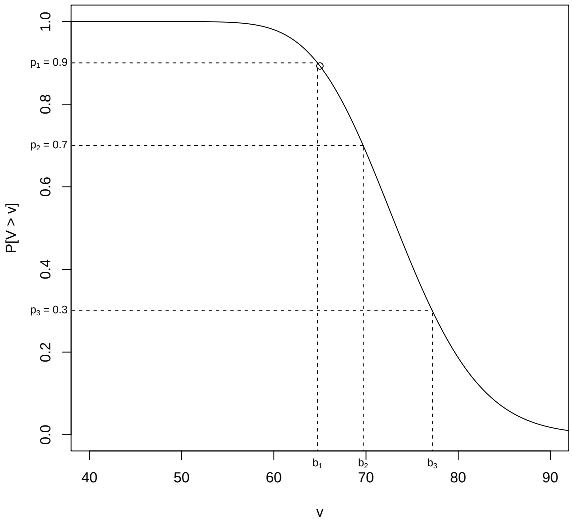
In the first release we had 4-week iterations, and we planned switching to 2-week iterations. With this in mind, the corrected velocity was $\mathbf{v}_1 = (8.5, 10, 9)$. With $\sigma_0 = 0.34$, the estimated velocity is $V_I \sim \log\mathcal{N}(2.2, 0.15^2)$. And, with 5 iterations, $V_R \sim \log\mathcal{N}(3.9, 0.031^2)$, the complementary cumulative distribution is shown in Figure D.5(a). Velocity for the second release was $\mathbf{v}_2 = (7, 16, 17, 9, 18)$. This irregular sequence was the result of doubling the size of our development team after the first iteration. Actual velocity was well outside the predicted 95% confidence interval, showing that our decision to hire additional programmers was effective. This is reflected in both increased expected velocity and increased uncertainty for the third release predicted on the basis of $\mathbf{v}_2$: $V_R \sim \log\mathcal{N}(4.3, 0.097^2)$, as shown in Figure D.5(b). For this release, actual velocity was slightly over b_1, the 'must have' budget.

In case the predicted velocity distribution is like the one in Figure D.5(a), the deterministic model we proposed previously [van Valkenhoef et al., 2010] may be appropriate due to the small budget for 'should have' and 'could have'. However, with greater uncertainty, such as in Figure D.5(b), a probabilistic method such as the one proposed here is more suited. Note that if for some reason the actual velocity differs greatly from the predicted, replanning may be required. The planning model can then be used to suggest a revised plan for the remaining iterations.

We applied the planning model retrospectively to the ADDIS 0.6 planning problem, obtaining a plan reasonably similar to the one actually implemented. For ADDIS 0.8, we used the model to create a preliminary plan that was adopted with some



(a) V_R for R2 (based on $\mathbf{v}_1$)



(b) V_R for R3 (based on $\mathbf{v}_2$)

Figure D.5: The complementary cumulative velocity distributions estimated for release 2 (from release 1 velocity) and release 3 (from release 2 velocity). Due to higher variability during release 2, the estimated velocity is much less certain. The ∘ shows the velocity that was actually achieved.

modifications by our customer. Although our planning model reduced the time and effort required in release planning, managing the stories and entering them in the model was quite cumbersome due to lack of usable software that would integrate story management.

D.7 Computational tests

We implemented the supporting release planning model using R [R Development Core Team, 2008] and lp_solve [Berkelaar et al.] (using a branch-and-bound algorithm). Our implementation is freely available online at `http://github.com/gertvv/xpplan`. The nested knapsack problem is NP-complete, but an exact solution is often feasible in practice due to small problem instance sizes. Knapsack problems are widely analyzed in the literature (see e.g. [Pisinger, 1997]), but our model has a non-standard structure. For this reason, we analyzed the running time with randomly generated problem instances with different numbers of stories and themes. The story values and sizes were sampled from $\{1, 2, 3, 4, 5\}$ and $\{1, 2, 3, 5, 8\}$, respectively. The completion probabilities were $p = (0.9, 0.7, 0.3)$, and the velocity distribution

$$V_R \sim \log \mathcal{N}(0.5c_T, 0.34^2) \ ; \ c_T = \sum_{i \in S} c_i.$$

The themes were valued through ordinal evaluation (Section D.4) with random ranks and $u_{n+j} = 5\pi(j)$. Each theme contained between 3 and 10 randomly assigned stories.

We tested running times for 10 to 50 stories (step size 5) with 2 to 10 themes (step size 2). For each problem size we ran 10 tests on an Intel Core 2 Duo 3Ghz CPU with Ubuntu 9.10 and no relevant extra load during the tests. We set a time-out of 2 hours for running the model. Figure D.6 illustrates the minimum, maximum, mean, and median running times of different problem sizes. The spikes in max and mean running times are due to problems that were not solved within the time limit. This occurred once for 35 stories with 10 themes, twice for 45 and 50 stories with 10 themes, and twice for 50 stories with 8 themes.

The mean running times in Figure D.6 show that the problem gets harder as the amount of themes increases. The maximum running times indicate that problem instances with 2-6 themes were solvable within one hour. With 2 or 4 themes the median running times are very low, which hints that a lot larger problems could be solved exactly. The minimum running times are relatively low for all problem sizes, so some instances are very easy to solve even for larger amounts of themes.

D.8 Conclusions

Release planning in Extreme Programming (XP) can cause prioritization stress for the customer and is impractical in larger projects. To remedy this, we developed an optimization model to support release planning. Our model evaluates the stories

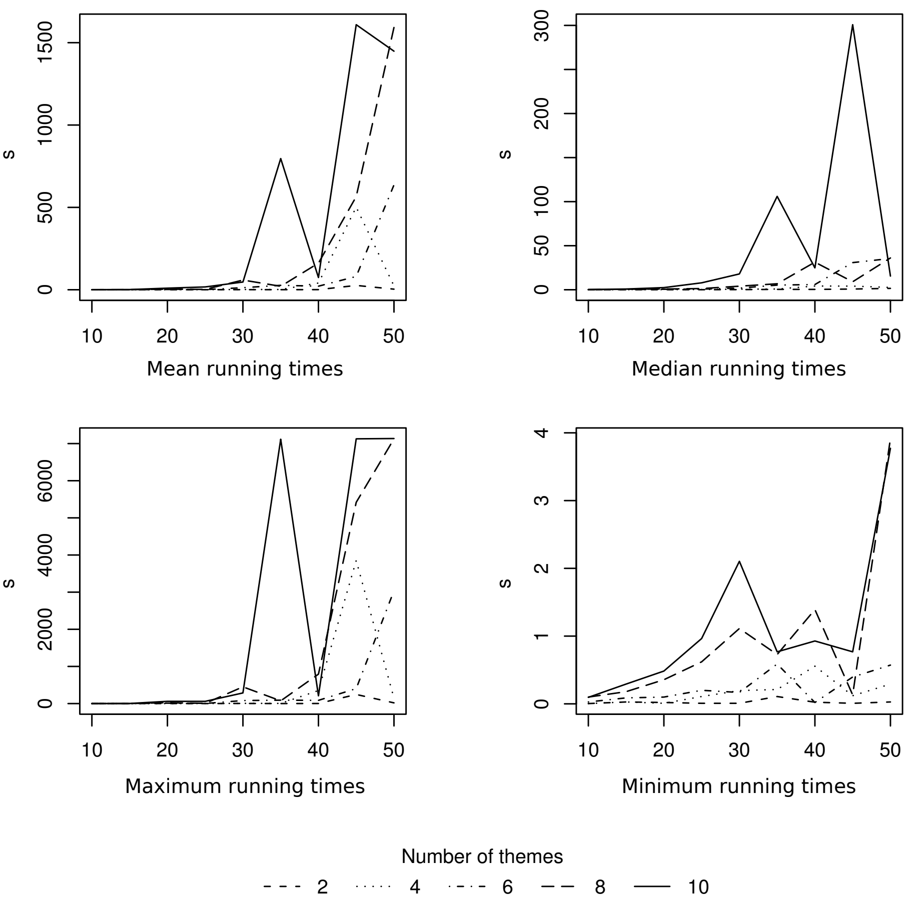


Figure D.6: Running time statistics for different model sizes. Number of stories is on the x-axis.

with regard to their size, business value, and technical and preference precedence relations, and incorporates synergy effects among stories with themes. The model assumes availability of a velocity distribution. We discussed simple rules of thumb for estimating an appropriate log-normal velocity distribution and evaluated them with an example. Our experiences suggest that the model is both feasible and useful in practice.

It has been argued that there is a lack of diversity among the existing release planning models [Svahnberg et al., 2010]. We address this by introducing a model compatible with XP values and practices [Beck, 2005] that is optimized for higher levels of requirements change and uncertainty in the velocity estimates by requiring less 'up front' specification and estimation. Most other models, in contrast, do not consider uncertainty in release velocity and operate at a more fine-grained level of detail (e.g. [Greer and Ruhe, 2004, Saliu and Ruhe, 2007, van den Akker et al., 2008]). One could also consider imprecise story sizes [Ruhe and Greer, 2003], but our approach, when combined with an appropriate velocity estimation method (e.g. [Hearty et al., 2009]), appears to be more general. In the framework of Boehm and Turner [Boehm and Turner, 2003], our model can be considered more 'agile', while the existing models exhibit more 'discipline'. Still, the use of our model introduces some additional discipline to enable XP outside its home ground.

The optimization model is of exponential complexity, but our computational tests showed that problems with up to 6 themes and 50 stories can be solved exactly in less than an hour with a standard PC. This is sufficient for practical use in up to medium sized projects. However, [Carlshamre, 2002] suggests that for quantitative models to be useful, running time should be minimized (immediate interaction) and several good solutions may be more valuable than a single optimal one. Future research should develop and evaluate approximate methods to address this. For example, the genetic algorithm proposed in [Greer and Ruhe, 2004] solves a similar problem, and might be adaptable to our model. In addition, we chose to model dependencies through precedence relations and themes, although other relations (e.g. exclusion [van den Akker et al., 2008]) could be considered as well. Future research should address which model features are actually useful in supporting XP release planning, and which ones just cause additional cognitive burden. Finally, usable software is important to enable the use of quantitative planning methodologies such as the one presented here. A simple planning model implemented in a "provotype" tool with a graphical user interface was presented and evaluated in [Carlshamre, 2002], and the requirements for release planning and story management software in an agile environment were discussed in [Koponen, 2008]. Our model can be implemented as a decision support tool in such a system. However, a planning model and software alone cannot solve the project management problems for larger projects. Other socio-technical aspects of agile software management should also be addressed. A recent review [Dybå and Dingsøyr, 2008] has shown that currently there aren't sufficient empirical studies on agile software management to draw firm conclusions on "goodness" of the methods. We acknowledge that this holds also with respect to our work, and empirical studies should be initiated to consider the model's applicability on different types of development projects.

Acknowledgements

First and foremost I thank dr. Tommi Tervonen, my daily supervisor for the first half of the project and partner in science ever since. He has constantly pushed me to get things done and do things better and was happy to have me push right back. A big thanks also goes to dr. Douwe Postmus for a great collaboration as well as his support in general. Equally to Jing Zhao, our MSc student and research assistant, for all the help and enthousiasm she offered. You have not just been colleagues, but also friends.

I thank my supervisors, prof. dr. Hans Hillege and prof. dr. Bert de Brock, for all the things they made possible, for their vision and support, and for the great confidence they placed in me. I thank my co-authors in Bristol, prof. dr. Tony Ades, dr. Nicky Welton, dr. Sofia Dias, and Guobing Lu, for accepting my plea for collaboration and for pushing my understanding of evidence synthesis to the next level.

During most of the project, I had the great fortune to manage a small and ever changing team of programmers: Tijs Zwinkels, Maarten Jacobs, Hanno Koeslag, Daan Reid, Florin Schimbinschi, Ahmad Kamal, Wouter Reckman, and Joël Kuiper. Their hard work has made this project possible – thanks! I am grateful to my colleagues of the Escher project for the many fruitful workshops that provided invaluable feedback throughout the project. I specifically thank Michelle Putzeist for almost co-authoring a paper with me and Hans van Leeuwen and Marcel Hekking for their input on many parts of the project. Thanks also to dr. Nalan Baştürk for a great collaboration.

I also thank my colleagues at the Faculty of Economics and Business, University of Groningen and at the Department of Epidemiology, University Medical Center Groningen, as well as my friends (especially Anna Muijzer and Arjen Zonneveld) and family. A happy mind and a positive attitude are indispensable in an undertaking such as this and they made sure I kept both! Finally, a special thank you to my girlfriend, Brechtsje Kingma, for being awesome!

Gert van Valkenhoef
Groningen
October 21, 2012

Bibliography

P. Abrahamsson, J. Warsta, M. T. Siponen, and J. Ronkainen. New directions on agile methods: a comparative analysis. In *IEEE Proceedings of the International Conference on Software Engineering*, pages 244–254, Portland, Oregon, USA, 2003. doi: 10.1109/ICSE.2003.1201204.

A. Al-Emran, P. Kapur, D. Pfahl, and G. Ruhe. Studying the impact of uncertainty in operational release planning-An integrated method and its initial evaluation. *Information and Software Technology*, 52(4):446–461, 2010. doi: 10.1016/j.infsof.2009.11.003.

AMIA. AMIA global trial bank, 2010. URL `https://www.amia.org/global-trial-bank`. Archived at `http://www.webcitation.org/5ma1fjB1A`.

Y. Amit and U. Grenander. Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, 37(2):197–222, 1991. doi: 10.1016/0047-259X(91)90080-L.

D. Avis and K. Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(1):295–313, 1995. doi: 10.1007/BF02293050.

K. Beck. *Extreme Programming Explained*. Addison-Wesley, 1st edition, 1999.

K. Beck. *Extreme Programming Explained*. Addison-Wesley, 2nd edition, 2005.

K. Beck and M. Fowler. *Planning Extreme Programming*. Addison-Wesley, 2001.

C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, 276(8):637–639, 1996. doi: 10.1001/jama.1996.03540080059030.

V. Belton and T. J. Stewart. *Multiple Criteria Decision Analysis - An Integrated Approach*. Kluwer Academic Publishers, Dordrecht, 2002.

S. K. Berberian. *Introduction to Hilbert Space*. American Mathematical Society, 1999. ISBN 0-8218-1912-7.

M. Berkelaar, K. Eikland, and P. Notebaert. lpsolve 5.5 – open source (mixed-integer) linear programming system. URL `http://lpsolve.sf.net/`.

J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 1995. doi: 10.1214/ss/1177010123.

Biomedical Research Integrated Domain Group (BRIDG). *BRIDG Model Release 3.0.3 User's Guide*, 2010. URL `http://gforge.nci.nih.gov/frs/?group_id=342`.

P. Bleicher. Clinical trial technology: at the inflection point. *Biosilico*, 1(5):163–168, 2003. doi: 10.1016/S1478-5382(03)02373-4.

O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–270, 2004. doi: 10.1093/nar/gkh061.

B. Boehm and R. Turner. *Balancing agility and discipline: a guide to the perplexed*. Addison Wesley, 2003.

F. Boudin, J. Y. Nie, J. C. Bartlett, R. Grad, P. Pluye, and M. Dawes. Combining classifiers for robust PICO element detection. *BMC Medical Informatics and Decision Making*, 10:29, 2010. doi: 10.1186/1472-6947-10-29.

P. Bourque, S. Oligny, A. Abran, and B. Fournier. Developing project duration models in software engineering. *Journal of Computer Science and Technology*, 22(3):348–357, 2007. doi: 10.1007/s11390-007-9051-5.

C. Brandt, P. Nadkarni, L. Marenco, B. Karras, C. Lu, L. Schacter, J. Fisk, and P. Miller. Reengineering a database for clinical trials management: Lessons for system architects. *Control Clin Trials*, 21:440–461, 2000.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998. doi: 10.1080/10618600.1998.10474787.

J. Butler, J. Dia, and J. Dyer. Simulation techniques for the sensitivity analysis of multi-criteria decision models. *European Journal of Operational Research*, 103(3):531–545, 1997.

D. M. Caldwell, N. J. Welton, and A. E. Ades. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *Journal of Clinical Epidemiology*, 63(8):875 – 882, 2010. doi: 10.1016/j.jclinepi.2009.08.025.

D. M. Caldwell, A. E. Ades, and J. P. T. Higgins. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*, 331(7521):897–900, 2005. doi: 10.1136/bmj.331.7521.897.

L. Cao, K. Mohan, and P. Xu. How extreme does extreme programming have to be? adapting XP practices to large-scale projects. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, Waikoloa, Hawaii, 2004. doi: 10.1109/HICSS.2004.1265237.

S. Carini, B. H. Pollock, H. P. Lehmann, S. Bakken, E. M. Barbour, D. Gabriel, H. K. Hagler, C. R. Harper, S. A. Mollah, M. Nahm, H. H. Nguyen, R. H. Scheuermann, and I. Sim. Development and evaluation of a study design typology for human research. *AMIA Annu Symp Proc*, 2009: 81–85, 2009.

P. Carlshamre. Release planning in market-driven software product development: Provoking an understanding. *Requirements Engineering*, 7(3):139–151, 2002. doi: 10.1007/s007660200010.

CDISC. CDISC 2004 research project on attitudes, adoption, and usage of data collection technologies and data interchange standards; executive summary, Sept 2005.

CDISC and FDA. Walking down the critical path: The application of data standards to FDA submissions. A Discussion Paper by CDISC and FDA, February 2005.

I. Chalmers. The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: the need for systematic reviews. In P. Rothwell, editor, *Treating individuals: from randomised trials to personalised medicine*, chapter 2, pages 37–58. Elsevier, Edinburgh, Scotland, 2007.

A. Chan, A. Hrobjarttson, M. Haahr, P. Gotzsche, and D. Altman. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, 291(20):2457–2465, 2004.

A.-W. Chan. Bias, spin, and misreporting: Time for full access to trial protocols and results. *PLoS Medicine*, 5(11):e230, 11 2008. doi: 10.1371/journal.pmed.0050230.

A. W. Chan and D. G. Altman. Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, 365(9465):1159–1162, 2005. doi: 10.1016/S0140-6736(05)71879-1.

E. U. Choo, B. Schoner, and W. C. Wedley. Interpretation of criteria weights in multicriteria decision making. *Computers & Industrial Engineering*, 37(3):527–541, 1999. ISSN 0360-8352. doi: 10.1016/S0360-8352(00)00019-X.

T. Chow and D.-B. Cao. A survey study of critical success factors in agile software projects. *Journal of Systems and Software*, 81(6):961–971, 2008. doi: 10.1016/j.jss.2007.08.020.

J. J. Cimino. Review paper: coding systems in health care. *Methods of Information in Medicine*, 35(4–5):273–284, 1996.

A. Cipriani, T. A. Furukawa, G. Salanti, J. R. Geddes, J. P. T. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I. M. Omori, H. McGuire, M. Tansella, and C. Barbui. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet*, 373(9665):746 – 758, 2009. doi: 10.1016/S0140-6736(09)60046-5.

H. G. Claycamp. Rapid benefit-risk assessment: no escape from expert judgments in risk management. *Risk Analysis*, 26(1), 2006.

ClinicalTrials.gov. Clinicaltrials.gov protocol data element definitions (draft), November 2009a. URL `http://prsinfo.clinicaltrials.gov/definitions.html`. Archived at `http://www.webcitation.org/5mYe6OxvP`.

ClinicalTrials.gov. Linking to clinicaltrials.gov, April 2009b. URL `http://clinicaltrials.gov/ct2/info/linking`. Archived at `http://www.webcitation.org/5mZlpYkGW`.

ClinicalTrials.gov. Clinicaltrials.gov “basic results” data element definitions (draft), September 2009c. URL `http://prsinfo.clinicaltrials.gov/results_definitions.html`.

ClinicalTrials.gov. Linking to clinicaltrials.gov, December 2011. URL `http://clinicaltrials.gov/ct2/info/linking`. Archived at `http://www.webcitation.org/66HHPj8RI`.

ClinPage. Trial-friendly EHR systems, February 2009. URL `http://www.clinpage.com/article/trial-friendly_ehr_systems/C9`. Archived at `http://www.webcitation.org/5mjBWgITQ`.

A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6:57–71, 2005. doi: 10.1093/bib/6.1.57.

M. Cohn. *Agile Estimating and Planning*. Robert C. Martin Series. Prentice Hall PTR, 2005. ISBN 0131479415.

N. Cooper, D. Coyle, K. Abrams, M. Mugford, and A. Sutton. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *Journal of Health Services Research & Policy*, 10(4):245–250, 2005. doi: 10.1258/135581905774414187.

P. M. Coplan, R. A. Noel, B. S. Levitan, J. Ferguson, and F. Mussen. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clinical Pharmacology and Therapeutics*, 89(2):312–315, 2011. doi: 10.1038/clpt.2010.291.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2nd edition, 2001. ISBN 0-262-03293-7.

E. T. Crumley, N. Wiebe, K. Cramer, T. P. Klassen, and L. Hartling. Which resources should be used to identify RCT/CCTs for systematic reviews: a systematic review. *BMC Med Res Methodol*, 5:24, 2005. doi: 10.1186/1471-2288-5-24.

Current Controlled Trials Ltd. Current controlled trials: Frequently asked questions, 2010. URL `http://www.controlled-trials.com/information/faqs`. Archived at `http://www.webcitation.org/5mbV14caj`.

M. Denne and J. Cleland-Huang. *Software by Numbers: Low-Risk, High-Return Development*. Prentice-Hall, 2003.

M. Denne and J. Cleland-Huang. The incremental funding method: Data-driven software development. *IEEE Software*, 21:39–47, 2004. doi: 10.1109/MS.2004.1293071.

N. Deo, G. M. Prabhu, and M. S. Krishnamoorthy. Algorithms for generating fundamental cycles in a graph. *ACM Transactions on Mathematical Software*, 8(1):26–42, 1982. doi: 10.1145/355984.355988.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188, 1986. doi: 10.1016/0197-2456(86)90046-2.

S. Dias, N. J. Welton, A. J. Sutton, and A. E. Ades. NICE DSU technical support document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. Technical report, 2011. URL `http://www.nicedsu.org.uk/`. updated August 2011.

S. Dias, N. J. Welton, D. M. Caldwell, and A. E. Ades. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29(7-8, Sp. Iss. SI):932–944, 2010. doi: 10.1002/sim.3767.

K. Dickersin and D. Rennie. Registering clinical trials. *Journal of the American Medical Informatics Association*, 290(4):516–523, 2003. doi: 10.1001/jama.290.4.516.

K. Dickersin, E. Manheimer, S. Wieland, K. A. Robinson, C. Lefebvre, and S. McDonald. Development of the Cochrane Collaboration's central register of controlled clinical trials. *Evaluation & the Health Professions*, 25(38):38–64, 2002. doi: 10.1177/016327870202500104.

K. Dudziński and S. Walukiewicz. Exact methods for the knapsack problem and its generalizations. *European Journal of Operational Research*, 28(1):3–21, 1987. doi: 10.1016/0377-2217(87)90165-2.

T. Dybå and T. Dingsøyr. Empirical studies of agile software development: A systematic review. *Information and Software Technology*, 50:833–859, 2008. doi: 10.1016/j.infsof.2008.01.006.

M. Egger, G. D. Smith, and A. N. Phillips. Meta-analysis: principles and procedures. *BMJ*, 315 (7121):1533–1537, 1997.

H.-G. Eichler, F. Pignatti, B. Flamion, H. Leufkens, and A. Breckenridge. Balancing early market access to new drugs with the need for benefit/risk data: a mounting dilemma. *Nature Reviews Drug Discovery*, 7(10):818–836, 2008. doi: 10.1038/nrd2664.

H.-G. Eichler, B. Aronsson, E. Abadie, and T. Salmonson. New drug approval success rate in europe in 2009. *Nat Rev Drug Discov*, 9:355–356, 2010. doi: 10.1038/nrd3169.

K. El Emam, E. Jonker, M. Sampson, K. Krleza-Jeric, and A. Neisa. The use of electronic data capture tools in clinical trials: Web-survey of 259 canadian trials. *J Med Internet Res*, 11(1):e8, 2009. doi: 10.2196/jmir.1120.

EMEA. ICH Topic E 2 C (R1), clinical safety data management: Periodic safety update reports for marketed drugs, step 5: European Medicines Agency (EMEA), 1997.

EMEA. A guideline on summary of product characteristics: The rules governing medicinal products in the European Union. EudraLex, Vol. 2C, 2005.

EMEA. Reflection paper on expectations for electronic source documents used in clinical trials (draft). European Medicines Agency (EMEA). Report No.: EMEA/505620/2007, 2007.

EMEA. Reflection paper on benefit-risk assessment methods in the context of the evaluation of marketing authorization applications of medicinal products for human use. Committee for medicinal products for human use, European Medicines Agency (CHMP-EMEA). Report No.: EMEA/CHMP/15404/2007, 2007. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069634.pdf`.

European Commision. A guideline on summary of product characteristics (SmPC), September 2009. URL `http://ec.europa.eu/enterprise/sectors/pharmaceuticals/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf`.

European Medicines Agency. PIM: data exchange standard, 2010a. URL `http://pim.emea.europa.eu/des/index.html`. Archived at `http://www.webcitation.org/5nlDWKtab`.

European Medicines Agency. Human medicines - quality review of documents (QRD), 2010b. URL `http://www.emea.europa.eu/htms/human/qrd/qrdtemplate.htm`. Archived at `http://www.webcitation.org/5nlDmjKRR`.

European Medicines Agency. EudraCT public web report for february 2011, 2011a. URL `https://eudract.ema.europa.eu/docs/statistics/EudraCT_Statistics_February.pdf`.

European Medicines Agency. About EU clinical trials register, 2011b. URL `https://www.clinicaltrialsregister.eu/about.html`. Archived at `http://www.webcitation.org/5yNkGPZZX`.

European Medicines Agency. Eu clinical trials register, 2012. URL `https://www.clinicaltrialsregister.eu/`. Archived at `http://www.webcitation.org/66HHLkoV1`.

Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association*, 268(17):2420–2425, 1992. doi: 10.1001/jama.1992.03490170092032.

FDA. US Food and Drug Administration Amendments Act (FDAAA), Section 801, 2007.

FDA. Guidance for industry: Providing regulatory submissions in electronic format – drug establishment registration and drug listing. US Food and Drug Administration (FDA). OMB Control No. 0910-0045, May 2009.

G. Fegan and T. Lang. Could an open-source clinical trial data-management system be what we have all been looking for? *PLoS Medicine*, 5(3):347–349, 2008. doi: 10.1371/journal.pmed.0050006.

J. C. Felli, R. A. Noel, and P. A. Cavazzoni. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. *Medical Decision Making*, 29(1):104–115, 2009. doi: 10.1177/0272989X08323299.

L. Fenton. The sum of log-normal probability distributions in scatter transmission systems. *Institute of Radio Engineers Transactions on Communication Systems*, CS-8(1):57–67, March 1960.

Food and Drug Administration. Janus operational pilot, 2010a. URL `http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm`. Archived at `http://www.webcitation.org/5nmLXVuZE`.

Food and Drug Administration. Structured product labeling resources, 2010b. URL `http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm`. Archived at `http://www.webcitation.org/5nlE9bM4Y`.

M. Fowler. *UML Distilled*. Addison-Wesley, 3rd edition, 2003.

D. Fridsma, J. Evans, S. Hastak, and C. Mead. The BRIDG project: A technical report. *J Med Inform Assoc*, 15(2):130–137, 2008.

J. Friedman and L. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979. doi: 10.1214/aos/1176344722.

H. N. Gabow and E. W. Myers. Finding all spanning trees of directed and undirected graphs. *SIAM Journal on Computing*, 7(3):280–287, 1978. doi: 10.1137/0207024.

L. P. Garrison, Jr., A. Towse, and B. W. Bresnahan. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. *Health Affairs*, 26(3):684–695, 2007.

W. H. Geerts, R. M. Jay, K. I. Code, E. Chen, J. P. Szalai, E. A. Saibil, and P. A. Hamilton. A comparison of low-dose heparin with low-molecular-weight heparin as prophylaxis against venous thromboembolism after major trauma. *New England Journal of Medicine*, 335:701–707, 1996. doi: 10.1056/NEJM199609053351003.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136.

D. Ghersi, M. Clarke, J. Berlin, A. M. Guelmezoglu, R. Kush, P. Lumbiganon, D. Moher, F. Rockhold, I. Sim, and E. Wager. Reporting the findings of clinical trials: a discussion paper. *Bulletin of the World Health Organization*, 86(6):492–493, 2008. doi: 10.2471/BLT.08.053769.

T. L. Graves. *An Introduction to YADAS*, September 2008. URL `http://yadas.lanl.gov/`.

S. Greco, V. Mousseau, and R. Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):415–435, 2008. doi: 10.1016/j.ejor.2007.08.013.

S. Greco, V. Mousseau, and R. Słowiński. Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(4):1455–1470, 2010. doi: 10.1016/j.ejor.2010.05.021.

D. Greer and G. Ruhe. Software release planning: an evolutionary and iterative approach. *Information and Software Technology*, 46(4):243 – 253, 2004. doi: 10.1016/j.infsof.2003.07.002.

J. M. Grimshaw, N. Santesso, M. Cumpston, A. Mayhew, and J. McGowan. Knowledge for knowledge translation: the role of the Cochrane Collaboration. *Journal of Continuing Education in the Health Professions*, 26(1):55–62, 2006. doi: 10.1002/chp.51.

P. S. Grisham and D. E. Perry. Customer relationships and extreme programming. In *HSSE '05: Proceedings of the 2005 workshop on Human and social factors of software engineering*, pages 1–6, St. Louis, Missouri, USA, 2005. doi: 10.1145/1083106.1083113.

L. Grobler, N. Siegfried, L. Askie, L. Hooft, P. Tharyan, and G. Antes. National and multinational prospective trial registers. *Lancet*, 372(9645):1201–1202, 2008. doi: 10.1016/S0140-6736(08)61498-1.

J. J. Guo, S. Pandey, J. Doyle, B. Bian, Y. Lis, and D. W. Raisch. A Review of Quantitative Risk-Benefit Methodologies for Assessing Drug Safety and Efficacy-Report of the ISPOR Risk-Benefit Management Working Group. *Value in Health*, 13(5):657–666, 2010. doi: 10.1111/j.1524-4733.2010.00725.x.

R. A. Hansen, G. Gartlehner, K. N. Lohr, B. N. Gaynes, and T. Carey. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. *Annals of Internal Medicine*, 143(6):415–426, 2005.

D. Hartmann and R. Dymond. Appropriate agile measurement: using metrics and diagnostics to deliver business value. In *Agile Conference, 2006*, pages 6 pp. –134, 2006. doi: 10.1109/AGILE.2006.17.

R. Hassin and A. Tamir. On the minimum diameter spanning tree problem. *Information Processing Letters*, 53(2):109–111, 1995. doi: 10.1016/0020-0190(94)00183-Y.

R. B. Haynes, P. J. Devereaux, and G. H. Guyatt. Clinical expertise in the era of evidence-based medicine and patient choice. *Evidence Based Medicine*, 7:36–38, 2002. doi: 10.1136/ebm.7.2.36.

R. B. Haynes, K. A. McKibbon, N. L. Wilczynski, S. D. Walter, and S. R. Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*, 330(7501):1179, 2005. doi: bmj.38446.498542.8Fv1.

P. Hearty, N. Fenton, D. Marquez, and M. Neil. Predicting project velocity in XP using a learning dynamic bayesian network model. *IEEE Transactions on Software Engineering*, 35(1):124–137, 2009. doi: 10.1109/TSE.2008.76.

L. V. Hedges and J. L. Vevea. Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4):486–504, 1998. doi: 10.1037/1082-989X.3.4.486.

J. Higgins and S. Green, editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]*. The Cochrane Collaboration, 2009. Available from http://www.cochrane-handbook.org.

J. P. T. Higgins and A. Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24):2733–2749, 1996. doi: 10.1002/(SICI)1097-0258(19961230)15:24⟨2733::AID-SIM562⟩3.0.CO;2-0.

J. Hokkanen, R. Lahdelma, K. Miettinen, and P. Salminen. Determining the implementation order of a general plan by using a multicriteria method. *Journal of Multi-Criteria Decision Analysis*, 7(5):273–284, 1998. doi: 10.1002/(SICI)1099-1360(199809)7:5⟨273::AID-MCDA198⟩3.0.CO;2-1.

J. Hokkanen, R. Lahdelma, and P. Salminen. A multiple criteria decision model for analyzing and choosing among different development patterns for the helsinki cargo harbor. *Socio-Economic Planning Sciences*, 33:1–23, 1999.

W. L. Holden. Benefit-risk analysis: a brief review and proposed quantitative approaches. *Drug Safety*, 26(12):853–862, 2003. doi: 10.2165/00002018-200326120-00002.

P. K. Honig. Systematic reviews and meta-analyses in the new age of transparency. *Clinical Pharmacology and Therapeutics*, 88(2):155–158, 2010. doi: 10.1038/clpt.2010.124.

B. Hughes. 2009 FDA drug approvals. *Nature Reviews Drug Discovery*, 9(2):89–92, 2010. doi: 10.1038/nrd3101.

ICTRP. About trial registration: Organizations with policies, 2010. URL `http://www.who.int/ictrp/trial_reg/en/index2.html`. Archived at `http://www.webcitation.org/5mZpckdBf`.

J. P. A. Ioannidis. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *Canadian Medical Association Journal*, 181(8):488–493, 2009. doi: 10.1503/cmaj.081086.

A. Irs, T. Janse de Hoog, and L. Rägo. Development of marketing authorisation procedures for pharmaceuticals. In N. Freemantle and S. Hill, editors, *Evaluating Pharmaceuticals for Health Policy and Reimbursement*, pages 3–23. Blackwell Science Ltd, 2004. doi: 10.1002/9780470994719.ch2.

ISO TC 215 (Health Informatics). ISO/PRF TS 29585: Deployment of a clinical data warehouse, 2011. URL `http://www.iso.org/iso/catalogue_detail.htm?csnumber=45582`.

A. R. Jadad, D. J. Cook, A. Jones, T. P. Klassen, P. Tugwell, M. Moher, and D. Moher. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *Journal of the American Medical Association*, 280:278–280, 1998. doi: 10.1001/jama.280.3.278.

J. Jia, G. W. Fischer, and J. S. Dyer. Attribute weighting methods and decision quality in the presence of reponse error: a simulation study. *Journal of Behavioral Decision Making*, 11:85–105, 1998.

M. Jørgensen, K. Teigen, and K. Moløkken. Better sure than safe? over-confidence in judgement based software development effort prediction intervals. *Journal of Systems and Software*, 70: 79–93, 2004. doi: 10.1016/S0164-1212(02)00160-7.

J. Kaiser. Making clinical data widely available. *Science*, 322(5899):217–218, 2008. doi: 10.1126/science.322.5899.217.

A. Kangas, J. Kangas, R. Lahdelma, and P. Salminen. Using SMAA-2 method with dependent uncertainties for strategic forest planning. *Forest Policy and Economics*, 9:113–125, 2006. doi: 10.1016/j.forpol.2005.03.012.

J. Kangas and A. Kangas. Multicriteria approval and SMAA-O method in natural resources decision analysis with both ordinal and cardinal criteria. *Journal of Multi-Criteria Decision Analysis*, 12(1):3–15, 2003. doi: 10.1002/mcda.344.

J. Kangas, J. Hokkanen, A. Kangas, R. Lahdelma, and P. Salminen. Applying stochastic multicriteria acceptability analysis to forest ecosystem management with both cardinal and ordinal criteria. *Forest Science*, 49(6):928–937, 2003.

T. J. Kaptchuk, E. Friedlander, J. M. Kelley, M. N. Sanchez, E. Kokkotou, J. P. Singer, M. Kowalczykowski, F. G. Miller, I. Kirsch, and A. J. Lembo. Placebos without deception: A randomized controlled trial in irritable bowel syndrome. *PLoS ONE*, 5(12):e15591, 2010. doi: 10.1371/journal.pone.0015591.

S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, 10:58, 2010. doi: 10.1186/1472-6947-10-58.

J. Karlsson and K. Ryan. A cost-value approach for prioritizing requirements. *IEEE Computer*, 14(5):67–74, 1997. doi: 10.1109/52.605933.

R. Keeney and H. Raiffa. *Decisions with multiple objectives: preferences and value tradeoffs*. Wiley, New York, 1976.

S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making*, 10:56, 2010. doi: 10.1186/1472-6947-10-56.

Y. M. Kong, C. Dahlke, Q. Xiang, Y. Qian, D. Karp, and R. H. Scheuermann. Toward an ontology-based framework for clinical research databases. *Journal of Biomedical Informatics*, 44(1):48–58, 2011. doi: 10.1016/j.jbi.2010.05.001.

J. Koponen. Agile Release Planning in a Product Backlog Tool. MSc thesis, 2008. URL `http://www.tml.tkk.fi/~anttiyj/Koponen-Agile.pdf`.

K. Krleza-Jeric, A.-W. Chan, K. Dickersin, I. Sim, J. Grimshaw, C. Gluud, and the Ottawa Group. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). *BMJ*, 330(7497):956–958, 2005. doi: 10.1136/bmj.330.7497.956.

R. Lahdelma and P. Salminen. SMAA-2: Stochastic multicriteria acceptability analysis for group decision making. *Operations Research*, 49(3):444–454, 2001. doi: 10.1287/opre.49.3.444.11220.

R. Lahdelma and P. Salminen. Classifying efficient alternatives in SMAA using cross confidence factors. *European Journal of Operational Research*, 170(1):228–240, 2006a. doi: 10.1016/j.ejor.2004.07.039.

R. Lahdelma and P. Salminen. Stochastic multicriteria acceptability analysis using the data envelopment model. *European Journal of Operational Research*, 170(1):241–252, 2006b. doi: 10.1016/j.ejor.2004.07.040.

R. Lahdelma, J. Hokkanen, and P. Salminen. SMAA - stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143, 1998. doi: 10.1016/S0377-2217(97)00163-X.

R. Lahdelma, P. Salminen, and J. Hokkanen. Locating a waste treatment facility by using stochastic multicriteria acceptability analysis with ordinal criteria. *European Journal of Operational Research*, 142(2):345–356, 2002. doi: 10.1016/S0377-2217(01)00303-4.

K. Lee, P. Bacchetti, and I. Sim. Publication of clinical trials supporting successful new drug applications: A literature analysis. *PLoS Medicine*, 5(9):e191, 09 2008. doi: 10.1371/journal.pmed.0050191.

S. Lewis and M. Clarke. Forest plots: trying to see the wood and the trees. *BMJ*, 322:1479–1480, 2001. doi: 10.1136/bmj.322.7300.1479.

C. Li, M. van den Akker, S. Brinkkemper, and G. Diepen. An integrated approach for requirement selection and scheduling in software release planning. *Requirements Engineering*, 15(4):375–396, 2010. doi: 10.1007/s00766-010-0104-x.

L. Liberti, A. Breckenridge, H. G. Eichler, R. Peterson, N. McAuslane, and S. Walker. Expediting patients' access to medicines by improving the predictability of drug development and the regulatory approval process. *Clinical Pharmacology and Therapeutics*, 87:27–31, Jan 2010. doi: 10.1038/clpt.2009.179.

D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Methods Inf Med*, 32:281–291, 1993.

R. Los, A. van Ginneken, and J. van der Lei. Extracting data recorded with OpenSDE: Possibilities and limitations. *Int J Med Inform*, 74:473–480, 2005.

L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999. doi: 10.1007/s101079900093.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006. doi: 10.1137/S009753970544727X.

G. Lu and A. E. Ades. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004. doi: 10.1002/sim.1875.

G. Lu and A. E. Ades. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006. doi: 10.1198/016214505000001302.

G. Lu and A. E. Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009. doi: 10.1093/biostatistics/kxp032.

G. Lu, A. E. Ades, A. J. Sutton, N. J. Cooper, A. H. Briggs, and D. M. Caldwell. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics in Medicine*, 26(20):3681–3699, 2007. doi: 10.1002/sim.2831.

G. Lu, N. J. Welton, J. P. T. Higgins, I. R. White, and A. E. Ades. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. *Research Synthesis Methods*, 2(1):43–60, 2011. doi: 10.1002/jrsm.34.

T. Lumley. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16):2313–2324, 2002. doi: 10.1002/sim.1201.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000. doi: 10.1023/A:1008929526011.

L. D. Lynd and B. J. O'Brien. Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. *Journal of Clinical Epidemiology*, 57(8):795–803, 2004. doi: 10.1016/j.jclinepi.2003.12.012.

R. Marks. Validating electronic source data in clinical trials. *Control Clin Trials*, 25(5):437–446, 2004. doi: 10.1016/j.cct.2004.07.001.

G. Marsaglia. Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics*, 43(2):645–646, 1972. doi: 10.1214/aoms/1177692644.

A. Martin, R. Biddle, and J. Noble. The XP customer role in practice: three studies. In *Agile Development Conference (ADC2004)*, Salt Lake City, Utah, USA, 2004. doi: 10.1109/ADEVC.2004.23.

A. T. McCray and N. C. Ide. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association*, 7(3):313–323, 2000. doi: 10.1136/jamia.2000.0070313.

M. McGregor and J. J. Caro. QALYs: are they helpful to decision makers? *Pharmacoeconomics*, 24(10):947–952, 2006.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144, 2008.

E. Miranda. Planning and executing time-bound projects. *Computer*, 35:73–79, 2002. doi: 10.1109/2.989933.

A. Mullard. 2010 fda drug approvals. *Nature Reviews Drug Discovery*, 10(2):82–85, 2011. doi: 10.1038/nrd3370.

C. D. Mulrow. Rationale for systematic reviews. *BMJ*, 309(6954):597–599, Sep 1994.

F. Mussen, S. Salek, and S. Walker. A quantitative approach to benefit-risk assessment of medicines – part 1: The development of a new model using multi-criteria decision analysis. *Pharmacoepidemiology and Drug Safety*, 16(Suppl. I):S12–S15, 2007. doi: 10.1002/pds.1435.

P. Nadkarni, C. Brandt, S. Frawley, F. Sayward, R. Einbinder, D. Zelterman, L. Schacter, and P. Miller. Managing attribute-value clinical trials data using the ACT/DB client-server database system. *Journal of the American Medical Informatics Association*, 5(2):139–151, 1998.

P. Nadkarni, C. Brandt, and L. Marenco. TrialDB: a clinical studies data management system, 2010. URL `http://ycmi.med.yale.edu/trialdb/`. Archived at `http://www.webcitation.org/5mm4G1K1S`.

P. M. Nadkarni and J. D. Darer. Determining correspondences between high-frequency MedDRA concepts and SNOMED: a case study. *BMC Medical Informatics and Decision Making*, 10: 66, 2010. doi: 10.1186/1472-6947-10-66.

NASA. *Manager's handbook for software development*. Software Engineering Laboratory. NASA Software Engineering Laboratory, Goddard Space Flight Center, Greenbelt, MD, 1990.

S. Nelson, M. Schopen, A. Savage, J.-L. Schulman, and N. Arluk. The MeSH translation maintenance system: Structure, interface design, and implementation. In *Proceedings of the 11th World Congress on Medical Informatics*, pages 67–69, San Francisco, 2004.

C. B. Nemeroff and M. E. Thase. A double-blind, placebo-controlled comparison of venlafaxine and fluoxetine treatment in depressed outpatients. *Journal of Psychiatric Research*, 41:351–359, 2007.

A. Ngo-The and G. Ruhe. A systematic approach for solving the wicked problem of software release planning. *Soft Computing*, 12(1):95–108, 2008. doi: 10.1007/s00500-007-0219-2.

A. Ngo-The and G. Ruhe. Optimized resource allocation for software release planning. *IEEE Transactions on Software Engineering*, 35(1):109–123, 2009. doi: 10.1109/TSE.2008.80.

S.-L. T. Normand. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999. doi: 10.1002/(SICI)1097-0258(19990215)18:3⟨321::AID-SIM28⟩3.0.CO;2-P.

A. Oliva. Janus update, 2009. URL `http://gforge.nci.nih.gov/docman/view.php/142/17571/Oliva%2020090310%20DIA%20Janus%20Update.pdf`. Presentation.

A. Oliveira and N. Salgado. Design aspects of a distributed clinical trials information system. *Clin Trials*, 3(4):385–396, 2006. doi: 10.1177/174077450609156.

Oracle Corp. Oracle buys phase forward, 2010. URL `http://www.oracle.com/us/corporate/press/068204`.

D. Ouellet. Benefit-risk assessment: the use of clinical utility index. *Expert Opinion on Drug Safety*, 9(2):289–300, 2010. doi: 10.1517/14740330903499265.

J. Paul, R. Seib, and T. Prescott. The internet and clinical trials: Background, online resources, examples and issues. *J Med Internet Res*, 7(1):e5, 2005. doi: 10.2196/jmir.7.1.e5.

Phase Forward, Inc. Clinical data repository, 2010. URL `http://www.phaseforward.com/products/cdc/cdr/`. Archived at `http://www.webcitation.org/5p8AxFMwO`.

L. D. Phillips. Decision conferencing. In W. Edwards, R. Miles Jr., and D. von Winterfeldt, editors, *Advances in Decision Analysis: from foundations to applications*. Cambridge University Press, 2007.

H. E. Pigott, A. M. Leventhal, G. S. Alter, and J. J. Boren. Efficacy and effectiveness of antidepressants: current status of research. *Psychotherapy and Psychosomatics*, 79:267–279, 2010. doi: 10.1159/000318293.

D. Pisinger. A minimal algorithm for the 0-1 knapsack problem. *Operations Research*, 45(5): 758–767, 1997. doi: 10.1287/opre.45.5.758.

M. Plummer. *JAGS Version 1.0.3 manual*, April 2009. URL `http://www-fis.iarc.fr/~martyn/software/jags`.

H. Prokosch and T. Ganslandt. Perspectives for medical informatics reusing the electronic medical record for clinical research. *Methods Inf Med*, 48(1):38–44, 2009. doi: 10.3414/ME9132.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

H. Rang, editor. *Drug Discovery and Development: Technology In Transition*. Churchill Livingstone, 2005.

I. Roberts, A. Po, and L. Chalmers. Intellectual property, drug licensing, freedom of information, and public health. *Lancet*, 352(9129):726–729, 1998.

B. Roy. *Multicriteria Methodology for Decision Analysis*. Kluwer Academic Publishers, Dordrecht, 1996.

D. L. Rubin, N. H. Shah, and N. F. Noy. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90, 2008. doi: 10.1093/bib/bbm059.

G. Ruhe and D. Greer. Quantitative studies in software release planning under risk and resource constraints. In *Proceedings of the 2003 International Symposium on Empirical Software Engineering (ISESE2003)*, pages 262–270, 2003. doi: 10.1109/ISESE.2003.1237987.

G. Ruhe and M. Saliu. The art and science of software release planning. *IEEE Software*, 22(6): 47–53, 2005. doi: 10.1109/MS.2005.164.

B. Rumpe and A. Schröder. Quantitative survey on extreme programming projects. In *Proceedings of the Third International Conference on Extreme Programming and Flexible Processes in Software Engineering*, pages 26–30, Sardinia, Italy, 2002.

D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1 1996. doi: 10.1136/bmj.312.7023.71.

G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3):279–301, 2008a. doi: 10.1177/0962280207080643.

G. Salanti, F. K. Kavvoura, and J. P. A. Ioannidis. Exploring the geometry of treatment networks. *Annals of Internal Medicine*, 148(7):544–553, 2008b.

G. Salanti, A. E. Ades, and J. P. A. Ioannidis. Graphical methods and numerical summaries for presenting results from multiple-treatments. *Journal of Clinical Epidemiology*, 64(2):163–171, 2011. doi: 10.1016/j.jclinepi.2010.03.016.

G. Salanti, V. Marinho, and J. P. T. Higgins. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of Clinical Epidemiology*, 62(8): 857–864, 2009. doi: 10.1016/j.jclinepi.2008.10.001.

M. O. Saliu and G. Ruhe. Bi-objective release planning for evolving software systems. In *Proceedings of the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, ESEC-FSE '07, pages 105–114, 2007. ISBN 978-1-59593-811-4. doi: 10.1145/1287624.1287641.

R. H. Scheuermann. Ontology-based extensible data model, 2010. URL `http://pathcuric1.swmed.edu/Research/scheuermann/OBX.html`. Archived at `http://www.webcitation.org/5y3tihqvB`.

K. F. Schulz, D. G. Altman, D. Moher, and for the CONSORT group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Medicine*, 7 (3):e1000251, 2010. doi: 10.1371/journal.pmed.1000251.

S. Schulz, S. Hanser, U. Hahn, and J. Rogers. The semantics of procedures and diseases in SNOMED CT. *Methods of Information in Medicine*, 45:354–358, 2006.

E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3–4):351–379, 1975. doi: 10.1016/0025-5564(75)90047-4.

I. Sim. *Trial Banks: An Informatics Foundation for Evidence-Based Medicine*. PhD thesis, Stanford University, 1997.

I. Sim. CTSA human studies database project. Abstract for All Hands Meeting, October 2008.

I. Sim and D. Detmer. Beyond trial registration: A global trial bank for clinical trial reporting. *PLoS Medicine*, 2(11):1090–1092, 2005. doi: 10.1371/journal.pmed.0020365.

I. Sim, D. K. Owens, P. W. Lavori, and G. D. Rennels. Electronic trial banks: A complementary method for reporting randomized trials. *Medical Decision Making*, 20(4):440–450, 2000. doi: 10.1177/0272989X0002000408.

I. Sim, B. Olasov, and S. Carini. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. *Journal of Biomedical Informatics*, 37(2):108–119, 2004. doi: 10.1016/j.jbi.2004.03.001.

I. Sim, A. W. Chan, A. M. Gulmezoglu, T. Evans, and T. Pang. Clinical trial registration: transparency is the watchword. *The Lancet*, 367(9523):1631–1633, 2006. doi: 10.1016/S0140-6736(06)68708-4.

I. Sim, C. G. Chute, H. Lehmann, R. Nagarajan, M. Nahm, and R. H. Scheuermann. Keeping raw data in context. *Science*, 323(5915):713a, 2009. doi: 10.1126/science.323.5915.713a.

I. Sim, S. Carini, S. Tu, R. Wynden, P. BH, S. Mollah, D. Gabriel, H. Hagler, R. Scheuermann, H. Lehmann, K. Wittkowski, M. Nahm, and S. Bakken. The human studies database project: Federating human studies design data using the ontology of clinical research. In *Proceedings of the AMIA CRI Summit 2010*, 2010.

R. J. Simes. Publication bias - the case for an international registry of clinical-trials. *Journal of Clinical Oncology*, 4(10):1529–1541, 1986.

R. L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984. doi: 10.1287/opre.32.6.1296.

S. Smith and A. Jain. Testing for uniformity in multidimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):73–81, 1984. doi: 10.1109/TPAMI.1984.4767477.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, version 1.4*, January 2003. URL `http://www.mrc-bsu.cam.ac.uk/bugs`.

J. Stapleton. *DSDM, dynamic systems development method: the method in practice*. Addison-Wesley Professional, 1997.

M. Starr and I. Chalmers. The evolution of The Cochrane Library, 1988–2003, 2003. URL `http://www.update-software.com/history/clibhist.htm`.

C. Stettler, S. Allemann, S. Wandel, A. Kastrati, M. C. Morice, A. Schoemig, M. E. Pfisterer, G. W. Stone, M. B. Leon, J. Suarez de Lezo, J.-J. Goy, S.-J. Park, M. Sabate, M. J. Suttorp, H. Kelbaek, C. Spaulding, M. Menichelli, P. Vermeersch, M. T. Dirksen, P. Cervinka, M. De Carlo, A. Erglis, T. Chechi, P. Ortolani, M. J. Schalij, P. Diem, B. Meier, S. Windecker, and P. Juni. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. *BMJ*, 337:a1331, 2008. doi: 10.1136/bmj.a1331.

J. G. Storosum, A. J. Elferink, B. J. van Zwieten, W. van den Brink, and J. Huyser. Natural course and placebo response in short-term, placebo-controlled studies in major depression: a meta-analysis of published and non-published studies. *Pharmacopsychiatry*, 37:32–36, Jan 2004. doi: 10.1055/s-2004-815472.

A. Sutton, A. E. Ades, N. Cooper, and K. Abrams. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics*, 26(9):753–767, 2008. doi: 10.2165/00019053-200826090-00006.

A. J. Sutton and J. P. T. Higgins. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008. doi: 10.1002/sim.2934.

A. J. Sutton, N. J. Cooper, and D. R. Jones. Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology*, 9(29):e–publication, 2009. doi: 10.1186/1471-2288-9-29.

M. Svahnberg, T. Gorschek, R. Feldt, R. Torkar, S. Bin Saleem, and M. U. Shafique. A systematic review on strategic release planning models. *Information and Software Technology*, 52(3):237–248, 2010. doi: 10.1016/j.infsof.2009.11.006.

M. Swertz, R. Oostergo, and B. de Brock. How to design a platform for the uniform integration, extraction, and querying of clinical and biobank data sets? *IEEE Trans Soft Eng*, (Submitted Manuscript), 2010.

R. Temple and S. S. Ellenberg. Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues. *Annals of Internal Medicine*, 133 (6):455–463, 2000.

T. Tervonen. JSMAA: an open source software for SMAA computations. In C. Henggeler Antunes, D. Rios Insua, and L. Dias, editors, *Proceedings of the 25th Mini-EURO conference on Uncertainty and Robustness in Planning and Decision Making (URPDM2010), Coimbra, Portugal*, 2010. URL `http://drugis.org/files/tervonen-urpdm2010.pdf`.

T. Tervonen and J. R. Figueira. A survey on stochastic multicriteria acceptability analysis methods. *Journal of Multi-Criteria Decision Analysis*, 15(1–2):1–14, 2008. doi: 10.1002/mcda.407.

T. Tervonen and R. Lahdelma. Implementing stochastic multicriteria acceptability analysis. *European Journal of Operational Research*, 178(2):500–513, 2007. doi: 10.1016/j.ejor.2005.12.037.

T. Tervonen, H. Hakonen, and R. Lahdelma. Elevator planning with Stochastic Multicriteria Acceptability Analysis. *Omega*, 36(3):352–362, 2008. doi: 10.1016/j.omega.2006.04.017.

T. Tervonen, J. R. Figueira, R. Lahdelma, J. Almeida Dias, and P. Salminen. A stochastic method for robustness analysis in sorting problems. *European Journal of Operational Research*, 192(1): 236–242, 2009a. doi: 10.1016/j.ejor.2007.09.008.

T. Tervonen, I. Linkov, J. Steevens, M. Chappell, J. R. Figueira, and M. Merad. Risk-based classification system of nanomaterials. *Journal of Nanoparticle Research*, 11(4):757–766, 2009b. doi: 10.1007/s11051-008-9546-1.

T. Tervonen, B. de Brock, P. A. de Graeff, and H. L. Hillege. Current status and future perspectives on drug information systems. In *Proceedings of the 18th European Conference on Information Systems, ECIS 2010*, Pretoria, South Africa, 2010.

T. Tervonen, G. van Valkenhoef, E. Buskens, H. L. Hillege, and D. Postmus. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428, 2011. doi: 10.1002/sim.4194.

T. Tervonen, G. van Valkenhoef, N. Baştürk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 2012. doi: 10.1016/j.ejor.2012.08.026. (in press).

The Cochrane Foundation. About the cochrane library, 2010. URL `http://www.thecochranelibrary.com/view/0/AboutTheCochraneLibrary.html`. Archived at `http://www.webcitation.org/5tEEbQuNz`.

The International Health Terminology Standards Development Organisation. SNOMED CT components, 2011. URL `http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-components/`. Archived at `http://www.webcitation.org/5yNmLpPkv`.

The Ministerial summit on health research. The Mexico statement on health research, 2004. URL `http://www.who.int/rpc/summit/agenda/en/mexico_statement_on_health_research.pdf`. Mexico City, November 16-20.

F. Thiers, A. Sinskey, and E. Berndt. Trends in the globalization of clinical trials. *Nature Reviews Drug Discovery*, 7(1):13–14, 2008.

T. Tse, R. J. Williams, and D. A. Zarin. Reporting "Basic Results" in ClinicalTrials.gov. *Chest*, 136(1):295–303, 2009. doi: 10.1378/chest.08-3022.

S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim. A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics*, 44(2):239–250, 2011. doi: 10.1016/j.jbi.2010.09.007.

Tufts, CSDD and CDISC. Study on the adoption and attitudes of electronic clinical research technology solutions and standards; summary of results, 2007. URL `http://www.cdisc.org/stuff/contentmgr/files/0/e35818eab8d8cc7b9d159ffeba5cdda5/misc/tuftstop30rdkjan08.pdf`.

M. van den Akker, S. Brinkkemper, G. Diepen, and J. Versendaal. Software product release planning through optimization and what-if analysis. *Information and Software Technology*, 50 (1-2):101–111, 2008. ISSN 0950-5849. doi: 10.1016/j.infsof.2007.10.017.

G. van Valkenhoef, T. Tervonen, E. O. de Brock, and D. Postmus. Product and release planning practices for extreme programming. In *Proceedings of the 11th International Conference on Agile Software Development (XP2010)*, Trondheim, Norway, 2010. doi: 10.1007/978-3-642-13054-0_25.

G. van Valkenhoef, T. Tervonen, B. de Brock, and D. Postmus. Quantitative release planning in extreme programming. *Information and Software Technology*, 53(11):1227–1235, 2011. doi: 10.1016/j.infsof.2011.05.007.

G. van Valkenhoef, G. Lu, B. de Brock, H. Hillege, A. E. Ades, and N. J. Welton. Automating network meta-analysis. *Research Synthesis Methods*, 2012a. doi: 10.1002/jrsm.1054. (in press).

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Clinical trials evidence in drug development and regulation: a survey of existing systems and standards. SOM Research Report 12003-Other, School of Management, Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands, 2012b. URL `http://irs.ub.rug.nl/dbi/4fcf224db9977`.

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Deficiencies in the transfer and availability of clinical evidence in drug development and regulation. *BMC Medical Informatics and Decision Making*, 2012c. doi: 10.1186/1472-6947-12-95. (in press).

G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Algorithmic parameterization of mixed treatment comparisons. *Statistics and Computing*, 22(5):1099–1111, 2012d. doi: 10.1007/s11222-011-9281-9.

G. van Valkenhoef, T. Tervonen, J. Zhao, B. de Brock, H. L. Hillege, and D. Postmus. Multi-criteria benefit-risk assessment using network meta-analysis. *Journal of Clinical Epidemiology*, 65(4):394–403, 2012e. doi: 10.1016/j.jclinepi.2011.09.005.

G. van Valkenhoef, T. Tervonen, T. Zwinkels, B. de Brock, and H. Hillege. ADDIS: a decision support system for evidence-based medicine. *Decision Support Systems*, 2012f. doi: 10.1016/j.dss.2012.10.005. (in press).

J.-C. Vansnick. On the problem of weights in multiple criteria decision making (the noncompensatory approach). *European Journal of Operational Research*, 24(2):288–294, 1986. doi: 10.1016/0377-2217(86)90051-2.

P. Vervuren. Visions of a drug development data warehouse. In *Proceedings of the First Conference of the Pharmaceutical Users Software Exchange (PhUSE 2005)*, 2005.

P. Vervuren and F. Dietvorst. Contours of a drug development data warehouse. In *Proceedings of the Second Conference of the Pharmaceutical Users Software Exchange (PhUSE 2006)*, 2006.

N. Victor and J. Hasford. Risk-benefit analyses of drugs: fundamental considerations and requirements from the point of view of the biometrician. Problems in the assessment of the combination of trimethoprim with sulfamethoxazole. *Infection*, 15:236–240, 1987. doi: 10.1007/BF01643196.

A. Vikstrom, Y. Skaner, L. E. Strender, and G. H. Nilsson. Mapping the categories of the swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Medical Informatics and Decision Making*, 7:9, 2007. doi: 10.1186/1472-6947-7-9.

A. Vitry. Comparative assessment of four drug interaction compendia. *British Journal of Clinical Pharmacology*, 63(6):709–714, 2006.

R. von Nitzch and M. Weber. The effect of attribute ranges on weights in multiattribute utility measurement. *Management Science*, 39:937–943, 1993.

B. T. Walsh, S. N. Seidman, R. Sysko, and M. Gould. Placebo response in studies of major depression. *Journal of the American Medical Association*, 287(14):1840–1847, 2002. doi: 10.1001/jama.287.14.1840.

N. J. Welton, D. M. Caldwell, E. Adamopoulos, and K. Vedhara. Mixed Treatment Comparison Meta-Analysis of Complex Interventions: Psychological Interventions in Coronary Heart Disease. *American Journal of Epidemiology*, 169(9):1158–1169, 2009. doi: 10.1093/aje/kwp014.

I. R. White. Multivariate random-effects meta-regression: Updates to mvmeta. *Stata Journal*, 11 (2):255–270, 2011.

A. J. J. Wood. Progress and deficiencies in the registration of clinical trials. *New England Journal of Medicine*, 360(8):824–830, 2009. doi: 10.1056/NEJMsr0806582.

W. Wood and W. Kleb. Exploring XP for scientific research. *IEEE Software*, 20(3):30–36, 2003. doi: 10.1109/MS.2003.1196317.

P. Wu, K. Wilson, P. Dimoulas, and E. J. Mills. Effectiveness of smoking cessation therapies: a systematic review and meta-analysis. *BMC Public Health*, 6:300, 2006. doi: 10.1186/1471-2458-6-300.

D. A. Zarin and T. Tse. Medicine - moving toward transparency of clinical trials. *Science*, 319 (5868):1340–1342, 2008. doi: 10.1126/science.1153632.

D. A. Zarin, N. C. Ide, T. Tse, W. R. Harlan, J. C. West, and D. A. B. Lindberg. Issues in the registration of clinical trials. *Journal of the American Medical Association*, 297(19):2112–2120, 2007. doi: 10.1001/jama.297.19.2112.

Making better use of clinical trials

Health care policy decision makers routinely evaluate the health impact of alternative treatment options. Here benefit-risk assessment is key, consisting of weighing the favorable effects (benefits) and unfavorable effects (risks) of the alternatives. For example, before a new drug is allowed on the market, regulators evaluate its benefit-risk balance in comparison to placebo or competing drugs.

Ideally, benefit-risk assessments are based on the best available evidence, typically meaning randomized controlled trials. However, finding the evidence and explicitly linking it to the assessment is complicated by several factors. First, the results of clinical trials are mainly made available in text-based documents that can not be processed automatically. Second, the data from these trials must be combined into a consistent basis for benefit-risk analysis. Third, quantitative decision models are required to directly link decisions to the underlying evidence and to make trade-off decisions explicit.

This thesis addresses these topics through the development of the Aggregate Data Drug Information System (ADDIS), an integrated system for decision support based on databases of structured clinical trials data. Novel algorithms are presented to automate network meta-analysis to combine clinical trials' results and multi-criteria decision models are developed to support benefit-risk assessment.

ADDIS is open source software, available from http://drugis.org/

Part of: TI PHARMA escher project