Statistical mechanics approaches to network meta-analysis and dynamic prediction with time-varying covariates

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2022

Annabel L Davies

School of Natural Sciences

Contents

Word count 81466

List of Tables

List of Figures

The University of Manchester

Statistical mechanics approaches to network meta-analysis and dynamic prediction with time-varying covariates

Annabel L Davies

Doctor of Philosophy

May 17, 2022

Abstract

Outside the domain of traditional physics, statistical mechanics provides a framework to describe the behaviour of systems from a diverse range of disciplines. In this thesis, we investigate several problems in medical statistics. In particular, we focus on the topics of network meta-analysis (NMA) and dynamic prediction. NMA is a technique for combining data from multiple medical trials that compare different combinations of treatment options. Dynamic prediction on the other hand, is a topic in survival analysis. Specifically, it refers to the process of making survival predictions based on the history of time-varying covariate measurements and updating prognosis as more observations are made.

In the first part of the thesis we present a statistical physics perspective on network meta-analysis. As well as introducing the technical details of the methodology for a physics audience, we compile existing analogies between the two fields, and discuss ideas for how statistical mechanics may be useful for NMA in the future.

A particular source of interest for statistical physicists lies in the representation of the treatments and trials as a network graph. In Chapters 3 and 4 we present two research projects on NMA that each stem from considerations of this graph.

First, we investigate the effect of network topology on NMA outcomes via a simulation study. The results of this study indicate that irregularity in the number of trials each treatment is involved in is negatively associated with the accuracy and precision of parameter estimates.

Second, we use the graph representation of NMA to introduce an analogy between NMA and random walks. The analogy provides insight into NMA methodology and leads to an analytical derivation of the so-called 'proportion contribution matrix' that overcomes limitations of previous algorithms used to construct this quantity.

In the next part of the thesis, we introduce the topic of survival analysis. Then, in Chapter 6 we develop an approach to dynamic prediction that, in comparison to standard methods, makes full use of the available data while remaining relatively parsimonious. In applications to clinical data sets, we find that our model performs similarly to standard approaches in terms of predictive accuracy.

The work in this thesis explores an interdisciplinary connection between statistical mechanics and medical statistics. I hope that this work is interesting for physicists and statisticians alike, and that it demonstrates that statistical physics ideas can make useful contributions to medical statistics.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University's Policy on Presentation of Theses.

Acknowledgements

In March 2020 I was in the second year of my PhD, on the brink of submitting my first paper when the Covid-19 pandemic forced the university to close its doors. As I lugged my computer screen and books back on the bus that day, I could not have known that the kitchen table I was heading back to would be the place from which I submitted my thesis almost exactly two years later.

Although I have spent the majority of my PhD working alone from my flat, there are a number of people who have made this effort possible. First, and most importantly, is my supervisor Tobias Galla from whom I have learnt so much and whose enthusiasm, expertise and dedication has been an inspiration. Thank you for your constant support, for challenging me, believing in me and for making this experience so rewarding.

I am extremely grateful to my collaborators Gerta Rücker, Ton Coolen, Adriani Nikolakopoulou and Thodoris Papakonstantinou for their guidance, encouragement, and for many stimulating and interesting discussions. I would also like to thank Niels Walet and Colin Timperley for their technical and administrative assistance, especially during the university closures. Thanks as well to the University of Manchester, EPSRC and the UoM PDS award for making this research possible.

A big thank you to my office mates: Joe B, Gabriel, Rory, Henry, Ernesto, Francisco, Kelsea and Laura who welcomed me to their group and made the first year and half of my PhD so much fun. Thanks especially to Joe B for his continued mentorship and for answering my many questions.

Finally, thank you to my friends and family for providing a loving support network. In particular; Holly for understanding and always checking in, and my parents and sister, Charlotte, for their unwavering belief and pride in me. The biggest thanks must go to my partner, Joe. Thank you for distracting me, feeding me, consoling me, understanding me and loving me. I couldn't have done this without you.

Dedication

This thesis is dedicated to my grandparents, three of whom saw me begin my PhD but none of whom have seen me finish it. First, to grandad who I never met but who I hope would be proud. To Granny, for being feisty and fabulous and for showing me how to live life to the full. To Grandma, for smothering me in love and care - and for all the cake and giggles. And to Grandpa; my biggest champion and best friend. I wish you were here to share this.

Chapter 1

Introduction

1.1 A brief overview of statistical mechanics

Most of the physical systems we observe in every day life are, at a microscopic level, comprised of many interacting components. A typical example is a gas which is made up of a large number of individual atoms or molecules. In thermodynamics, interest lies in the observable macroscopic properties of the gas such as its volume, pressure, and temperature. The thermodynamic laws governing the relations between these properties are phenomenological, obtained empirically from experiments on macroscopic systems [1]. Statistical mechanics was born out of the desire to derive the properties of a macroscopic system from the properties of its microscopic constituents [1, 2]. Essentially, this is done by averaging over unobservable microscopic coordinates [1]. For example, one can obtain the temperature of an ideal gas from the average kinetic energy of the atoms that make up that gas. In this way it is the statistical description of the individual components that tells us something about the macroscopic behaviour of the system as a whole.

In principle, one might imagine that the behaviour of a gas could be determined by modelling the movement of every individual particle in the system. Due to the vast number of particles, not only are the equations of motion for such a system far too complex to solve, the amount of information this would yield would be indigestible [2]. Even if we somehow had access to the position and momentum of every atomic component at each infinitesimal time step, to derive any meaningful understanding from this would require a statistical description of the information, e.g. the number

of collisions per second, the typical distance between atoms, or the number of atoms with velocities in a certain range [2]. Statistical mechanics provides a framework for obtaining this statistical description without the need to calculate the detailed individual behaviours. This is achieved by modelling the behaviour of the particles in a probabilistic manner.

The statistical description of particles provided by statistical mechanics explains a number of observable macroscopic behaviours that might otherwise appear in conflict with the microscopic laws governing the individual atoms. Perhaps most notably is the observation that macroscopic systems evolve in accordance with a specific 'arrow of time'. For example, a gas initially confined to half a container spreads out to fill the space, while a cup of tea left in a room eventually cools down. Without external interference, one does not observe the reverse behaviours. This phenomenon is encapsulated in the second law of thermodynamics which states that the entropy of a closed system never decreases. The thermodynamic arrow of time therefore points in the direction of increasing entropy. The apparent relevance of temporal direction conflicts with the fact that the physical laws governing the motion of individual particles are time reversible. For example, imagine that we could take a video of a collection of particles moving over time. One would then observe that the laws governing the motion of the particles in the video as it is played forward are the same as the laws of motion for the particles when the video is played in reverse.

This apparent conflict is reconciled by the idea that it is the collective, rather than the individual behaviour of the microscopic elements that determines the macroscopic properties of a system. In large systems, collective phenomena are governed by the probable behaviour of the interacting components. Therefore, in statistical mechanics we make probabilistic predictions for these systems. In a simple coin flip experiment, the proportion of times we obtain 'heads' from an unbiased coin approaches $1/2$ as the number of flips increases. In the same way, as the number of microscopic components of a system becomes large, the probabilistic description of their behaviour leads to accurate predictions of the macroscopic phenomena. Under this framework, the second law of thermodynamics can be understood from probabilistic arguments. Specifically, states of a system that lead to a decrease in entropy are less likely and become exceedingly improbable for very large systems. In fact, they are so improbable that we (almost)

never observe them.

The macroscopic properties of a system that arise from the collective behaviour of many individual components are said to be 'emergent'. These emergent phenomena are properties of the system as a whole. To use an example from Phillip Anderson [3, 4], a single atom of lead cannot be superconducting; instead, superconductivity is a property only of the macroscopic entity. Another example of emergent behaviour is the phase transition of a gas to a liquid or solid state. In this scenario the individual atoms themselves are neither a gas nor a solid. Rather, it is the collective behaviour of the atoms that leads to the emergent macroscopic properties of the solid, such as elasticity. Therefore, statistical mechanics involves modelling the emergent properties of the macroscopic system by considering the underlying statistical behaviour of the individual components and the forces governing their interactions.

In essence, the approach of statistical mechanics, and theoretical physics more generally, boils down to abstraction. We cannot know the exact reasons why an individual particle follows a particular trajectory but we can come up with a probabilistic model which accurately describes the overall system we are interested in.

1.2 Interdisciplinary applications of statistical mechanics

There is something almost fundamental in asking whether we can determine observable macroscopic properties from the stochastic behaviour of individual components. Indeed, the techniques developed in statistical mechanics are applicable to almost all physical systems. In this sense, statistical mechanics naturally lends itself to interdisciplinary work [5]. For example, in biology, statistical mechanics has made significant contributions to topics such as pattern formation [6], neural networks [7], and anti-microbial resistance [8]. In economics, properties of financial markets can be modelled via the statistical dynamics of individual stock prices [9]. There are even applications to sociology, where interaction based models of individual behaviour can explain social phenomena such as voting patterns, language formation and crowd dynamics [10–12].

The work in this thesis relates to an application of statistical mechanics that has seen much less exploration; namely, the field of medical statistics. While it is not

uncommon for researchers with a background in physics to move into the world of data analysis and statistical modelling, an interdisciplinary connection between the fields is not well established.

In medicine, the health outcome for an individual is subject to some level of effective (or 'epistemic') randomness. Anyone visiting a doctor's surgery will be familiar with this fact; prognosis is commonly discussed in terms of probability, chance and risk. In the most basic sense, medical statistics aims to extract some underlying 'truth' from observations of the health outcomes of many individuals. This process involves averaging over noisy individual observations to make statements about the population as a whole, such as the relative efficacy of medical interventions or the effect of certain variables on survival. There is a level of abstraction here. The models used to analyse this data cannot encapsulate the true complex reality of the forces governing any one person's medical experience. Instead, we must rely on simpler, more abstract models based on probabilistic ideas.

As in statistical mechanics, it is this abstraction that can tell us something interesting about reality. Indeed, it was statistician George Box [13] who is credited with the famous quote *"all models are wrong, but some are useful"*, a philosophy that holds relevance in many scientific disciplines. In fact, when elaborating on this idea in a contribution to a book entitled 'Robustness of Statistics' [13], Box makes reference to thermodynamics,

> *"...cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules."*

Evidently then, there is some overlap between statistical analysis of data and the statistical physicist's approach to understanding physical systems.

The applications of statistical mechanics to the fields of biology, finance and sociology have developed over time with input from numerous researchers across the disciplinary divides. These interdisciplinary approaches have become a fundamental aspect of their respective fields, with entire journals and research facilities dedicated to their study.

I believe that medical statistics and statistical physics have the potential to benefit similarly from an exchange of ideas. The work described here takes some initial steps towards this aim. In particular, this thesis aims to explore if a physicist's approach to research and techniques from statistical mechanics can contribute in a meaningful way to medical statistics methodology.

1.3 Medical statistics topics in this thesis

The projects presented in this thesis pertain to two medical statistics methodologies; network meta-analysis (NMA), and dynamic prediction. Network meta-analysis is a technique involved in so-called 'evidence synthesis', while dynamic prediction falls under the topic of 'survival analysis'.

1.3.1 Evidence synthesis

Evidence synthesis refers to the process of bringing together all relevant information on a specific research question. By collating the existing evidence base, evidence synthesis methods summarise what is known about a particular topic. In the context of medical research, this type of analysis promotes well-informed healthcare policy and clinical decision making [14].

A 'systematic review' is a form of evidence synthesis that uses transparent, repeatable methods to identify, evaluate and summarise the findings of all individual studies relevant to the specific topic [15]. The review follows a strict pre-specified protocol based on a well-defined research question. The process includes identification and appraisal of eligible studies, data extraction, and data synthesis [14]. Motivated by the work of Archie Cochrane in the 1970s and 80s [16, 17], systematic reviews in medicine have seen rapid uptake and development over recent years [14, 18].

In the final step of a systematic review, one aims to provide a summary of the data collated from the individual studies. The statistical synthesis of the relevant numerical results is usually completed using a so-called 'meta-analysis'.

Meta-analysis. A meta-analysis is a quantitative method for combining the results of multiple scientific studies that answer the same research question [19, 20]. In this thesis, we focus on the meta-analysis of medical trials.

In medicine, a randomised controlled trial is used to compare a set of interventions for the treatment of a particular medical condition. Participants in the trial are sampled from the population of patients with the condition and each participant is randomly allocated one of the treatment options. By comparing the subsequent health outcomes between groups of patients assigned to the different treatment options, one can assess which interventions are the most effective.

Clinical trials often involve small sample sizes drawn from a subsection of the population of interest. By synthesising the results of all trials comparing the same treatment interventions, meta-analysis provides a more precise estimate of their relative effect compared with the results from individual trials [19–21].

A standard (pairwise) meta-analysis is performed on a set of trials that each compare the same two treatment options. This provides an overall statistic that summarises the effectiveness of one treatment compared to the other [22]. Typically, the effect of an experimental treatment is compared with a control treatment such as a placebo or a standard of care.

Network meta-analysis. For any given medical condition, there usually exists more than two possible treatments. There is then interest in comparing the full set of treatment options in order to work out which is the most effective. A naïve approach is to perform a separate meta-analysis for each pairwise combination of treatments. However, this is rarely possible as, in general, not all pairs of treatments will have been compared in a head-to-head trial. Furthermore, this naïve method ignores indirect information arising from comparisons to some common third treatment. In response to this, network meta-analysis (NMA) emerged as a technique to combine so-called direct and indirect evidence (concepts which we will explain in the next paragraph) from trials comparing different combinations of treatment options [23].

Indirect evidence refers to the idea that when two treatments, A and B, have not been compared in any head-to-head trials, one can infer their relative effect using information from trials in which these treatments have been compared to some 'common comparator' treatment C (see Figure 1.1) [24]. Essentially, if we know that treatment A is more effective than treatment C (from trials comparing A to C) and that C is more effective than B (from trials comparing B to C) then we can infer that A is more effective than B. Direct evidence on the comparison of A and B refers to trials in which

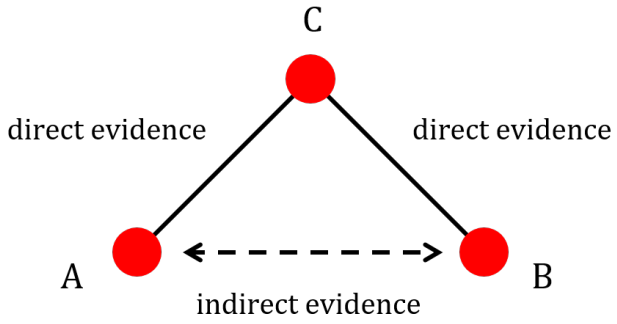


Figure 1.1: An illustration of indirect evidence. Vertices represent different treatment options (A, B, C) and edges represent comparisons between treatments in trials.

these two treatments are compared head-to-head.

Higgins and Whitehead (1996) [25] assumed that both direct and indirect evidence provide estimates of the same parameter (i.e. the relative treatment effect). Based on this idea they proposed models for combining evidence from trials comparing different combinations of treatments A, B and C, including trials that compare all three treatments. In another seminal paper, Lumley (2002) [26] considered that indirect evidence about A and B could arise with reference to various common comparators (e.g. treatments C, D, E...). Combining these indirect estimates, along with any available direct evidence, then yields an overall estimate for the relative effect of A and B based on all the available data. Building on this work, Lu and Ades (2004, 2006) [27, 28] proposed different parameterisations of the Higgins and Whitehead model that allowed for the simultaneous evaluation of the relative effects between all pairs of treatments in a wide range of scenarios. Since then, numerous other formulations have also been developed (e.g. [29–35]).

The term 'network meta-analysis' was proposed by Lumley [26] who used a network graph to visualise the various indirect comparisons between A and B. In this visualisation, vertices represent treatment options, and edges connecting the vertices indicate comparisons between treatments in the trials.

The representation of treatments and trials as a network graph has natural appeal to statistical physicists. Physical systems can be modelled by networks that encapsulate structural patterns of connection between the interacting components of the system. Statistical mechanics then provides a framework for modelling dynamic processes on networks [36, 37], and for describing the structural properties, time evolution and

statistical features of the networks themselves [38, 39]. Our work on network meta-analysis focuses on the network properties of the collection of treatments and trials. Chapter 3 investigates the effect of network topology on the outcome of the analysis while Chapter 4 considers the properties of a random walker moving on the network of evidence.

1.3.2 Survival analysis

In short, survival analysis involves studying the expected time until the occurrence of an event [40]. While applications to a range of disciplines exist (for example in economics [41], engineering [42] and sociology [43]), the work in this thesis focuses on the survival of patients in a medical trial.

Individuals in the trial are typically characterised by a set of measurable quantities called covariates. Each individual is then observed over time until they experience some irreversible medical event such as death or the onset of disease. Patients who do not experience an event by the end of the observation period are said to be censored [44].

The principal idea is to model the probability of survival as a function of time by extracting the relationship between patient covariates and survival based on these time-to-event measurements. One can then predict the probability that a new patient, for whom we have a set of covariate measurements, survives to some specified future time. This analysis has particular clinical relevance as it provides a framework with which to quantify and communicate personalised patient prognosis.

In the simplest case covariates are treated as constants and their values are measured at the baseline time, i.e. at study entry. Using a parameterised or semi-parameterised model, one then estimates the regression coefficients representing the association between these covariates and the probability of survival. In reality however, one often has access to repeated measurements of time-varying covariates from patients attending follow-up appointments. The task is then to use the longitudinal trajectory of the covariate measurements to specify a survival model and to make predictions. In particular, clinical interest lies in predicting patient survival probabilities based on the full history of their covariate measurements and updating prognosis as each new observation is made. This process is known as 'dynamic prediction' [45–47].

The time until an event occurs represents a stochastic process. In survival analysis, the stochasticity is accounted for using a probabilistic model and the process is parameterised by the covariates. It is this parameterisation that we attempt to extract from the data. The stochastic nature of this process lends itself to statistical mechanics. Indeed, concepts such as first hitting time models, that have their origin in the study of statistical physics, have previously been used to model survival [48].

In dynamic prediction one aims to make predictions about the time to an event in a stochastic process based on observations that evolve over time. From a statistical mechanics perspective, these processes offer an interesting topic for exploration. In Chapter 6 we propose an approach to dynamic prediction that builds on standard survival analysis models.

1.4 Thesis structure and format

This thesis is presented in 'journal format' in accordance with The University of Manchester guidelines. Chapters 2, 3, 4, and 6 contain the content of manuscripts that have either been published or have been submitted for publication. Each of these chapters is prefaced with a reference to the relevant journal article or pre-print, along with the author list, and a description of the role each author played. A list of works pertaining to these chapters is given in Section 1.5 and is ordered chronologically (different from the ordering of the chapters).

The presentation of the manuscripts in Chapters 2, 3, 4, and 6 has been modified to fit the requirements of The University of Manchester's thesis format. Any typographical errors that were spotted after submission have been corrected. Each of these chapters uses the notation from the original manuscript meaning there are some small notational differences between chapters. However, the notation used in each chapter is introduced in full and is self-consistent within the chapter. We now briefly summarise the content of each chapter.

Chapter 2: Network Meta-Analysis: A Statistical Physics Perspective. This chapter provides a technical background to network meta-analysis. Here, we introduce the necessary mathematical details for the projects discussed in Chapters 3 and 4. This includes the main NMA models and their assumptions, Bayesian and

frequentist approaches to inference, details of relevant numerical algorithms, and methods for reporting NMA results and ranking treatments. The original manuscript to which this chapter relates had the aim of introducing network meta-analysis to statistical physicists. Therefore, the chapter also details existing points of contact between meta-analysis and topics from physics, and discusses ideas for how statistical physics might contribute to NMA in the future.

Chapter 3: Degree irregularity and rank probability bias in network meta-analysis. In this chapter we present a simulation study investigating the effect of network topology on the outcomes of an NMA. In particular, we characterise network topology in terms of asymmetry in the distribution of trials between the treatments. We find that this asymmetry is associated with variation in the precision of treatment effect estimates and a systematic bias in estimates of rank probabilities. We discuss how these findings may be used to inform the planning future trials.

Chapter 4: Network meta-analysis and random walks. In statistical mechanics, random walks are a popular tool for analysing networks. A random walk on a graph is a stochastic process that describes a path made up of a succession of random 'hops' between vertices that are connected by an edge. In this chapter we demonstrate a novel analogy between random walks and NMA. Using the existing interdisciplinary analogies between (i) NMA and electrical networks [29], and (ii) electrical networks and random walks [49], we construct a random walk on the meta-analytic network. By analysing the average movement of the random walker, we obtain information about the flow of evidence through the network, and the influence of certain comparisons on the network estimates. The analogy leads to an analytical expression for the so-called 'proportion contribution matrix' which overcomes the limitations of previous algorithms used to construct this quantity.

Chapter 5: Introduction to Survival Analysis. This chapter provides a bridge between the projects on NMA and the final project in Chapter 6 on the topic of survival analysis. We give an overview of some of the main topics in survival analysis and introduce some technical concepts relevant for Chapter 6 such as the survival function, hazard rates, censoring mechanisms, and the Cox proportional hazards model.

Chapter 6: Retarded kernels for longitudinal survival analysis and dynamic prediction. Here we develop a new approach to dynamic prediction with

time-varying covariates. We assume that the effect of a covariate change at a certain time decays exponentially according to some characteristic (covariate-specific) time scale. Based on this, and requiring that our models contain standard models as a special case, we specify two time-varying association kernels that assign less weight to measurements made further in the past and greater weight to more recent measurements. We compare our models to two standard approaches via application to three clinical data sets.

Chapter 7: Conclusions. In conclusion, we review the findings from Chapters 3, 4, and 6. We present some final remarks on the thesis as a whole including the wider contributions of the work, and potential avenues for future research.

Chapter 8: Supplementary material for 'Degree irregularity and rank probability bias in network meta-analysis'. This chapter includes supplementary simulations and figures for Chapter 3. We separate these from the rest of the thesis so as not to interrupt the flow of the text.

1.5 List of works

- A. L. Davies and T. Galla, "Degree irregularity and rank probability bias in network meta-analysis", Research Synthesis Methods **12**, 316-332 (2021). 10.1002/jrsm.1454

- A. L. Davies, T. Papakonstantinou, A. Nikolakopoulou, G. Rücker and T. Galla, "Network meta-analysis and random walks", Statistics in Medicine, 1-24 (2022). 10.1002/sim.9346

- A. L. Davies, A. C. C. Coolen and T. Galla, "Retarded kernels for longitudinal survival analysis and dynamic prediction", *arXiv preprint.* arXiv:2110.11196 (2021). [**Submitted to Statistical Methods in Medical Research**]

- A. L. Davies and T. Galla, "Network Meta-Analysis: A Statistical Physics Perspective", *arXiv preprint.* arXiv:2203.11741 (2022). [**Submitted to Journal of Statistical Mechanics: Theory and Experiment**]

1.6 Contributions to software

- G. Rücker, U. Krahn, J. König, O. Efthimiou, A. L. Davies, T. Papakonstantinou and G. Schwarzer. "netmeta: Network Meta-Analysis using Frequentist Methods", *R Foundation for Statistical Computing* (2021). R package version 2.0-0. CRAN.R-project.org/package=netmeta

Bibliography

[1] F. Mandl, *Statistical physics*, 2nd ed., Manchester Physics Series (Wiley, Chichester, UK, 1988).

[2] N. Davidson, *Statistical mechanics* (Dover Publications Inc, Mineola, NY, USA, 2003).

[3] P. W. Anderson, *More and different: notes from a thoughtful curmudgeon* (World Scientific Publishing Company, Singapore, 2011).

[4] P. W. Anderson, "More is different", Science **177**, 393–396 (1972).

[5] D. Stauffer, "Introduction to statistical physics outside physics", Physica A **336**, Proceedings of the XVIII Max Born Symposium "Statistical Physics Outside Physics", 1–5 (2004).

[6] P. K. Maini, T. E. Woolley, R. E. Baker, E. A. Gaffney, and S. S. Lee, "Turing's model for biological pattern formation and the robustness problem", Interface focus **2**, 487–496 (2012).

[7] H. Sompolinsky, "Statistical mechanics of neural networks", Phys. Today **41**, 70–80 (1988).

[8] R. Allen and B. Waclaw, "Antibiotic resistance: a physicist's view", Phys. Biol. **13**, 045001 (2016).

[9] R. N. Mantegna, Z. Palágyi, and H. E. Stanley, "Applications of statistical mechanics to finance", Physica A **274**, 216–221 (1999).

[10] S. N. Durlauf, "How can statistical mechanics contribute to social science?", Proc. Natl. Acad. Sci. USA **96**, 10582–10584 (1999).

[11] D. B. Bahr and E. Passerini, "Statistical mechanics of collective behavior: macrosociology", J. Math. Sociol. **23**, 29–49 (1998).

[12] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics", Rev. Mod. Phys. **81**, 591–646 (2009).

[13] G. E. P. Box, "Robustness in the strategy of scientific model building", in *Robustness in statistics* (Academic Press, New York, NY, USA, 1979), pp. 201–236.

[14] M. Clarke, "History of evidence synthesis to assess treatment effects: personal reflections on something that is very much alive", J. Roy. Soc. Med. **109**, 154–163 (2016).

[15] Centre for Reviews and Dissemination, *Systematic reviews: CRD's guidance for undertaking reviews in healthcare*, 3rd ed. (CRD, University of York, York, UK, 2009).

[16] A. L. Cochrane, "1931-1971: a critical review, with particular reference to the medical profession", in *Medicine for the year 2000*, edited by G. Teeling-Smith and N. Wells (Office of Health Economics, London, UK, 1979), pp. 1–11.

[17] A. L. Cochrane, "Foreword", in *Effective care in pregnancy and childbirth*, edited by I. Chalmers, M. Enkin, and M. J. N. C. Keirse (Oxford University Press, Oxford, UK, 1989).

[18] A. B. Haidich, "Meta-analysis in medical research", Hippokratia **14**, 29–37 (2010).

[19] T. C. Smith, D. J. Spiegelhalter, and A. Thomas, "Bayesian approaches to random effects meta analysis: a comparative study", Stat. Med. **14**, 2685–2699 (1995).

[20] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials", Control. Clin. Trials **7**, 177–188 (1986).

[21] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, eds., *Cochrane handbook for systematic reviews of interventions*, 2nd ed. (Wiley, Chichester, UK, 2019).

[22] J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter, "A re-evaluation of random effects meta-analysis", J. R. Stat. Soc. **172**, 137–159 (2009).

[23] S. Dias, A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton, *Network meta-analysis for decision making* (Wiley, Oxford, UK, 2018).

[24] H. C. Bucher, G. H. Guyatt, L. E. Griffith, and S. D. Walter, "The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials", J. Clin. Epidemiol. **50**, 683–691 (1997).

[25] J. P. T. Higgins and A. Whitehead, "Borrowing strength from external trials in a meta-analysis", Stat. Med. **15**, 2733–2749 (1996).

[26] T. Lumley, "Network meta-analysis for indirect treatment comparisons", Stat. Med. **21**, 2313–2324 (2002).

[27] G. Lu and A. E. Ades, "Combination of direct and indirect evidence in mixed treatment comparisons", Stat. Med. **23**, 3105–3124 (2004).

[28] G. Lu and A. E. Ades, "Assessing evidence inconsistency in mixed treatment comparisons", J. Am. Stat. Assoc. **101**, 447–459 (2006).

[29] G. Rücker, "Network meta-analysis, electrical networks and graph theory", Res. Synth. Meth. **3**, 312–324 (2012).

[30] O. Efthimiou, G. Rücker, G. Schwarzer, J. P. T. Higgins, M. Egger, and G. Salanti, "Network meta-analysis of rare events using the Mantel-Haenszel method", Stat. Med. **38**, 2992–3012 (2019).

[31] T. Stijnen, T. H. Hamza, and P. Özdemir, "Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data", Stat. Med. **29**, 3046–3067 (2010).

[32] S. Dias, A. J. Sutton, A. E. Ades, and N. J. Welton, "Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials", Med. Decis. Making. **33**, 607–617 (2013).

[33] G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis, "Evaluation of networks of randomized trials", Stat. Methods. Med. Res. **17**, 279–301 (2008).

[34] R. D. Riley, D. Jackson, G. Salanti, D. L. Burke, M. Price, J. Kirkham, and I. R. White, "Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples", BMJ **358**, j3932 (2017).

[35] G. Salanti, "Indirect and mixed treatment comparison, network, or multiple treatments meta analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool", Res. Synth. Meth. **3**, 80–97 (2012).

[36] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press, Cambridge, UK, 2013).

[37] T. Gross and B. Blasius, "Adaptive coevolutionary networks: a review", J. R. Soc. Interface **5**, 259–271 (2008).

[38] J. Park and M. E. J. Newman, "Statistical mechanics of networks", Phys. Rev. E **70**, 066117 (2004).

[39] B. A. Desmarais and S. J. Cranmer, "Statistical mechanics of networks: estimation and uncertainty", Physica A **391**, 1865–1876 (2012).

[40] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part I: basic concepts and first analyses", Brit. J. Cancer **89**, 232–238 (2003).

[41] J. J. Heckman and B. Singer, "Econometric duration analysis", J. Econometrics **24**, 63–132 (1984).

[42] I. A. Ushakov, ed., *Handbook of reliability engineering* (John Wiley & Sons, New York, NY, USA, 1994).

[43] H.-P. Blossfeld, A. Hamerle, and K. U. Mayer, *Event history analysis: statistical theory and application in the social sciences*, 1st ed. (Psychology Press, Taylor & Francis Group, New York, NY, USA, 1989).

[44] D. G. Kleinbaum and M. Klein, *Survival analysis. A self-learning text* (Springer-Verlang, New York, NY, USA, 2005).

[45] H. C. Van Houwelingen, "Dynamic prediction by landmarking in event history analysis", Scand. J. Stat. **34**, 70–85 (2007).

[46] D. Rizopoulos, "Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data", Biometrics **67**, 819–829 (2011).

[47] H. Van Houwelingen and H. Putter, *Dynamic prediction in clinical survival analysis* (Taylor & Francis Group, Boca Raton, FL, USA, 2011).

[48] M.-L. T. Lee and G. A. Whitmore, "Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary", Stat. Sci. **21**, 501–513 (2006).

[49] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, arXiv:math/0001057, 2000.

Chapter 2

Network Meta-Analysis: A Statistical Physics Perspective

Preface

The contents of this chapter constitute a manuscript submitted to the Journal of Statistical Mechanics: Theory and Experiment. The preprint edition is available on ArXiv[1]. This is a perspective review article and serves as a technical introduction to network meta-analysis. The manuscript was authored by Annabel L Davies[2] and Tobias Galla[2,3].

ALD wrote the bulk of the first draft of the manuscript except for Sections 2.7.3-2.7.7 which were primarily written by TG. Both ALD and TG contributed to discussions guiding the work and edited the manuscript. ALD produced all of the figures except Figure 2.4 which was created by TG.

[1]A. L. Davies and T. Galla, "Network Meta-Analysis: A Statistical Physics Perspective", *arXiv preprint.* arXiv:2203.11741 (2022).

[2]Theoretical Physics, School of Physics and Astronomy, The University of Manchester, Manchester, M13 9PL, United Kingdom.

[3]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

Abstract

Network meta-analysis (NMA) is a technique used in medical statistics to combine evidence from multiple medical trials. NMA defines an inference and information processing problem on a network of treatment options and trials connecting the treatments. We believe that statistical physics can offer useful ideas and tools for this area, including from the theory of complex networks, stochastic modelling and simulation techniques. The lack of a unique source that would allow physicists to learn about NMA effectively is a barrier to this. In this article we aim to present the 'NMA problem' and existing approaches to it coherently and in a language accessible to statistical physicists. We also summarise existing points of contact between statistical physics and NMA, and describe our ideas of how physics might make a difference for NMA in the future. The overall goal of the article is to attract physicists to this interesting, timely and worthwhile field of research.

2.1 Introduction

Physicists in general, and statistical physicists in particular, have a propensity to draw inspirations from problems across the borders of traditional disciplines. The application of ideas and methods from physics to questions in biology, economics and the social sciences is therefore well established [1, 2]. The following quote by the late Dietrich Stauffer encapsulates this [3]: *'The basic theorem of interdisciplinary research states: Physicists not only know everything; they know everything better.'*[1] Arguably, not all of these invasions into the territory of other disciplines are useful, and physicists have been criticised for their, at times, ill-informed attempts to address questions outside their area of expertise [4]. On the other hand, it is also hard to deny that physics approaches have made useful contributions to a number of different fields.

In this perspective review we highlight network meta-analysis (NMA), a topic from medical statistics, as a field for which we think physics ideas might be useful. Meta-analysis is a statistical technique used to combine the results of multiple trials [5–7].

[1]The quote continues: *'This theorem is wrong; it is valid only for computational statistical physicists like me'.*

The aim of such trials is to establish and compare how effective different treatment options are. To do this, the different treatments are administered to groups of subjects in medical trials. Individual trials often have small sample sizes and involve subjects taken from a reduced population. Because of this, it is desirable to systematically integrate results from different trials to obtain an overall estimate of the effect of a given treatment and to compare treatment options. This is complicated by the fact that trials taking place at different locations will generally involve demographically different subject groups. The aggregation of data from different trials is not straightforward.

Conventional meta-analysis focuses on pairwise comparisons of treatments. More recently however, NMA (also referred to as 'indirect meta-analysis', and 'multiple' or 'mixed treatment comparison' [8, 9]) has emerged as a technique for making inferences about multiple competing treatments. NMA allows one to combine data from multiple trials even when different trials test different sets of treatment options. The term 'network meta-analysis' derives from a graphical representation of the treatments and trials. The nodes of the graph are the different treatment options and the connecting edges represent comparisons made between the treatments, an illustration can be found in Figure 2.1. NMA combines direct and indirect evidence for the assessment of treatments. This makes it possible to compare treatments that have not been tested together in any trial. For a textbook on NMA see [9].

NMA is based on two main concepts: microscopic models for the outcomes of the different trials in the graph, and algorithms or procedures to carry out the actual NMA inference.

Microscopic models. The microscopic model captures the main assumptions made on the process leading to real-world trial outcomes. Each trial tests a subset of treatments. In the so-called 'random effects model' the relative treatment effects of the treatments in a particular trial are drawn from an underlying distribution. As a consequence, the effect of one treatment relative to another is not necessarily the same in two different trials. This reflects variability in local characteristics, for example, the fact that patient groups are chosen from different demographic subsets at different locations. In simulations of the random effects model, rates of success and failure for each trial arm are then constructed (the treatments tested in a trial are referred to as the 'arms' of the trial). A treatment success or failure occurs for a particular patient with the

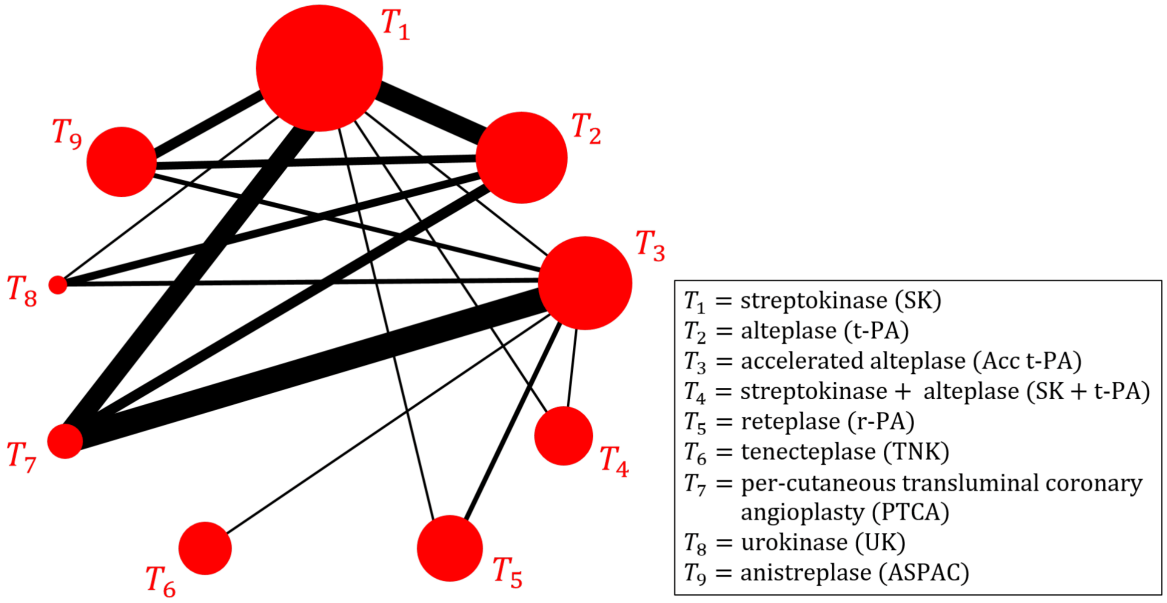


Figure 2.1: The network for the 'thrombolytic drug data' set [10–12] comparing nine treatments for acute myocardial infarction (heart attack). The treatments $T_1, \ldots, T_9$ are labelled in the box. They consist of eight thrombolytic drugs and one angioplasty intervention (T_7). The thickness of the edges in the network indicate the number of trials making that comparison. The area of the node is proportional to the number of patients allocated to that treatment. The network consists of 50 trials; two 3-arm trials (comparing T_1, T_3, T_4 and T_1, T_2, T_9, respectively) and 48 2-arm trials. The multi-arm trials are not explicitly indicated on the graph.

corresponding probability. This generates multiple layers of randomness in simulations: random treatment effects, and a binomial distribution of successes and failures for each trial arm.

NMA inference. The purpose of NMA is to estimate model parameters from given trial outcomes. These can either be real-world data or synthetic data (i.e. data generated in simulation studies used for methodological evaluation). The NMA process also provides confidence levels for these parameters. These can be used to construct a 'ranking' of treatments, as best, second best and so on. In more sophisticated approaches, probabilities are assigned that each treatment has a particular rank, reflecting the uncertainty on inferred treatment effects. Different ranking methods are still very much under discussion.

The NMA inference itself can either be carried out in a frequentist or a Bayesian setting. In this paper we will describe both approaches. In Bayesian NMA prior distributions are assumed for key model parameters, and posterior distributions are constructed from these and the trial outcomes. This needs to be done numerically, using

Markov Chain Monte Carlo methods (in NMA specifically, the Metropolis-in-Gibbs algorithm is often used [13]). In the frequentist approach described here, one defines a linear regression model dependent on the model parameters. The model can then be fitted using generalised least squares regression or maximum likelihood. For continuous outcomes under the assumption of normally distributed random errors these methods are equivalent. Other procedures are also possible [14, 15], but are not discussed here.

We believe that NMA has a natural appeal to statistical physicists. Those with experience in complex networks will find it interesting to connect the structure of treatment-trial networks with the outcome of NMA. Computational physicists may contribute to optimising the inference process and required sampling methods. Those interested in stochastic simulations can naturally connect with data generation methods used to obtain synthetic data for a given network of treatments and trials. NMA is an inference problem on a networked structure, and we expect that physicists working at the border to computer science and machine learning will become excited about it; for example it is conceivable that message passing methods can become a useful tool for NMA. Our own work (with collaborators) shows that random walks on the meta-analytic network and related graphs can lead to additional insights and improve methods to establish how evidence flows in NMA [16].

One main bottleneck appears to be that there is no unique source which would allow a physicist to enter this field efficiently. While textbooks and review articles on NMA exist [8, 9, 17–24], these are often written for medical practitioners, or users of existing software packages. The mathematical details are frequently suppressed, or not presented in a language physicists are used to. This can make it hard to get a good grip on the actual mechanics of NMA. This perspective review is our attempt at rectifying this. Our objective is to provide a technical introduction to NMA, accessible to physicists. We have aimed to make this self-contained, but at the same time this review is not a textbook and we have tried to keep the length to a reasonable limit. We hope we have found a sensible middle ground. We necessarily had to make a selection of topics we can cover, and attempted to choose those that are most helpful for others entering this area. We also aim to point out ideas from physics which we believe to be most promising to make a difference to NMA. We hope that this will facilitate future work by the physics community in this timely and worthwhile area of research.

The paper is organised as follows: Section 2.2 sets the scene, defines the necessary notation and states the 'NMA problem'. In Sections 2.3 and 2.4 we then present the mathematical procedures used to carry out an NMA in a Bayesian and frequentist setting respectively. Section 2.5 summarises how the results of an NMA are reported. In Section 2.6 we present existing analogies connecting NMA to different systems in physics, including resistor networks and random walks. In Section 2.7 we then outline some more general connections between the two fields and speculate on ways in which we think physicists may contribute to NMA in the future. Section 2.8 contains a brief summary and discussion.

2.2 Networks of medical trials

2.2.1 General background: randomised controlled trials, meta-analysis, and network meta-analysis

In this section we first give an informal description of the key concepts in NMA. We turn to a more formal mathematical setup in Section 2.2.4.

2.2.1.1 Randomised controlled trials

For our purposes a trial is an experiment in which a group of subjects is used to compare a given set of treatment options. The different treatments are referred to as the *arms* of the trial. In particular, a 'controlled' clinical trial is one with at least two arms. Typically, this involves one or more 'experimental' treatment groups representing new treatments being tested. These are compared to the so-called 'control' group(s) which could be alternative (existing) treatments, a placebo or no treatment [25].

The allocation of subjects to the different arms is randomised to avoid any bias in treatment assignment. For example, the treatment assigned to a given subject can be chosen with equal probability from the arms of the trial. In this scenario the trial is a 'randomised controlled trial' (RCT).

Once assigned to a trial arm, each subject receives the respective treatment, and undergoes follow-up. In the simplest case the outcome for each subject is binary (dichotomous), e.g. 'treatment successful' vs. 'treatment not successful'. We can also

think of this as 'an event has occurred' vs. 'no event has occurred'. For the purposes of this introductory review we focus on this case of binary outcomes. We exclude censoring (e.g. patients withdrawing from the trial or otherwise not being followed up, and therefore not producing data). More complex outcomes in trial data may consist of a discrete set of more than two alternatives (e.g. 'ordinal outcomes' on a 5-point scale), or the outcome may be continuous, see e.g. [17].

2.2.1.2 Meta-analysis

Meta-analysis in general is concerned with combining evidence from multiple trials. The simplest case is 'pairwise' or 'standard' meta-analysis. Two treatment options are compared in a set of different trials. The purpose of meta-analysis is then to 'integrate' the outcomes of the different trials, and to estimate how effective the competing treatment options are. These estimates can then be used to decide if one of the two treatments is to be preferred over the other, and which one.

In this process it is important to bear in mind that the outcomes of different trials cannot always be aggregated directly. Clinical trials taking place at different locations will draw from a local patient pool, and as a result, the general characteristics of the subjects may differ from trial to trial (e.g. age, health or economic status, level of education etc.) which may affect the observed treatment effects. In order to combine evidence from multiple trials we require an underlying model – a stochastic process with unknown parameters leading to realisations of the data observed in trials. Two of the most common modelling approaches, fixed and random effects models, are discussed in Section 2.2.4. Once a given model assumption has been made, the objective of meta-analysis is to estimate parameters of the model from the data.

2.2.1.3 Networks of trials and NMA

General networks of treatments and trials capture more complex situations than the one described in Section 2.2.1.2. For example imagine that there are four different treatment options and several trials, each comparing a subset of treatments. Not every trial tests all four treatments, but the same pairwise comparison is perhaps made in different trials. This generates a network of treatment options and trials.

A possible scenario is illustrated in Figure 2.2. The network consists of three trials

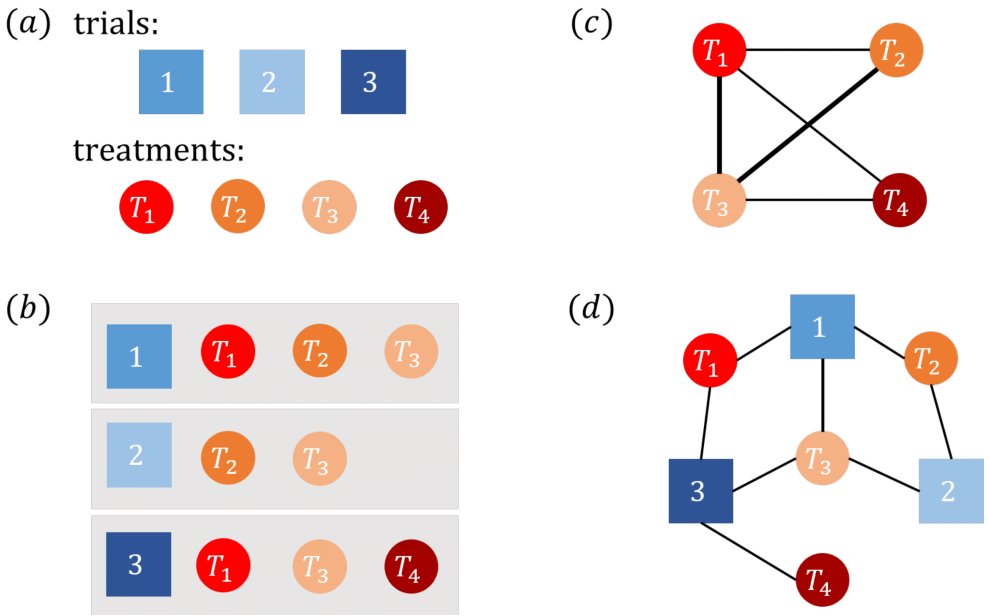


Figure 2.2: Illustration of a network of treatment options and trials. (a) There are three trials in the network (squares), and four treatments in total (circles). (b) Trial 1 has three arms (treatments T_1, T_2 and T_3), trial 2 is two-armed (treatments T_2 and T_3), and trial 3 tests treatments T_1, T_3 and T_4. (c) Presentation of the network as a graph with only one type of node. Each node represents one treatment, and two treatments are connected if they have been directly compared in at least one trial. The thickness of the edge connecting two nodes is proportional to the number of trials comparing those two treatments. This representation does not contain full information about the network of treatments and trials. (d) Representation as a bipartite graph of treatments and trials. This can also be understood as a hypergraph (related concepts include incidence or Levi graphs [26]).

(indicated by square boxes), comparing different subsets of four different treatments [panel (a)]. We label the treatments $T_1, \ldots, T_4$, and indicate them as circles on the graph. Panels (b), (c) and (d) show different graphical representations of this network; details can be found in the figure caption. Option (c) is most commonly used in practice.

Most notably in this example, there is no pairwise comparison between two fixed treatments that is made in all three trials (i.e. no pair of treatment arms features in all three trials). The pair T_2–T_3 appears in trials 1 and 2, so the use of conventional (pairwise) meta-analysis would be restricted to combining information regarding this particular pair from trials 1 and 2 only.

Network meta-analysis aims to integrate further information from the network. The information about the pair T_2–T_3 from trials 1 and 2 is referred to as *direct evidence* for the comparison of these two treatments. However, T_2 and T_3 are each also compared to T_1. For treatment option T_2 this happens in trial 1, and for T_3 in trials 1 and 3. These

comparisons to a common third treatment provide *indirect evidence* for the comparison of T_2 and T_3.

In the example, treatment options T_2 and T_4 are not compared directly in any trial. However, each of the two are directly compared to T_1 and T_3 [see Figure 2.2 (b)]. This indirect evidence can be used to infer information about the comparison between T_2 and T_4, even though there is no direct evidence.

Data example. Figure 2.1 shows the network for a real NMA data set comparing nine treatments for acute myocardial infarction (heart attack) [10–12]. The treatments are labelled $T_1, \ldots, T_9$ and their names are given in the figure. They consist of eight thrombolytic drugs and one angioplasty intervention (T_7). The data has therefore been referred to as the 'thrombolytic drug' data set [12, 27]. A heart attack occurs when blood flow to the heart is cut off, usually resulting from blockage of one or more of the coronary arteries. Thrombolytic drugs aim to dissolve blood clots that have blocked arteries whereas angioplasty is a procedure that tries to relieve blockage by widening the arteries. The data consists of 50 trials; two 3-arm trials (comparing T_1, T_3, T_4 and T_1, T_2, T_9, respectively) and 48 2-arm trials. Each trial records the number of deaths in each treatment arm that occur within 30 or 35 days of a heart attack. The network is represented by a weighted graph where for each pair of treatments, the thickness of the edge link is proportional to the number of trials comparing these two treatments. The area of each node is proportional to the number of patients allocated to the treatment represented by that node. Multi-arm trials are not explicitly indicated on the graph.

2.2.2 General notation for networks of trials and treatments

We write N for the total number of treatment options in the network. We label the different treatments $T_1, T_2, \ldots, T_N$, and we will use the indices a, b when we refer to elements of the set of treatments, i.e. $a, b \in \{T_1, \ldots, T_N\}$. The number of trials in the network is denoted by M, and we use the indices, i, j to refer to the different trials, i.e. $i, j \in \{1, \ldots, M\}$. We will use the words 'trial' and 'study' synonymously.

Each trial compares a subset of treatments. We write $A_i \subset \{T_1, \ldots, T_N\}$ for the set of treatment options compared in trial i. Hence $m_i \equiv |A_i|$ is the number of arms of study i. We number the arms of trial i by $\ell = 1, \ldots, m_i$, and denote the treatment

given to patients in arm ℓ of trial i by $t_{i,\ell}$. Each $t_{i,\ell}$ ($i = 1, \ldots, M$, $\ell = 1, \ldots, m_i$) is a treatment from the set $\{T_1, \ldots, T_N\}$.

In the illustration in Figure 2.2, we have $N = 4$ treatments and $M = 3$ trials. For trial 3, for example, we have $m_3 = 3$ (three-armed trial), $A_3 = \{T_1, T_3, T_4\}$ as well as $t_{3,1} = T_1, t_{3,2} = T_3$ and $t_{3,3} = T_4$.

We write $n_{i,\ell}$ for the number of subjects receiving the ℓ-th treatment in trial i, with $\ell = 1, \ldots, m_i$. Focusing on binary outcomes, the data available for each patient is whether an 'event' has occurred by the end of the study or not. (We note that, depending on the context, an event can either be treatment success or an adverse event as in the data example in Section 2.2.1.3.)

For each arm ℓ of trial i, the number of resulting events is recorded. We denote this by $r_{i,\ell}$. This quantity takes integer values in the range $0, 1, \ldots, n_{i,\ell}$

Summarising, trial i is defined by the treatments it compares, $t_{i,1}, \ldots, t_{i,m_i}$, by the number of patients in each arm, $n_{i,1}, \ldots, n_{i,m_i}$, and by the number of events in each arm $r_{i,1}, ..., , r_{i,m_i}$.

2.2.3 Absolute outcomes and relative treatment effects: the logit scale

We assume that the application of the treatment in arm ℓ of trial i generates events with probability $p_{i,\ell}$ independently for each of the $n_{i,\ell}$ patients at the end of this trial arm [28]. As a consequence, each $r_{i,\ell}$ is a binomial random variable,

$$\text{Prob}(r_{i,\ell} = r) = \begin{pmatrix} n_{i,\ell} \\ r \end{pmatrix} p_{i,\ell}^r (1 - p_{i,\ell})^{n_{i,\ell} - r}, \tag{2.1}$$

for $i = 1, \ldots, M$ and $\ell = 1, \ldots m_i$.

2.2.3.1 Absolute outcomes

The $\{p_{i,\ell}\}$ can be interpreted as the 'absolute outcomes' for each treatment group, they capture how likely it is that the different treatments produce 'events'. The word 'absolute' indicates that these values are not expressed with reference to any other treatment or baseline.

In the context of binary data, absolute outcomes are frequently expressed in terms of so-called 'log-odds'. For a probability p (i.e. a number $p \in [0, 1]$) the term 'odds' refers to the ratio $p/(1-p)$, and the log-odds or 'logit' ('logistic unit') is defined as

$$\text{logit}(p) = \ln \frac{p}{1-p}. \tag{2.2}$$

While the original probability p is restricted to the range $0 \leq p \leq 1$, the logit of p can take values on the entire real axis, with $\lim_{p\to 0} \text{logit}(p) = -\infty$, and $\lim_{p\to 1} \text{logit}(p) = \infty$.

Accordingly, we can express the treatment outcomes in terms of $\lambda_{i,\ell} \equiv \text{logit}(p_{i,\ell}) = \ln p_{i,\ell} - \ln(1-p_{i,\ell})$. In a slight abuse of terminology we will refer to both the $\{\lambda_{i,\ell}\}$ and to the $\{p_{i,\ell}\}$ as the absolute outcomes. From the context it will be clear what we mean.

The logit transformation in Equation (2.2) is an example of a so-called 'link function'. Outcomes from medical trials come in many forms (e.g. time-to-event, ordered categories, continuous measurements) and are generated from a range of distributions (e.g. normal, binomial, Poisson). By using a link function to transform the treatment outcome associated with a particular type of data onto the continuous scale we can then use the same basic model for a range of different data types. The choice of logit in Equation (2.2) is a practical choice for binomial data. Exchanging the definitions of event vs no event (i.e. $p \leftrightarrow 1-p$) only results in a sign reversal, i.e. it makes no difference for the mathematical model and inference process. This is not true for some other choices of the link function [29, 30].

In the following we describe the NMA model for binomial data with a logit link. The likelihood function (defined further below) is based on a binomial distribution. To analyse other types of data, one can use the same basic NMA model but the likelihood and link functions vary depending on the type of this data. See References [17, 31] for an overview of different data types and their corresponding link functions.

2.2.3.2 Relative treatment effects, transitivity and common baseline

We now introduce the so-called *relative treatment effect* for treatments a and b. If these two treatments have absolute outcomes λ_a and λ_b, then we write

$$d_{ab} \equiv \lambda_b - \lambda_a \tag{2.3}$$

for the relative treatment effect for this pair. This definition implies $d_{ab} = -d_{ba}$, and $d_{aa} = 0$.

In formulating this setup we assume that the relative treatment effects fulfil the transitivity relation $d_{ab} = d_{ac} + d_{cb}$ for all triplets of treatments a, b and c. Alternatively, this can be written as

$$d_{ab} = d_{cb} - d_{ca}. \tag{2.4}$$

Using transitivity, the relative treatment effects of all pairs in a network of N treatments, $T_1, \ldots, T_N$, are fully specified by $N-1$ numbers. For example, we can designate treatment T_1 as the overall 'global' baseline. It is then sufficient to know $d_{T_1 a}$, for $a \in \{T_2, \ldots, T_N\}$. These values are termed the 'basic parameters' [32, 33] and we collect them in the $(N-1)$-component vector $\boldsymbol{d} = (d_{T_1T_2}, d_{T_1T_3}, \ldots, d_{T_1T_N})^\top$ (the notation $(\ldots)^\top$ indicates transposition). The relative treatment effect d_{ab} for any pair of treatments a, b can then be determined from Equation (2.4) using $c = T_1$.

Transitivity describes an assumption made in setting up the model. It is applicable to all possible comparisons in the network. The 'statistical manifestation' [19] of transitivity in observed data is referred to as consistency. I.e. if direct and indirect evidence exist in the data for a particular comparison, then the data is consistent if there is no discrepancy in the treatment effects obtained via the two types of evidence.

2.2.4 Fixed and random effects models

In this section we describe the random and fixed effects models used in meta-analysis. We will abbreviate these as RE and FE models respectively. The FE model is a limiting case of the RE model.

2.2.4.1 General idea: Modelling relative effects

The usual approach to NMA is to model the relative treatment effects rather than the absolute outcomes. Individual trials are likely to have different characteristics in terms of population demographics (e.g. age, socio-economic status, baseline health) and trial procedures (such as treatment dosage or administration). It is therefore likely that absolute outcomes, such as the number of patients who experience an event, will vary substantially depending on these characteristics. For example, we are likely to observe a higher proportion of deaths in a trial with an older population compared to a younger population. It is then unrealistic to model the absolute outcomes as

being comparable across all trials. A less restrictive model is to assume that *differences* between treatment outcomes are similar across the trials. For example, if treatment T_1 is more effective than treatment T_2 then it is likely that in each trial we will observe fewer deaths in arm T_1 than in arm T_2 even if the overall number of deaths in each trial is very different.

With this in mind, we assume that each trial comparing treatments a and b is associated with some unknown relative treatment effect $\Delta_{i,ab}$ which represents the 'true' difference in effectiveness between a and b in trial i. Trial i then provides information about $\Delta_{i,ab}$ subject to some sampling error (due to the finite number of participants in that trial). This information is said to be 'observed'. The FE and RE models differ in the assumptions placed on these 'trial-specific' relative effects.

The FE model assumes that each trial has the same underlying relative treatment effects, i.e. $\Delta_{i,ab} = d_{ab}\ \forall i$. The RE model, on the other hand, is more flexible. Rather than requiring that the trial-specific relative effects are the same in every trial (as in the FE model), they are instead assumed to be 'exchangeable' [5, 34, 35]. In other words, the true relative treatment effects in each trial are treated as random variables, drawn from an underlying distribution [7]. This reflects the fact that differences in trial characteristics may mean that a particular treatment option is comparatively more effective in one trial than in another. For example, treatment T_1 may be the most effective treatment for participants of all ages, but it could yield even better results for younger patients. Trials with a younger demographic may then observe a larger relative effect between T_1 and T_2 compared with a trial of older participants. More specifically, the RE model assumes that the relative effect of two treatments a and b is drawn from the *same* distribution for any trial involving these two treatments, i.e. this distribution is the same for all i. We are interested in the mean of this distribution. This indicates the typical relative effect between the treatments.

The RE model therefore consists of two levels of randomness; one due to variations *between* trials and the other due to sampling *within* a given trial. In the FE model, we only allow for the latter.

2.2.4.2 Transitivity of trial-specific relative effects

We assume that the trial-specific relative treatment effects fulfil the transitivity relations in Equation (2.4). For a trial with m arms it is therefore sufficient to designate a trial-specific baseline treatment and its absolute outcome, and the treatment effects of the $m-1$ remaining arms in the trial relative to this baseline. This will fully determine the true absolute outcomes associated with all treatments in the trial.

In our model there are m_i arms in trial i, labelled $\ell = 1, \ldots, m_i$. Without loss of generality we use $\ell = 1$ as the trial-specific baseline treatment. We then write

$$\Delta_{i,1\ell} = \ln \frac{p_{i,\ell}}{1-p_{i,\ell}} - \ln \frac{p_{i,1}}{1-p_{i,1}} \tag{2.5}$$

for the effect of treatment $t_{i,\ell}$ in the trial relative to this baseline.

The absolute outcomes $p_{i,\ell}$ of all m_i arms in the trial can then be obtained from the true absolute outcome of the baseline, $p_{i,1}$, and $\Delta_{i,12}, \ldots, \Delta_{i,1m_i}$.

2.2.4.3 Model definitions

We now formalise our model. As we have seen, the RE model consists of two level of randomness. One is between-trial variation and the other is due to sampling in a given trial. We describe these in turn[2].

Between-trial variation. We assume that the relative treatment effects for a given trial i are drawn from a multivariate normal distribution,

$$\begin{pmatrix} \Delta_{i,12} \\ \vdots \\ \Delta_{i,1m_i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} d_{t_{i,1}t_{i,2}} \\ \vdots \\ d_{t_{i,1}t_{i,m_i}} \end{pmatrix}, \boldsymbol{\Sigma_i} \right). \tag{2.6}$$

The first argument is the mean, the second argument is the covariance matrix. We will now describe these first and second moments in more detail.

Given Equation (2.6), the relative effect between two treatments a and b in a given trial is a Gaussian random variable. In particular, the relative effect between these

[2]In our description the first level of randomness is the between-trial variation, whereas the second level is the sampling randomness within a trial. This reflects the 'mechanistic' view a physicist might take, and focuses on how one would generate synthetic trial data in a simulation. A statistician might take a reverse view and see the trials as the starting point (hence sampling noise is the first level of randomness). The synthesis of several trials then follows later, and the between-trial randomness therefore comes second for the statistician.

treatments varies across different trials involving a and b. The model parameters d_{ab} are the averages of these random numbers. More precisely, for a given pair a and b the parameters d_{ab} can be interpreted as the 'typical' relative effect one should expect to see in a trial involving these treatments.

Using the transitivity relation in Equation (2.4), we can write the vector of mean treatment effects in trial i, $\boldsymbol{d}_i = (d_{t_{i,1}t_{i,2}}, \ldots, d_{t_{i,1}t_{i,m_i}})^\top$ in terms of the vector of basic parameters via

$$\boldsymbol{d}_i = \mathbf{X}_i \boldsymbol{d}. \tag{2.7}$$

We note that $\boldsymbol{d}$ is a vector with $N-1$ entries and that $\boldsymbol{d}_i$ has $m_i - 1$ components. The $(m_i - 1) \times (N-1)$ matrix $\mathbf{X}_i$ describes which treatments are compared in trial i and is called the 'design matrix' of the trial. Each of the $N-1$ columns of $\mathbf{X}_i$ represents a treatment $a \in \{T_2, \ldots, T_N\}$. The $m_i - 1$ rows represent the comparisons of treatments $t_{i,\ell}$ $(\ell = 2, \ldots, m_i)$ to the trial-specific baseline $t_{i,1}$. In a given row ℓ all entries of $\mathbf{X}_i$ are zero, except those corresponding to the treatments that are being compared. More precisely, we distinguish two cases: (i) the trial-specific baseline treatment is not the global baseline $(t_{i,1} \neq T_1)$, and (ii) the trial-specific baseline is the global baseline $(t_{i,1} = T_1)$.

We focus first on case (i). In row ℓ the matrix entry for the treatment $t_{i,\ell}$ that is being compared against the trial-specific baseline is set to $+1$. The entry in the column corresponding to the trial-specific baseline [which is among the $\{T_2, \ldots, T_N\}$ in case (i)] is set to -1. All other $N-3$ entries in row ℓ are zero. In situation (ii), we again set the entry for the treatment $t_{i,\ell}$ that is being compared against the trial-specific baseline to $+1$. All other $N-2$ entries in row ℓ are zero.

To illustrate this, consider the example network in Figure 2.3. This network consists of $N = 5$ treatments and $M = 2$ trials. The global baseline treatment is T_1. The vector of basic parameters is $\boldsymbol{d} = (d_{T_1T_2}, d_{T_1T_3}, d_{T_1T_4}, d_{T_1T_5})^\top$. Trial $i = 1$ compares treatments $A_1 = \{T_2, T_3, T_4, T_5\}$ with trial-specific baseline $t_{1,1} = T_2$. This is an example of option (i) and its design matrix is

$$\mathbf{X}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}. \tag{2.8}$$

Trial $i = 2$ compares treatments $A_2 = \{T_1, T_3, T_5\}$ and its trial-specific baseline is

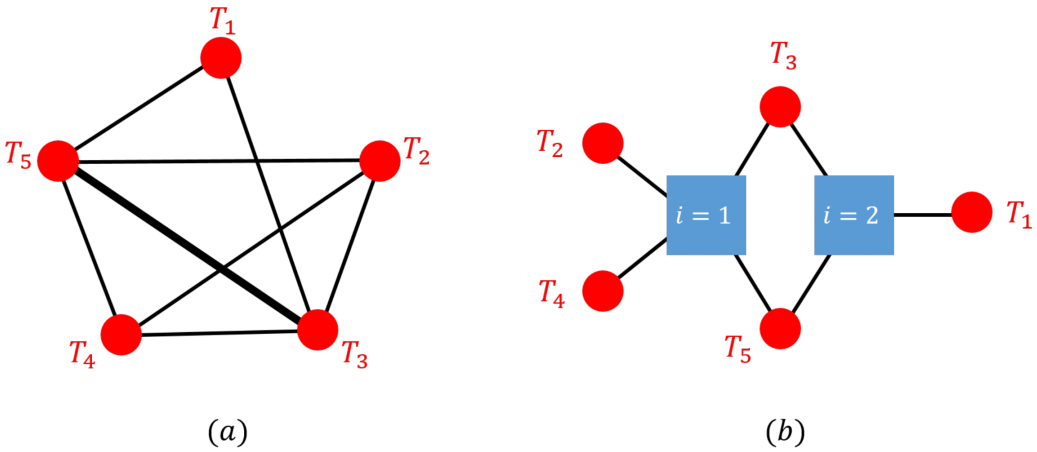


Figure 2.3: A fictional network with $N = 5$ treatments and $M = 2$ trials. Trial $i = 1$ compares treatments $A_1 = \{T_2, T_3, T_4, T_5\}$ and trial $i = 2$ compares $A_2 = \{T_1, T_3, T_5\}$. (a) Standard network representation where the thickness of each edge relates to the number of trials that make that comparison. Here, only the pair $\{T_3, T_5\}$ appears in both trials. (b) Network representation as a bipartite graph.

$t_{2,1} = T_1$. This is an example of option (ii) and its design matrix is

$$\mathbf{X}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.9}$$

The $(m_i - 1) \times (m_i - 1)$ matrix $\boldsymbol{\Sigma_i}$ in Equation (2.6) describes the variance of the relative treatment effects, $\Delta_{i,1\ell}$ $(\ell = 2, \ldots, m_i)$, and their correlations. Following References [36–38], we will assume that its diagonal elements are all identical. We write τ^2 for their common value. This is the variance of each $\Delta_{i,1\ell}$. We will further assume that the covariance between any two treatment effects is $\tau^2/2$ (these are the off-diagonal elements of $\boldsymbol{\Sigma_i}$). This ensures that the relative effect $\Delta_{i,1\ell} - \Delta_{i,1\ell'}$ between *any two* treatments $\ell \neq \ell'$ in trial i has variance τ^2.[3]

The between-trial variance τ^2 is termed the *heterogeneity* variance. We will refer to its square root, τ, as the heterogeneity parameter. Occasionally we will simply use 'heterogeneity' to refer to either the parameter or the variance but it should be clear from the context what we mean. Usually this distinction in not important.

The aim of network meta-analysis is to estimate the mean relative treatment effects d_{ab} for all pairs $a \neq b$, and the heterogeneity parameter, τ. Given the transitivity assumption in Equation (2.4) not all d_{ab} are independent. As explained earlier, we can use treatment $a = T_1$ as the overall global baseline treatment, and it is sufficient to

[3]This can be seen from $\mathrm{Var}(\Delta_{i,1\ell} - \Delta_{i,1\ell'}) = \mathrm{Var}(\Delta_{i,1\ell}) + \mathrm{Var}(\Delta_{i,1\ell'}) - 2\mathrm{Cov}(\Delta_{i,1\ell}, \Delta_{i,1\ell'})$.

estimate d_{T_1a} for $a = T_2, \ldots, T_N$ [33].

Sampling noise in a given trial. In a second level of randomness the model assumes that the application of the treatment in arm ℓ of trial i generates events with probability $p_{i,\ell}$ independently for each of the $n_{i,\ell}$ patients at the end of this trial arm [28]. Each $r_{i,\ell}$ is then a binomial random variable, as described in Equation (2.1).

The RE model is summarised and illustrated in Figure 2.4. The FE model is simply a special case of the RE model where $\tau = 0$. As previously explained, for any fixed pair of treatments, there is then no variation in the relative treatment effects between trials.

2.2.5 Generation of synthetic data in simulations

The different levels of randomness in the model can be understood by thinking about how one would simulate synthetic trial data in line with the model assumptions.

We begin such a process by defining the fixed parameters of the network. First, we pick the network configuration; the number of treatments N, the total number of studies, M, the number of arms in each trial $\{m_i\}$, the set of treatments these arms relate to $\{t_{i,\ell}\}$, and the number of participants in each arm $\{n_{i,\ell}\}$. We then assign the 'true' values of the model parameters, i.e. of $\boldsymbol{d} = (d_{T_1T_2}, d_{T_1T_3}, \ldots, d_{T_1T_N})^\top$ and τ.

Following this set-up, we generate independent realisations $\nu = 1, 2, ..., \Omega$ of synthetic trial outcomes. Specifically, for each ν:

(1) For all trials i, randomly sample the parameters $\Delta_{i,1\ell}$, $\ell = 2, \ldots, m_i$, from the multivariate normal distribution in Equation (2.6).

(2) Using the $\Delta_{i,1\ell}$, $\ell = 2, \ldots, m_i$, and a 'data generating model' still to be defined (see below), construct the probabilities $p_{i,\ell}$, $\ell = 1, \ldots, m_i$, for all trials i in the network.

(3) For each trial arm, generate random event data ('observations'), $r_{i,\ell}$, from the binomial distribution in Equation (2.1).

In step (2) of the simulation procedure, the relative treatment effects $\Delta_{i,1\ell}$ ($\ell = 2, \ldots, m_i$) in any one trial i do not uniquely define the absolute outcomes $p_{i,\ell}$ ($\ell = 1, \ldots, m_i$) required for step (3). Equation (2.5) can be re-arranged to give

$$p_{i,\ell} = p_{i,\ell}[p_{i,1}, \Delta_{i,1\ell}] = \frac{p_{i,1}\mathrm{e}^{\Delta_{i,1\ell}}}{1 + p_{i,1}\left(\mathrm{e}^{\Delta_{i,1\ell}} - 1\right)}, \tag{2.10}$$

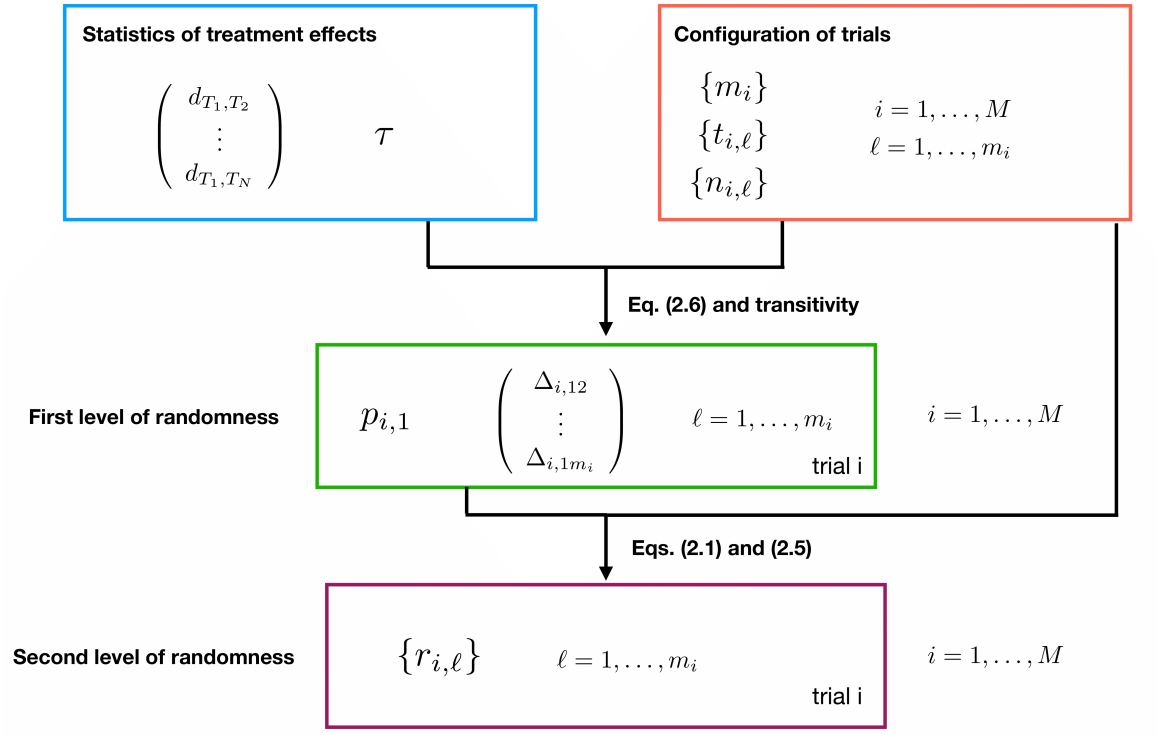


Figure 2.4: Diagram summarising the random effects model of NMA. The main input parameters are the configuration of trials and the statistics of treatment effects. The trial configuration is set by the number of trials M in the network, the number of arms in each trial (m_i), the treatment options used in these arms ($t_{i,\ell}$) and the number of patients in each arm ($n_{i,\ell}$). The statistics of the treatment effects are parameterised by the mean effect of each treatment $T_2, \ldots, T_N$ relative to the overall baseline treatment T_1, and the heterogeneity parameter τ. In a first step of randomness realisations of the random variables describing the treatment effects in the different trials ($p_{i,1}$ and $\Delta_{i,12}, \ldots, \Delta_{i,1m_i}$) are drawn for each trial from the distribution in Equation (2.6), supplemented by a distribution for each $p_{i,1}$. These are then used along with Equation (2.5) to construct the absolute outcomes of the treatments in each trial. From these, and using the number of participants (the $\{n_{i,\ell}\}$), the number of events in each arm (the $\{r_{i,\ell}\}$) are then drawn from the binomial distributions in Equation (2.1). The fixed effect model is the special case $\tau = 0$. The distribution in Equation (2.6) then turns into a delta-distribution. In this scenario, the true relative effect between two treatments a and b does not vary between trials and is given by d_{ab}.

so that $p_{i,1}$ together with the $\Delta_{i,1\ell}$ ($\ell = 2, \ldots, m_i$) specifies all absolute outcomes in trial i.

To fully define step (2) in the above algorithm it is therefore sufficient to specify the construction of $p_{i,1}$. A discussion of possible data generating models for $p_{i,1}$ is given in Seide et al (2019) [39]. We briefly describe two possible methods, highlighting the resulting symmetry or asymmetry introduced.

One simple procedure involves sampling $p_{i,1}$ for each trial from some specified distribution. For example one could choose a uniform distribution between two limits (perhaps zero and one to sample the full range of outcomes) or from a normal distribution truncated at zero at the lower end, and at one at the upper end. One then obtains the other absolute outcomes via Equation (2.10). By using a different method for generating the absolute outcome of the trial-specific baseline ($p_{i,1}$) compared to the other absolute outcomes in the trial ($p_{i,\ell\neq 1}$), this method introduces asymmetry into the generation procedure. A simple way of reducing the effect of this asymmetry on the synthetic data is to randomly select the trial-specific baseline treatments at each iteration ν.

An alternative data generation model was proposed by Seide et al (2019) [39]. Here, one chooses the absolute outcome for the baseline treatment to be the value that minimises the Euclidean distance of the vector $(p_{i,1}, \ldots, p_{i,m_i})$ from the vector $(1/2, \ldots, 1/2)$, i.e.

$$p_{i,1} = \min_q \left[\left(q - \frac{1}{2} \right)^2 + \sum_{\ell=2}^{m_i} \left(p_{i,\ell}\left[q, \Delta_{i,1\ell} \right] - \frac{1}{2} \right)^2 \right], \tag{2.11}$$

where $p_{i,\ell}[\cdot,\cdot]$ is the expression given in Equation (2.10). This method maintains symmetry since all m_i absolute outcomes in trial i are determined simultaneously.

2.2.6 The process of carrying out a network meta-analysis – Brief overview

2.2.6.1 Frequentist versus Bayesian network meta-analysis

In Sections 2.3 and 2.4 below we describe the two main approaches to carrying out an NMA, i.e. the steps that are used to infer parameters such as relative treatment effects from the observed trial data. The two approaches correspond to the two main

branches of inference in general, frequentist and Bayesian inference.

Much has been written about the difference between Bayesian and frequentist approaches to inference (e.g. [40–44]). One central point distinguishing the two is the conception of probability. Frequentist inference defines the probability of some event in terms of how frequently the event occurs if we repeat some process (e.g. an experiment) many times [40]. The Bayesian approach instead uses probability to describe the degree of belief in a statement [45, 46]. In the Bayesian framework, parameters such as treatment effects are considered random variables, where the randomness reflects the remaining uncertainty after the inference process. If the distribution for a parameter is very sharp, then this indicates that we can be fairly certain that the inferred parameter is in a given range around the mode of that distribution. If the distribution is wide then the strength of our beliefs is weak. In the Bayesian approach probability therefore becomes subjective. It is not a property of the system only, but also of the prior beliefs, and the information available to the observer. Given that probability in Bayesian statistics reflects the degree of belief in the value of parameters, we can make statements such as 'Given our prior beliefs and the data we have observed, we think that treatment A is more effective than treatment B with probability 70%'. In frequentist methodology, probability is not an expression of our beliefs and therefore an equivalent statement would be, for example, 'Based on hypothesised repetitions of the experiment, treatment A would be estimated to be more effective than treatment B 70% of the time'.

A concept used in both Bayesian and frequentist inference is the 'likelihood function'. For observed data $\mathbf{D}$, the likelihood function is the conditional probability (or probability density) of observing this data given a specific set of model parameters $\boldsymbol{\theta}$, $P(\mathbf{D}|\boldsymbol{\theta})$[4]. The likelihood function is so named because it describes how likely it is to observe the given data for different values of parameters. In fact, the likelihood is viewed as a function of the parameters rather than the data and – somewhat confusingly – is often written as $L(\boldsymbol{\theta}|\mathbf{D})$.

[4]One way of thinking about this is as follows: Consider the map $P : (\mathbf{D}, \boldsymbol{\theta}) \mapsto P(\mathbf{D}, \boldsymbol{\theta})$ as a real-valued function of the two arguments $\mathbf{D}$ and $\boldsymbol{\theta}$ (which each may be multi-dimensional). We can then look at this from two perspectives: (i) Fixing $\boldsymbol{\theta}$ one obtains a map $\mathbf{D} \mapsto P(\mathbf{D}, \boldsymbol{\theta})$ describing the probability distribution (or density) of the data for given fixed parameters. Equation (2.1) is an example. (ii) If we fix $\mathbf{D}$ we obtain a map $\boldsymbol{\theta} \mapsto P(\mathbf{D}, \boldsymbol{\theta})$. This is the likelihood for the parameter $\boldsymbol{\theta}$, given the (fixed) data $\mathbf{D}$.

2.2.6.2 Arm-based versus contrast-based data

In the setup so far we have treated the $\{r_{i,\ell}\}$ (the number of events in the arm of the trials) as the trial outcome or 'data' in an NMA. Since each of the $r_{i,\ell}$ is associated with an arm in a trial, data of this type is referred to as 'arm-level' data.

The event probabilities associated with these measurements can also be expressed on the logit scale and used to calculate the log odds ratio (LOR) representing the relative effect between two treatments. For example, we write $y_{i,1\ell}$ for the *observed* LOR between the effect of treatment ℓ in trial i and the baseline treatment in that trial,

$$y_{i,1\ell} = \ln \frac{r_{i,\ell}/n_{i,\ell}}{1 - r_{i,\ell}/n_{i,\ell}} - \ln \frac{r_{i,1}/n_{i,1}}{1 - r_{i,1}/n_{i,1}}. \tag{2.12}$$

Sometimes a trial will only report the log odds ratio of each treatment relative to the trial-specific baseline (the $\{y_{i,1\ell}\}$), and not the detailed number of events in each arm (the $\{r_{i,\ell}\}$). The log odds ratio is a so-called 'summary statistic' and data of this type is called 'summary-level data'.

In NMA we can choose to model data on the arm level or the summary level. We refer to these approaches as 'arm-based' (AB) and 'contrast-based' (CB) models respectively [22]. Arm-level data is modelled using the binomial distribution [Equation (2.1)]. The likelihood function of the data is then also based on this binomial distribution. This is sometimes referred to as the 'exact-likelihood' or 'AB-likelihood' [47] approach. In contrast-based models the LORs from each trial are modelled as following a normal distribution. This is an approximation, and the approach is also referred to as the 'approximate-likelihood' or the 'CB-likelihood' [47] model[5].

Both frequentist and Bayesian inference methods can be used to evaluate both AB and CB models. In practice, frequentist models [37, 48, 49] are usually based on contrast-level summaries while Bayesian models [9, 17, 38] tend to use arm-level data [50]. Clearly, if trials only report summary-level data then we are restricted to CB models.

In the next section we summarise the general Bayesian approach to inference,

[5]The terms 'contrast-based' and 'arm-based' have also been used to distinguish between models of relative treatment effects (CB) and models of absolute outcomes associated with each treatment or 'arm' (AB) [35]. The latter are not standard practice. All models discussed in this article are based on relative treatment effects and we use CB/AB to distinguish between the 'level' of data that is used in constructing the model.

and describe how this is applied in an arm-based NMA model. In Section 2.4 we first describe the contrast-based NMA model and give an overview of the frequentist approach to inference. We then show how frequentist inference can be used to estimate relative treatment effects in a CB model.

2.3 Bayesian Network Meta-Analysis

In this section we discuss the Bayesian approach to network meta-analysis. We use an arm-based model and treat the number of events in each arm of each trial (the $r_{i,\ell}$) as the raw data in the model.

2.3.1 General approach

The process of Bayesian NMA converts prior beliefs on the distribution of model parameters into posterior distributions using the observed data. The approach is based on Bayes theorem, which in its simplest form can be stated as $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$, where A and B are outcomes of a probabilistic experiment. Writing $\boldsymbol{\theta}$ for the parameters, and $\mathbf{D}$ for the data, this becomes

$$P(\boldsymbol{\theta}|\mathbf{D}) = P(\mathbf{D}|\boldsymbol{\theta})\frac{P(\boldsymbol{\theta})}{P(\mathbf{D})}. \tag{2.13}$$

We are interested in the distribution of parameters given the observed data. In this context, we notice that the term $P(\mathbf{D})$ is not a function of $\boldsymbol{\theta}$, and so we can write

$$P(\boldsymbol{\theta}|\mathbf{D}) = \text{const} \times P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{2.14}$$

where the constant on the right is to be determined from normalisation. This is the fundamental equation for Bayesian NMA (and any other type of Bayesian inference). The object $P(\boldsymbol{\theta})$ on the right is known as the *prior* distribution of parameters. It reflects our beliefs about what the parameters might be, before we have taken into account the data $\mathbf{D}$. The expression on the left is the *posterior distribution* of the parameters, it represents our updated beliefs having observed and used the data $\mathbf{D}$. The factor that connects the two is the conditional probability, or likelihood function, $P(\mathbf{D}|\boldsymbol{\theta})$ (see Section 2.2.6.1).

2.3.2 Hierarchical structure of the random effects model

In NMA the parameters $\boldsymbol{\theta}$ are the true relative treatment effects $\boldsymbol{d}$ and the heterogeneity parameter τ. These are the 'parameters of interest' [40, 51] of the model and we will refer to them simply as the 'model parameters'. As described in Section 2.2.4 there are two levels of randomness in the RE model. The first layer generates trial-specific relative treatment effects $\boldsymbol{\Delta}_i$ $(i = 1, \ldots, M)$ and absolute outcomes for the baseline treatment in each trial (the $p_{i,1}$ in Figure 2.4). In a second layer, binomial outcomes are then produced for each trial arm.

The trial-specific effects $\boldsymbol{\Delta}_i$ and the $p_{i,1}$, are so-called 'nuisance parameters' [40, 51]. For the discussion of the general Bayesian approach we will call these $\boldsymbol{\nu}$. They are random variables, and their distribution is parameterised by the model parameters $\boldsymbol{\theta}$. The nuisance parameters in turn determine the distribution of the output data $\mathbf{D}$. This is captured by the following relation

$$P(\mathbf{D}|\boldsymbol{\theta}) = \int d\boldsymbol{\nu}\ P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu}) P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta}). \tag{2.15}$$

We write $P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta})$ to describe the internal layer of the model (generation of nuisance parameters from the model parameters), and $P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu})$ for the 'output layer' (generation of data from the nuisance parameters).

Using Equation (2.14) we then have

$$P(\boldsymbol{\theta}|\mathbf{D}) = \text{const} \times \left(\int d\boldsymbol{\nu}\ P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu}) P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta}) \right) P(\boldsymbol{\theta}), \tag{2.16}$$

where $P(\boldsymbol{\theta})$ is the prior distribution of the parameters $\boldsymbol{\theta}$.

In Section 2.3.3 below, we focus on the construction of

$$U(\mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\theta}) \equiv P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu}) P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta}) P(\boldsymbol{\theta}). \tag{2.17}$$

This is the joint distribution of the model parameters $\boldsymbol{\theta}$, the nuisance parameters $\boldsymbol{\nu}$ and the data $\mathbf{D}$.

To obtain the posterior distribution $P(\boldsymbol{\theta}|\mathbf{D})$ one fixes $\mathbf{D}$ to be the observed data. The next step then is to integrate out the nuisance parameters in Equation (2.16). The normalisation constant in this equation can be determined at the end.

We have now reduced the problem of carrying out an NMA to two tasks:

1. We need to construct explicit forms for the factors on the right-hand side of Equation (2.17);

2. We need a method with which to integrate out the nuisance parameters, and to extract the posterior distribution for $\boldsymbol{\theta}$.

We will first discuss step 1. Numerical methods for step 2 are described in Section 2.3.4.

2.3.3 Construction of the joint distribution of model parameters, nuisance parameters and the data

2.3.3.1 Choice of priors for the model parameters

The parameters of interest in the model are the heterogeneity τ, and the true treatment effects of treatments $a \in \{T_2, \ldots, T_N\}$ relative to the overall baseline treatment T_1. The method requires distributions capturing prior beliefs on the values these parameters might take.

It is common to choose a Gaussian distribution as the prior for the relative treatment effects. The prior for τ has evoked more discussion [36, 52–55], but the usual practice is to use a uniform prior distribution between zero and some upper limit, $\tau_{\max}$, which can depend on the data [17, 33].

This results in the form

$$P(\boldsymbol{\theta}) = \frac{1}{\tau_{\max}} \mathbb{1}_{[0,\tau_{\max}]}(\tau) \times \prod_{a \in \{T_2,\ldots,T_N\}} \frac{\exp\left(-\frac{d_{T_1 a}^2}{2\sigma_d^2}\right)}{\sqrt{2\pi\sigma_d^2}}, \tag{2.18}$$

with the indicator function $\mathbb{1}_{[x,y]}(\tau) = 1$ for $x \leq \tau \leq y$, and $\mathbb{1}_{[x,y]}(\tau) = 0$ otherwise. The product over a has $N-1$ factors (one for each $a \in \{T_2, \ldots, T_N\}$) and indicates that the prior distribution for each of the true relative treatment effects $d_{T_1 a}$ is a Gaussian distribution with mean zero, and variance σ_d^2. In using this factorised form, we have assumed that these parameters are a priori pairwise independent [18, 21].

It is common to use so-called 'non-informative' priors. These are relatively broad distributions for each of the parameters, reflecting a situation in which little information about the parameters is known a priori. This can be achieved by choosing values for $\tau_{\max}$ and σ_d^2 that are large compared with the typical scale of the parameters τ and

$\boldsymbol{d}$ respectively. What constitutes as 'large' is informed by the type of data and the medical condition/clinical question of interest.

Occasionally it may be necessary to use an informative prior for the heterogeneity parameter [17, 36, 56–58]. If the network contains very few trials per comparison then there is little information about the variation between trials and the estimation of τ is likely to be imprecise [58]. In particular, when a flat prior is used with data that gives little information about between-trial variance then the posterior of τ will be dominated by the prior which may lead to unrealistically high estimates of heterogeneity [17]. Informative priors have been proposed for τ based on external data, for example databases of existing meta-analyses that relate to the relevant data type, medical condition and interventions [56, 57]. The use of such priors then allows us to incorporate external information about τ into the inference process.

2.3.3.2 Distribution of nuisance parameters for given model parameters

The model parameters ($\boldsymbol{d} = (d_{T_1,T_2}, \ldots, d_{T_1,T_N})^\top$ and τ) of the RE model determine the distribution of relative treatment effects $\boldsymbol{\Delta}_i$ for each of the trials $i = 1, \ldots, M$. As explained in more detail in Section 2.2.4.3, the RE model assumes that each entry of $\boldsymbol{\Delta}_i$ is drawn from a Gaussian distribution centred on $\mathbf{X}_i\boldsymbol{d}$, where $\mathbf{X}_i$ is the $(m_i - 1) \times (N - 1)$ design matrix for trial i. The variance of each component of each $\boldsymbol{\Delta}_i$ is τ^2. The correlation between any two different entries of $\boldsymbol{\Delta}_i$ is assumed to be $\tau^2/2$, and there are no correlations between $\boldsymbol{\Delta}_i$ and $\boldsymbol{\Delta}_j$ for two different trials $i \neq j$.

Putting this all together we have

$$P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta}) = \prod_{i=1}^{M} \left[\frac{\exp\left[-\frac{1}{2}(\boldsymbol{\Delta}_i - \mathbf{X}_i\boldsymbol{d})^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}_i - \mathbf{X}_i\boldsymbol{d})\right]}{(2\pi)^{(m_i-1)/2}\det(\boldsymbol{\Sigma}_i)^{1/2}} \times P_{\text{bl},i}(p_{i,1}) \right], \tag{2.19}$$

where for a given trial i, the matrix $\boldsymbol{\Sigma}_i$ is of size $(m_i - 1) \times (m_i - 1)$, and has diagonal entries τ^2, and off-diagonal entries $\tau^2/2$ (see Section 2.2.4.3). The term $P_{\text{bl},i}(p_{i,1})$ is the distribution for the absolute outcome associated with the trial-specific baseline. It is here common to use a non-informative distribution, such as a normal distribution for $\text{logit}(p_{i,1})$ with large variance [17, 21].

2.3.3.3 Distribution of data for given nuisance parameters

Finally, the data is given by the number of 'events' in each arm of each of the trials. The nuisance parameters $p_{i,1}$ and $\Delta_{i,1\ell}$ ($\ell = 2, \ldots, m_i$) for trial i translate into event probabilities for the treatments in this trial via Equation (2.10) which we re-state here,

$$p_{i,\ell} = p_{i,\ell}(\boldsymbol{\nu}) = \frac{p_{i,1} e^{\Delta_{i,1\ell}}}{1 + p_{i,1}(e^{\Delta_{i,1\ell}} - 1)}. \tag{2.20}$$

For binary outcomes the distribution of the data for given nuisance parameters is the binomial for each arm, with parameters $n_{i,\ell}$ and $p_{i,\ell}$. Combining this for all arms of all trials in the network we arrive at

$$P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu}) = \prod_{i=1}^{M} \prod_{\ell=1}^{m_i} \begin{pmatrix} n_{i,\ell} \\ r_{i,\ell} \end{pmatrix} p_{i,\ell}(\boldsymbol{\nu})^{r_{i,\ell}} [1 - p_{i,\ell}(\boldsymbol{\nu})]^{n_{i,\ell} - r_{i,\ell}}. \tag{2.21}$$

2.3.4 Computational techniques

The posterior distribution for the model parameters is obtained from

$$P(\boldsymbol{\theta}|\mathbf{D}) = \text{const} \times \int d\boldsymbol{\nu}\; U(\mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\theta}), \tag{2.22}$$

where U is defined in Equation (2.17). Even though U is known in closed form, the integral over $\boldsymbol{\nu}$ cannot be performed analytically. Due to the high-dimensionality of the integral, direct numerical integration is not always viable either.

One therefore resorts to computational methods to sample combined values for $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$, and then considers the resulting marginal distribution for $\boldsymbol{\theta}$. The most common techniques to do this in meta-analysis are Markov Chain Monte Carlo (MCMC) methods. Popular MCMC software include WinBUGS [13], JAGS [59] and Stan [60].

MCMC methods are a class of algorithm based on the construction of a Markov chain with a stationary distribution given by the target distribution. By observing the chain after a large number of steps using Monte Carlo simulations, one eventually produces samples from the target distribution.

The MCMC implemented in the WinBUGS software combines the celebrated Metropolis-Hastings algorithm with Gibbs Sampling, resulting in what is called 'Metropolis-in-Gibbs sampling'. We here briefly outline the main principles.

2.3.4.1 Metropolis-Hastings Algorithm

We discuss this topic in a more general sense (independent of NMA) and write the distribution from which we would like to sample as $p(\boldsymbol{x})$. The Metropolis-Hastings algorithm [61] is based on a Markov chain producing a trajectory $\boldsymbol{x}^t$, $t = 0, 1, 2, \ldots$. Each step consists of proposing a value for $\boldsymbol{x}^{t+1}$, followed by a decision whether to accept or reject this proposed value. The distribution of proposed values and the acceptance criterion only depend on $\boldsymbol{x}^t$, but not on states visited earlier in the sequence. They are constructed such that the resulting process has stationary distribution $p(\boldsymbol{x})$ [62]. Provided one allows for a sufficiently long equilibration time t_{eq}, the set $\{\boldsymbol{x}^t : t > t_{\text{eq}}\}$ represents a statistically faithful sample of the distribution $p(\boldsymbol{x})$.

The Metropolis-Hastings Algorithm:

1. Initialise $t = 0$, and $\boldsymbol{x}^t = \boldsymbol{x}^0$ for some starting value $\boldsymbol{x}^0$.

2. Generate a proposed value $\boldsymbol{x}'$ from the proposal distribution $q(\boldsymbol{x}'|\boldsymbol{x}^t)$.

3. Calculate the acceptance probability

$$p_a(\boldsymbol{x}'|\boldsymbol{x}^t) = \min\left(1, \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}^t)} \frac{q(\boldsymbol{x}^t|\boldsymbol{x}')}{q(\boldsymbol{x}'|\boldsymbol{x}^t)}\right). \tag{2.23}$$

   Then accept the proposal with probability $p_a(\boldsymbol{x}'|\boldsymbol{x}^t)$, i.e. set $\boldsymbol{x}^{t+1} = \boldsymbol{x}'$. Alternatively, with probability $1 - p_a(\boldsymbol{x}'|\boldsymbol{x}^t)$ reject the proposed update, and set $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t$.

4. Increment time by one, and go to 2.

The algorithm results in an overall probability $A(\boldsymbol{x}|\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})p_a(\boldsymbol{x}|\boldsymbol{y})$ to transition to $\boldsymbol{x}$ if the chain is currently at $\boldsymbol{y}$. Using the fact that exactly one of $p_a(\boldsymbol{x}|\boldsymbol{y})$ and $p_a(\boldsymbol{y}|\boldsymbol{x})$ is equal to one for each pair of states $\boldsymbol{x}$ and $\boldsymbol{y}$, one has $p(\boldsymbol{x})A(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y})A(\boldsymbol{x}|\boldsymbol{y})$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e. the detailed balance condition holds. This is sufficient to demonstrate that $p(\boldsymbol{x})$ is indeed a stationary distribution of the process. We also need to choose the hopping kernel q such that the stationary distribution is unique (i.e. the Markov chain must be irreducible and aperiodic). A sufficient condition to ensure this, is that q is positive everywhere [63]. For further details see also [64, 65].

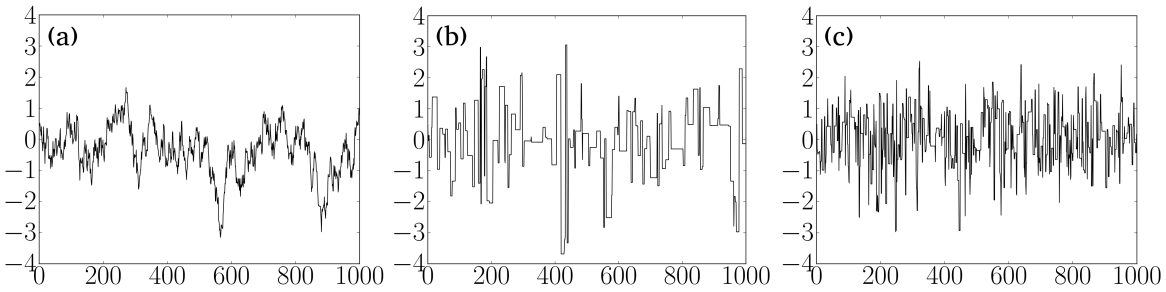


Figure 2.5: (a) Sample path of a Markov chain with small proposal variance and high acceptance rate. (b) Sample path with high proposal variance and low acceptance rate. (c) An efficient Markov chain with proposal variance tuned to obtain a 'reasonable' acceptance rate. The example is for a standard normal distribution $p(x)$. In panel (a) the standard deviation of the hopping kernel is 0.25, resulting in an acceptance rate of 0.925, panel (b) is for a standard deviation of 10 (acceptance rate 0.129), and panel (c) is for a standard deviation of 2.5 (acceptance rate 0.43). The optimal acceptance rate for a model in one dimension is approximately 0.44 [66].

The process simplifies if the proposal function (the hopping kernel) is symmetric, $q(\boldsymbol{x}'|\boldsymbol{x}^t) = q(\boldsymbol{x}^t|\boldsymbol{x}')$. In this case the acceptance probability in step 3 of the algorithm becomes

$$p_a(\boldsymbol{x}'|\boldsymbol{x}^t) = \min\left(1, \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}^t)}\right). \tag{2.24}$$

An example of a symmetric proposal function $q(x'|x^t)$ in a univariate system ($x^t \in \mathbb{R}$) is a normal distribution centred on the current sample value x^t with a fixed variance [67]. The choice of variance is not straightforward. Choosing a value that is too high means that most proposed values are rejected and the sequence of $\{x^t\}$ remains constant for long periods of time. This scenario is illustrated in Figure 2.5 (b). On the other hand, if the variance is too small the chain does not explore the state space and convergence to the stationary state is slow (see Figure 2.5 (a)) [68]. Both of these scenarios make the MCMC less efficient and mean that more iterations are required for the chain to reach the stationary state. If the proposal variance is tuned so that the chain has a 'reasonable' acceptance rate then the state space is explored efficiently. A chain with this characteristic is shown in Figure 2.5 (c).

For an n-dimensional target distribution the optimal acceptance rate has been found to be approximately 0.44 for $n = 1$ and declines to 0.23 as $n \to \infty$ [66, 68]. In the WinBUGS software, the acceptance rate of proposed values is tuned to between 0.2 and 0.4 [13].

2.3.4.2 Gibbs Sampling

Gibbs samplers are used to sample from multivariate distributions. They update one variable at a time, and are used when sampling from conditional probabilities of individual variables is easier than direct sampling from the multivariate distribution. The algorithm cycles through the individual variables, and samples from the conditional distribution of one variable in the target distribution given the current values of all other variables [67, 69]. This can be shown to generate a sequence of multivariate samples faithfully representing the joint distribution [69].

Here we describe details of the Gibbs sampling procedure for a target distribution $p(\boldsymbol{x})$ of n variables, where $\boldsymbol{x} = (x_1, x_2, ..., x_n)$. The conditional probability distribution for the variable x_i given all other variables is given by

$$p_i(x_i|\boldsymbol{x}_{-i}) = \frac{p(x_1, \ldots, x_n)}{p_{-i}(\boldsymbol{x}_{-i})}. \tag{2.25}$$

We have written $\boldsymbol{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ (i.e, $\boldsymbol{x}_{-i}$ is obtained from $\boldsymbol{x}$ by removing the i-th entry). The expression $p_{-i}(\boldsymbol{x}_{-i}) = \int dx_i\, p(x_1, \ldots, x_n)$ in the denominator is the marginal distribution for $\boldsymbol{x}_{-i}$.

The Gibbs Sampling Algorithm:

1. Initialise time at $t = 0$, and set $\boldsymbol{x}^{t=0}$ to a starting value.

2. Update entries of $\boldsymbol{x}$ in turn:

   First, sample x_1^{t+1} from the conditional distribution $p_1(\cdot|x_2^t, \ldots, x_n^t)$.

   Then sample x_2^{t+1} from $p_2(\cdot|x_1^{t+1}, x_3^t, \ldots, x_n^t)$.

   Then sample x_3^{t+1} from $p_3(\cdot|x_1^{t+1}, x_2^{t+1}, x_4^t \ldots, x_n^t)$.

   ...

   Then sample x_{n-1}^{t+1} from $p_{n-1}(\cdot|x_1^{t+1}, \ldots, x_{n-2}^{t+1}, x_n^t)$.

   Finally, sample x_n^{t+1} from $p_n(\cdot|x_1^{t+1}, \ldots, x_{n-1}^{t+1})$.

   At the end of this process $\boldsymbol{x}^{t+1}$ is available.

3. Increment time by one, and go to 2

2.3.4.3 Metropolis-in-Gibbs

In NMA we are interested in samples from the distribution $U(\mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\theta}) = P_{\text{out}}(\mathbf{D}|\boldsymbol{\nu}) \times P_{\text{in}}(\boldsymbol{\nu}|\boldsymbol{\theta})P(\boldsymbol{\theta})$ in Equation (2.17), for a fixed $\mathbf{D}$. To generate these samples we use the 'Metropolis-in-Gibbs' algorithm. This is a Gibbs sampling algorithm with a Metropolis-Hastings accept/reject step used to sample from the conditional distributions for the individual variables, i.e. we use the Metropolis-Hastings algorithm at each single-variable stage of the Gibbs Sampling algorithm.

Metropolis-in-Gibbs for a random effects NMA:

1. Initialise time at $t = 0$ and initialise the parameters: $\Delta^0_{i,\ell}$, $p^0_{i,1}$, τ^0 and $d^0_{T_1 a}$ for $i = 1, \ldots, M$, $\ell = 1, \ldots, m_i$ and $a = T_2, \ldots, T_N$

2. Update the trial-specific treatment effects $\Delta_{i,\ell}$. For each study $i = 1, 2, ..., M$ and each $\ell = 2, ..., m_i$:

   (a) Draw $\Delta'_{i,\ell}$ from the normal distribution $\mathcal{N}(\Delta^t_{i,\ell}, v_\Delta)$.

   (b) Set
   $$\Delta^{t+1}_{i,\ell} = \begin{cases} \Delta'_{i,\ell} & \text{with probability } p_\Delta \\ \Delta^t_{i,\ell} & \text{with probability } 1 - p_\Delta, \end{cases} \tag{2.26}$$
   where
   $$p_\Delta = \min\left(1, \frac{U(\Delta'_{i,\ell}|\Delta^{t+1}_{i,2}, ..., \Delta^{t+1}_{i,\ell-1}, \Delta^t_{i,\ell+1}, ..., \Delta^t_{i,m_i}, p^t_{i,1}, \boldsymbol{d}^t, \tau^t, \mathbf{r})}{U(\Delta^t_{i,\ell}|\Delta^{t+1}_{i,2}, ..., \Delta^{t+1}_{i,\ell-1}, \Delta^t_{i,\ell+1}, ..., \Delta^t_{i,m_i}, p^t_{i,1}, \boldsymbol{d}^t, \tau^t, \mathbf{r})}\right). \tag{2.27}$$
   NB: The parameters $\Delta_{i,\ell}$, specific to trial i, are independent (under the distribution U) of parameters that are specific to other trials $j \neq i$.

3. Update the trial-specific baselines. For each trial $i = 1, ..., M$

   (a) Draw $\operatorname{logit} p'_{i,1}$ from $\mathcal{N}(\operatorname{logit} p^t_{i,1}, v_b)$. From this obtain $p'_{i,1}$.

   (b) Set
   $$p^{t+1}_{i,1} = \begin{cases} p'_{i,1} & \text{with probability } p_b \\ p^t_{i,1} & \text{with probability } 1 - p_b, \end{cases} \tag{2.28}$$
   where
   $$p_b = \min\left(1, \frac{U(p'_{i,1}|\boldsymbol{\Delta}^{t+1}_i, \boldsymbol{d}^t, \tau^t, \boldsymbol{r})}{U(p^t_{i,1}|\boldsymbol{\Delta}^{t+1}_i, \boldsymbol{d}^t, \tau^t, \boldsymbol{r})}\right). \tag{2.29}$$

We note that $p_{i,1}$ is independent of all $p_{j,1}$, $j \neq i$ under the distribution U.

4. Update the heterogeneity parameter:

   (a) Draw τ' from $\mathcal{N}(\tau^t, v_\tau)$.

   (b) Set

$$\tau^{t+1} = \begin{cases} \tau' & \text{with probability } p_\tau \\ \tau^t & \text{with probability } 1 - p_\tau, \end{cases} \tag{2.30}$$

   where

$$p_\tau = \min\left(1, \frac{U(\tau'|\boldsymbol{\Delta}^{t+1}, \boldsymbol{p}_1^{t+1}, \boldsymbol{d}^t, \boldsymbol{r})}{U(\tau^t|\boldsymbol{\Delta}^{t+1}, \boldsymbol{p}_1^{t+1}, \boldsymbol{d}^t, \boldsymbol{r})}\right). \tag{2.31}$$

   We note that the acceptance probability p_τ in step (b) is zero by construction if $\tau' < 0$ [see Equation (2.18)].

5. Update the basic parameters. For each treatment $a = T_2, ..., T_N$

   (a) Draw d'_{T_1a} from $\mathcal{N}(d^t_{T_1a}, v_d)$.

   (b) Set

$$d^{(t+1)}_{T_1a} = \begin{cases} d'_{T_1a} & \text{with probability } p_d \\ d^t_{T_1a} & \text{with probability } 1 - p_d, \end{cases} \tag{2.32}$$

   where

$$p_d = \min\left(1, \frac{U(d'_{T_1a}|\boldsymbol{\Delta}^{t+1}, \boldsymbol{p}_1^{t+1}, \tau^{t+1}, d^{t+1}_{T_1T_2}, ..., d^{t+1}_{T_1T_{\alpha-1}}, d^t_{T_1T_{\alpha+1}}, ..., d^t_{T_1T_N}, \boldsymbol{r})}{U(d^t_{T_1a}|\boldsymbol{\Delta}^{t+1}, \boldsymbol{p}_1^{t+1}, \tau^{t+1}, d^{t+1}_{T_1T_2}, ..., d^{t+1}_{T_1T_{\alpha-1}}, d^t_{T_1T_{\alpha+1}}, ..., d^t_{T_1T_N}, \boldsymbol{r})}\right), \tag{2.33}$$

   with $a = T_\alpha$ and $\alpha = 2, \ldots, N$.

6. Increment time from t to $t + 1$. Go to 2

We note that the acceptance probabilities p in each step of the algorithm are of the form

$$p = \min\left(1, \frac{U(\text{parameter}'|\text{other parameters, data})}{U(\text{parameter}^t|\text{other parameters, data})}\right), \tag{2.34}$$

where parameter$'$ is the proposed value for the model or nuisance parameter that is being updated, and paramater^t its value in the previous iteration. Crucially, the 'other parameters' and the data in the numerator and denominator in Equation (2.34) are

the same. Using the definition $P(A|B) = P(A,B)/P(B)$ of conditional probabilities we can then write

$$p = \min\left(1, \frac{U(\text{parameter}', \text{other parameters, data})}{U(\text{parameter}^t, \text{other parameters, data})}\right). \tag{2.35}$$

This means that we can use joint probabilities (or probability densities) instead of conditional probabilities.

Using the product form of the distribution U in Equation (2.17) and the fact that P_{in}, P_{out} and the prior further factorise, the ratios in Equation (2.35) can be simplified even more by cancelling factors that do not depend on the parameter that is being updated.

2.3.4.4 Assessing Convergence

The MCMC dynamics define a stochastic process with a stationary probability distribution for the model parameters. The process is constructed such that this stationary distribution is the target distribution we set out to sample from. Formally, the stationary distribution is reached only at infinite time, but in practice samples are effectively drawn from the target distribution after sufficiently many iterations. In simulations, we therefore discard the samples of the first n_c iterations (this is referred to as the 'burn-in' in the statistics community, physicists know this as equilibration time or transient). We then make inferences about our parameters based on samples taken in the subsequent iterations.

To obtain accurate inferences on the parameter values we must, therefore, assess the number of iterations required to reach stationarity. A common method to assess this was developed by Gelman and Rubin [70] and later modified by Brooks and Gelman [71]. The latter article gives a detailed description of the approach, we summarise the main ideas here.

A Markov chain has converged (reached stationarity) when the statistics of the samples taken do not depend on the distribution of initial conditions for the process. The Brooks-Gelman-Rubin approach is therefore based on assessing the similarity of samples (more precisely, distributions of samples) obtained from multiple independent chains (realisations of the process) with different starting points.

Assume a target distribution for a scalar parameter with mean μ and variance σ^2.

We now consider m realisations of the process, $i = 1, \ldots, m$, with a set of over-dispersed (i.e, widely spread compared to the expected scale of the parameter) starting values. Each realisation is run for a burn-in of n iterations, followed by another n iterations during which samples are taken. This generates m sets of samples of the parameter. The sample mean of each realisation i provides an estimate $\hat{\mu}_i$ for the mean μ. We then have m inferences about the parameter μ from the m chains. The variance between samples *within* realisation i is labelled v_i (the 'within-chain variance').

We can also obtain an inference about μ from the combined set of mn samples from all realisations. The overall mean of this sample $\hat{\mu}$ is the mean of the $\hat{\mu}_i$. One can then construct a measure of the so-called 'pooled variance' $\hat{V}$ that accounts for the average within-chain variance, the variance in the value of $\hat{\mu}_i$ between realisations, and the sampling variability. For details of this procedure see [71].

As the number of iterations n (before and after burn-in) increases, we expect the value of the pooled variance and the average within-chain variance to stabilise and for these variances to converge to the same value [70]. To assess the number of iterations required for convergence we split each chain (total length $2n$) into n/b batches of length $2b$.

For a given integer $k = 1, \ldots, n/b$ we then calculate the pooled variance, the average within-chain variance and their ratio $\hat{R}$ based on the samples in the first k batches, where the first half of this data is discarded as burn-in. I.e. for $k = 1$, we use the first batch (length $2b$), discard the first b iterations of it, and compute the variances and their ratio based on iterations $b + 1, ..., 2b$. For $k = 2$, we use the first $4b$ iterations (two batches), but again discard the first half of this, and compute the variances from iterations $2b + 1, \ldots, 4b$. For higher values of k we proceed analogously.

Plotting the variances and their ratio against k (or $2kb$) as shown in Figure 2.6, we can assess the approximate number of iterations required for the variances to stabilise and for $\hat{R}$ to be sufficiently close to unity so that the chain can be assumed to have converged [71]. More recent refinements of this method are discussed in [69].

Once the Markov chain has reached stationarity, we make inferences about the model parameters from the samples taken in the simulations. Usually this means calculating a central value of the distribution of samples (such as the mean or median) and some measure of spread (such as standard deviation). We discuss how exactly the

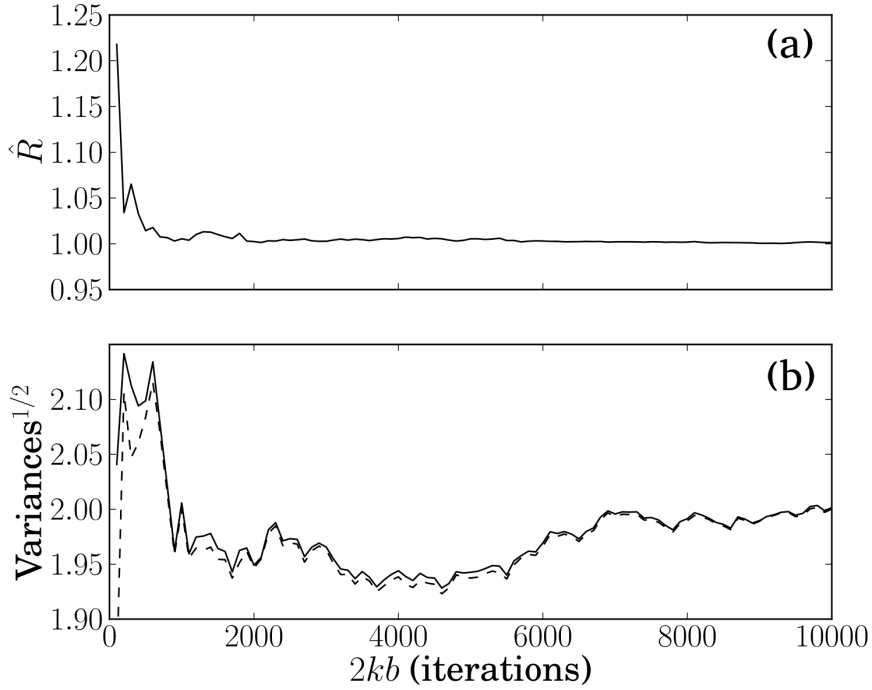


Figure 2.6: Examples of Brooks-Gelman-Rubin convergence plots with $m = 5$ chains and batch lengths of $b = 50$. (a) The ratio $\hat{R}$ of the pooled variance and the average within-chain variance against the number of iterations. (b) The solid line shows (the square root of) the pooled variance $\hat{V}$ as a function of the number of iterations. The dotted line is (the square root of) the average within-chain variance. For this example convergence is reached after approximately $2kb = 7000$ iterations (or a burn-in of 3500).

results of the NMA are reported in Section 2.5.

2.4 Frequentist Network Meta-Analysis

We now move on to the frequentist approach to network meta-analysis. In this section we use the contrast-based variant of NMA and treat the relative treatment effects measured in each trial as the raw data in the model. For binomial outcomes in each trial arm, these are the log odds ratios $y_{i,1\ell}$.

This section is structured as follows: In Section 2.4.1 we introduce the contrast-based NMA model and write it as a linear regression problem. Then, in Section 2.4.2 we discuss the frequentist approach in more general terms and describe how we estimate regression coefficients in a linear regression model. We start by explaining the ordinary least squares (OLS) method and then use this to derive the generalised least squares (GLS) problem. We then explain the maximum likelihood (ML) approach and show that this leads to the same condition as GLS. We solve the GLS/ML problem to obtain

the so-called 'Aitken estimator' of the regression coefficients. In Section 2.4.3 we go back to the NMA model. First, we use the Aitken estimator to estimate the relative treatment effects $\boldsymbol{d}$ (Section 2.4.3.1). Finally, we explain two common methods for estimating the heterogeneity variance τ^2 (Section 2.4.3.2).

2.4.1 Introduction and notation

We begin by outlining the assumptions of the contrast-based NMA model and writing the inference task as a linear regression problem.

As in previous sections, the vector of relative treatment effects in each trial is assumed to follow a normal distribution. We recall Equation (2.6), which we can write as

$$\Delta_{i,1\ell} = d_{t_{i,1},t_{i,\ell}} + \eta_{i,1\ell}, \tag{2.36}$$

where $i = 1, \ldots, M$, $\ell = 1, \ldots, m_i$ and $\eta_{i,1\ell}$ is a Gaussian random variable of mean zero. Using transitivity [Equation (2.4)] the mean $d_{t_{i,1},t_{i,\ell}}$ can be constructed via a linear relation from the vector of basic parameters $\boldsymbol{d} = (d_{T_1,T_2}, \ldots, d_{T_1,T_N})^\top$ [cf. Equation (2.7)]. The variance of $\eta_{i,1\ell}$ is given by the heterogeneity τ^2, and we note that for a fixed i the different $\eta_{i,1\ell}$, $\ell = 1, \ldots, m_i$ will in general be correlated [see Equation (2.6)]. We can collect the relations in Equation (2.36) for all trials i and all basic comparisons within each trial, and write more compactly

$$\boldsymbol{\Delta} = \mathbf{X}\boldsymbol{d} + \boldsymbol{\eta}. \tag{2.37}$$

Here, $\mathbf{X}$ is the design matrix of the *network* which can be constructed from the trial-specific design matrices described in Section 2.2.4.3, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)^\top$.

The matrix $\mathbf{X}$ has $N-1$ columns and $\sum_{i=1}^{M}(m_i - 1)$ rows, and each entry is either -1, 0 or 1. Each column of $\mathbf{X}$ represents one of the treatments $T_2, ..., T_N$ (treatment T_1 is the overall baseline). The rows represent comparisons to the trial-specific baseline in each study.

As before, we write $y_{i,1\ell}$ for the *observed* relative effects in each trial [e.g. the log odds ratios in Equation (2.12)]. These are assumed to follow a normal distribution centred on the mean value $\Delta_{i,1\ell}$ with some random sampling error, $\epsilon_{i,1\ell}$. That is,

$$y_{i,1\ell} = \Delta_{i,1\ell} + \epsilon_{i,1\ell}, \tag{2.38}$$

where the sampling errors $\epsilon_{i,1\ell}$ within a trial are correlated. Trial $i \in \{1, \ldots, M\}$ compares m_i treatments and therefore contributes $m_i - 1$ relative treatment effects (comparisons of treatments $\ell = 2, \ldots, m_i$ to the trial-specific baseline treatment).

Collecting the $\sum_{i=1}^{M}(m_i - 1)$ observations $y_{i,1\ell}$ in the vector $\boldsymbol{y}$, we can write the linear model as

$$\boldsymbol{y} = \boldsymbol{\Delta} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{d} + \boldsymbol{\eta} + \boldsymbol{\epsilon}. \tag{2.39}$$

The vectors $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ represent the two levels of stochasticity in the RE model described in Section 2.2.4; $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ models the trial-to-trial variation of relative treatment effects, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ models the noise on the observed relative treatment effects resulting from the sampling in the trial arms. We recall that we are working within a contrast-based model, where the sampling noise in the trial arms is assumed to be Gaussian. The covariance matrices for the two types of stochasticity are $\boldsymbol{\Sigma}$ and $\mathbf{V}$ respectively, we will discuss their mathematical form below. The two types of noise are independent of each other, and the overall covariance matrix is then $\mathbf{C} = \boldsymbol{\Sigma} + \mathbf{V}$, such that the model in Equation (2.39) can be written as

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{d}, \boldsymbol{\Sigma} + \mathbf{V}). \tag{2.40}$$

The covariance matrix associated with the sampling errors in the trials is of block diagonal form, $\mathbf{V} = \mathrm{diag}(\mathbf{V}_i)$, where each trial i contributes an $(m_i - 1) \times (m_i - 1)$ matrix,

$$\mathbf{V}_i = \begin{pmatrix} \sigma^2_{i,12} & \mathrm{Cov}(y_{i,12}, y_{i,13}) & \ldots & \mathrm{Cov}(y_{i,12}, y_{i,1m_i}) \\ \mathrm{Cov}(y_{i,13}, y_{i,12}) & \sigma^2_{i,13} & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(y_{i,1m_i}, y_{i,12}) & \ldots & \ldots & \sigma^2_{i,1m_i} \end{pmatrix}. \tag{2.41}$$

We stress that this describes sampling errors only, i.e. the matrix entries are the variances of the components $\epsilon_{i,1\ell}$, $\ell = 2, \ldots, m_i$, and the correlations between these variables. The measurements of relative treatment effects within a multi-arm trial are correlated because they involve a common treatment arm (the trial-specific baseline treatment). The values that make up the matrices $\mathbf{V}_i$ are assumed to be known (i.e. they are reported in the study, or can be directly calculated from the data - see Section 2.9 of the Appendix for details). Further details can also be found in [5, 7, 17, 28, 72, 73].

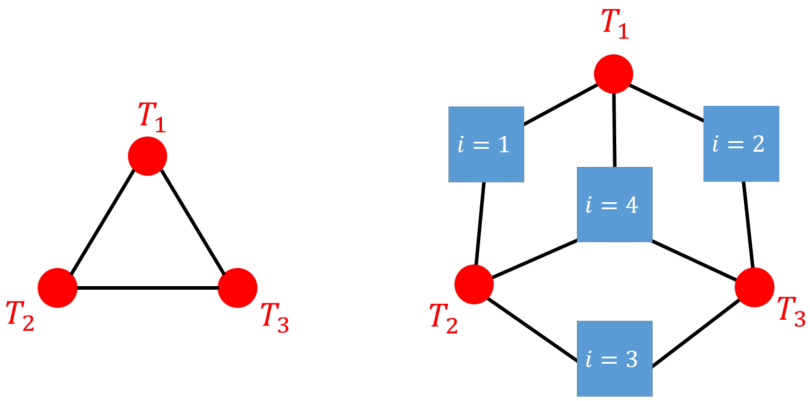


Figure 2.7: A fictional example of a network meta-analysis of $N = 3$ treatments, $\{T_1, T_2, T_3\}$, and $M = 4$ trials. (a) Standard network representation, all treatments are included in three trials (each pair of treatments appears in two trials). (b) Representation as a bipartite graph indicating which treatments are compared in each trial $i = 1, \ldots, 4$.

The covariance matrix associated with the random effects $\boldsymbol{\Sigma}$ represents the heterogeneity *between* trials. Similarly to $\mathbf{V}$, it has block diagonal form, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_i)$, where the blocks are the $(m_i - 1) \times (m_i - 1)$ matrices $\boldsymbol{\Sigma}_i$ defined in Section 2.2.4.3. The diagonal elements of $\boldsymbol{\Sigma}_i$ are the variances associated with the random effects and the off diagonal elements relate to the correlations between the random effects within a multi-arm trial. We assume that these are determined by the unknown heterogeneity variance, τ^2, as described in Section 2.2.4.3. Determining τ^2 is therefore part of the inference problem.

Example - NMA as a linear regression model:

Consider a network of $M = 4$ trials comparing $N = 3$ treatments, T_1, T_2, T_3. Trials $i = 1, 2, 3$ are two arm trials comparing (T_1, T_2), (T_1, T_3) and (T_2, T_3) respectively. Trial $i = 4$ is a three-arm trial comparing all three treatments. This network is shown in Figure 2.7. The regression model is then

$$\begin{pmatrix} y_{1,T_1T_2} \\ y_{2,T_1T_3} \\ y_{3,T_2T_3} \\ y_{4,T_1T_2} \\ y_{4,T_1T_3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} d_{T_1T_2} \\ d_{T_1T_3} \end{pmatrix} + \begin{pmatrix} \eta_{1,T_1T_2} \\ \eta_{2,T_1T_3} \\ \eta_{3,T_2T_3} \\ \eta_{4,T_1T_2} \\ \eta_{4,T_1T_3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,T_1T_2} \\ \epsilon_{2,T_1T_3} \\ \epsilon_{3,T_2T_3} \\ \epsilon_{4,T_1T_2} \\ \epsilon_{4,T_1T_3} \end{pmatrix}$$

where

$$\begin{pmatrix} \epsilon_{1,T_1T_2} \\ \epsilon_{2,T_1T_3} \\ \epsilon_{3,T_2T_3} \\ \epsilon_{4,T_1T_2} \\ \epsilon_{4,T_1T_3} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{1,T_1T_2} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{2,T_1T_3} & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{3,T_2T_3} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{4,T_1T_2} & \mathrm{Cov}(y_{4,T_1T_2}, y_{4,T_1T_3}) \\ 0 & 0 & 0 & \mathrm{Cov}(y_{4,T_1T_3}, y_{4,T_1T_2}) & \sigma^2_{4,T_1T_3} \end{pmatrix} \right),$$

where the matrix entries are the variances and covariances of the $\epsilon_{i,1\ell}$. We also have

$$\begin{pmatrix} \eta_{1,T_1T_2} \\ \eta_{2,T_1T_3} \\ \eta_{3,T_2T_3} \\ \eta_{4,T_1T_2} \\ \eta_{4,T_1T_3} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 & 0 & 0 & 0 \\ 0 & \tau^2 & 0 & 0 & 0 \\ 0 & 0 & \tau^2 & 0 & 0 \\ 0 & 0 & 0 & \tau^2 & \tau^2/2 \\ 0 & 0 & 0 & \tau^2/2 & \tau^2 \end{pmatrix} \right).$$

The aim of NMA is to estimate the unknown parameters $\boldsymbol{d}$ and τ^2. In the language of regression models, $\boldsymbol{d}$ are the 'regression coefficients' of the linear model and τ^2 is the 'variance parameter'. In frequentist NMA, the variance parameter is estimated first, then this estimate is used in the estimate of the regression coefficients. In the following we describe two frequentist approaches for estimating the regression coefficients assuming knowledge of the variance parameter. We then relate this to the contrast-based NMA model. Finally, we describe some common methods for estimating the variance parameter.

2.4.2 General frequentist approach

Two frequentist approaches to inferring the regression coefficients of a model are based on 'maximum likelihood' and 'least squares'.

In the maximum likelihood (ML) approach one finds the values of the parameters that maximise the likelihood or, equivalently, minimise the negative log likelihood. In least squares regression, we start from Equation (2.39). The vector $\boldsymbol{y}$ is observed from the trials, and the matrix $\mathbf{X}$ is known from the design of the trials (what treatments are tested in each trial). We therefore wish to find the vector $\boldsymbol{d}$ that best fits the

observed data via Equation (2.39). To do this a 'residual' is defined as the difference between the observed value of the response variable (here $\boldsymbol{y}$) and the mean value $\mathbf{X}\boldsymbol{d}$ predicted by the regression model. The model parameters (here $\boldsymbol{d}$) are then estimated by minimising the sum of the squared residuals, i.e. we find the values of the model parameters that 'best fit' the data. When measurements are associated with correlated random errors we must use so-called 'generalised least squares' (GLS) regression [74]. This will be explained in more detail later. For a linear regression model under the assumption of normally distributed errors, the ML estimates and GLS estimates are equivalent [75, 76] and can be found analytically.

We now derive these estimates using the GLS procedure and show that this is equivalent to obtaining the maximum likelihood estimates.

2.4.2.1 Ordinary least squares problem

We first describe this in general terms, and discuss the application to NMA further below. Assume we observe data $\{y_i, x_{i,j}\}$ on n statistical units such that $i = 1, \ldots, n$ and $j = 1, \ldots, p$. In the context of an NMA these would be the relative treatment effects and the design matrix elements. The latter are treated as part of the 'data' for the discussion in this section, as they are specific to individual instances of a real-world NMA. More generally, $\mathbf{X}$ may contain a set of observed covariates.

The values of the response variable are collected in the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$. The predictor variables are placed in the $n \times p$ design matrix $\mathbf{X} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_n^\top)^\top$, where $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^\top$ and $\mathbf{X}$ is assumed to have full rank (in NMA this is true by construction). We consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.42}$$

where $\boldsymbol{\beta}$ is a column vector of length p containing the parameters that we wish to estimate. The error term is assumed to be normally distributed, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{C}(\boldsymbol{\phi}))$, with an $n \times n$ covariance matrix $\mathbf{C}(\boldsymbol{\phi})$ that depends on a set of parameters $\boldsymbol{\phi}$. For convenience we will write $\mathbf{C}$ for $\mathbf{C}(\boldsymbol{\phi})$. The parameters $\boldsymbol{\beta}$ are the regression coefficients of the model while $\boldsymbol{\phi}$ represents the variance parameters. In the simplest case we assume uncorrelated errors and equal variances such that $\mathbf{C}$ is a multiple of the identity matrix. This assumption is known as the ordinary least squares (OLS) condition.

The vector of residuals, $\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}$, represents the difference between the observed outputs, $\boldsymbol{y}$, and the mean values predicted by the model in Equation (2.42). The ordinary least squared estimates of the parameters are then obtained by minimising the sum of the squared residuals,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} [(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})_i]^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right]. \end{aligned} \tag{2.43}$$

2.4.2.2 Generalised least squares problem

In generalised least squares regression we relax the ordinary least squares assumption. That is, we make no assumptions on the form of the covariance matrix $\mathbf{C}$. If errors are uncorrelated but do not necessarily have equal variances then $\mathbf{C}$ is a diagonal matrix. Regression under these conditions is a special case of GLS known as 'weighted least squares'. If errors are correlated then $\mathbf{C}$ has non-zero off diagonal elements representing the covariance between error terms.

To find the generalised least squares estimator we carry out a transformation of the GLS model so that it fulfils the ordinary least squares condition. To this end we note that given that $\mathbf{C}$ is a covariance matrix, it must be symmetric and positive-definite. Therefore we can write $\mathbf{C} = \mathbf{K}^\top \mathbf{K} = \mathbf{K}\mathbf{K}$ where $\mathbf{K}$ is the (symmetric) square root of $\mathbf{C}$ [74]. We now multiply both sides of Equation (2.42) with $\mathbf{K}^{-1}$ from the left,

$$\mathbf{K}^{-1}\boldsymbol{y} = \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^{-1}\boldsymbol{\epsilon}. \tag{2.44}$$

Defining the variables

$$\tilde{\boldsymbol{y}} = \mathbf{K}^{-1}\boldsymbol{y}, \qquad \tilde{\mathbf{X}} = \mathbf{K}^{-1}\mathbf{X}, \qquad \tilde{\boldsymbol{\epsilon}} = \mathbf{K}^{-1}\boldsymbol{\epsilon} \tag{2.45}$$

we obtain the model

$$\tilde{\boldsymbol{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}. \tag{2.46}$$

Now let us inspect the error term $\tilde{\boldsymbol{\epsilon}}$ of this model to see if it fulfils the OLS condition. The expected value of the error is $\mathbb{E}(\tilde{\boldsymbol{\epsilon}}) = \mathbb{E}(\mathbf{K}^{-1}\boldsymbol{\epsilon}) = \mathbf{K}^{-1}\mathbb{E}(\boldsymbol{\epsilon}) = 0$, as required. To obtain the covariance matrix, we use the relation $\mathrm{Cov}(\mathbf{A}\boldsymbol{z}) = \mathbf{A}\mathrm{Cov}(\boldsymbol{z})\mathbf{A}^\top$ which is

valid for any random vector $\boldsymbol{z}$ and fixed matrix $\mathbf{A}$. Therefore,

$$\begin{aligned}\text{Cov}(\tilde{\boldsymbol{\epsilon}}) &= \text{Cov}(\mathbf{K}^{-1}\boldsymbol{\epsilon}) = \mathbf{K}^{-1}\text{Cov}(\boldsymbol{\epsilon})(\mathbf{K}^{-1})^\top \\ &= \mathbf{K}^{-1}\mathbf{C}\mathbf{K}^{-1} = \mathbf{I},\end{aligned} \tag{2.47}$$

where we have used $\text{Cov}(\boldsymbol{\epsilon}) = \mathbf{C}$, $\mathbf{K}^{-1}$ is symmetric, and $\mathbf{C} = \mathbf{K}\mathbf{K}$. The errors $\tilde{\boldsymbol{\epsilon}}$ therefore fulfil the OLS condition, and Equation (2.46) hence defines an ordinary least squares problem.

We now obtain the generalised least squares estimator by using the OLS estimator with Equation (2.46), that is,

$$\hat{\boldsymbol{\beta}}^{\text{GLS}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[(\tilde{\boldsymbol{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^\top(\tilde{\boldsymbol{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})\right]. \tag{2.48}$$

Using the definitions in Equation (2.45) we find

$$\begin{aligned}(\tilde{\boldsymbol{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^\top(\tilde{\boldsymbol{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) &= (\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}))^\top(\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})) \\ &= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{K}^{-1})^\top\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{2.49}$$

The GLS estimator is therefore

$$\hat{\boldsymbol{\beta}}^{\text{GLS}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right]. \tag{2.50}$$

This can be solved analytically as we will see in Section 2.4.2.4. Before we return to this we derive the maximum likelihood estimator and show that this leads to the same condition as in Equation (2.50).

2.4.2.3 Maximum likelihood approach

The linear model in Equation (2.42) with normally distributed errors $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{C})$ can be written equivalently as

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{C}), \tag{2.51}$$

where we place no assumptions on the covariance matrix $\mathbf{C}$ except that it depends on a set of variance parameters $\boldsymbol{\phi}$. The likelihood of this model is then simply the multivariate normal distribution with mean vector $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{C}$,

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y}, \mathbf{X}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{C})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right), \tag{2.52}$$

and the log likelihood is

$$\ln(L(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y}, \mathbf{X})) = -\frac{1}{2}\ln(\det \mathbf{C}) - \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \text{const.} \tag{2.53}$$

Treating the variance parameters as known, we infer the regression coefficients $\boldsymbol{\beta}$ by maximising the (log) likelihood with respect to $\boldsymbol{\beta}$. This is equivalent to minimising the term $(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$,

$$\hat{\boldsymbol{\beta}}^{\text{ML}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right], \tag{2.54}$$

which is identical to the generalised least squares problem in Equation (2.50).

2.4.2.4 Solution to the GLS and ML problem (The Aitken estimator)

Now we proceed solve Equation (2.50) [= Equation (2.54)] to find the estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{GLS}} = \hat{\boldsymbol{\beta}}^{\text{ML}}$. We first multiply out the product $(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$, and then set the partial derivative with respect to $\boldsymbol{\beta}$ equal to zero. This leads to

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{y}^\top \mathbf{C}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{C}^{-1}\boldsymbol{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta}\right) = 0. \tag{2.55}$$

We address this term by term. The first term $\boldsymbol{y}^\top \mathbf{C}^{-1}\boldsymbol{y}$ is independent of $\boldsymbol{\beta}$ and therefore differentiates to zero. The second and third terms take the forms $\boldsymbol{a}^\top\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\top\boldsymbol{a}$ respectively, where $\boldsymbol{a} = -\mathbf{X}^\top \mathbf{C}^{-1}\boldsymbol{y}$ is a column vector of length p. We have $\boldsymbol{a}^\top\boldsymbol{\beta} = \boldsymbol{\beta}^\top\boldsymbol{a}$, and the second and third terms of Equation (2.55) each evaluate to $-\mathbf{X}^\top \mathbf{C}^{-1}\boldsymbol{y}$. The last term on the right-hand side of Equation (2.55) is quadratic in $\boldsymbol{\beta}$. We also note that the matrix $\mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}$ is symmetric. Therefore the final term in Equation (2.55) evaluates to $2\mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta}$.

Combining these results, Equation (2.55) reduces to

$$-2\mathbf{X}^\top \mathbf{C}^{-1}\boldsymbol{y} + 2\mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta} = 0. \tag{2.56}$$

Solving for $\boldsymbol{\beta}$ yields the GLS and ML estimator of the vector of regression coefficients,

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{GLS}} = \hat{\boldsymbol{\beta}}^{\text{ML}} = \left(\mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{C}^{-1}\boldsymbol{y}, \tag{2.57}$$

also known as the 'Aitken estimator' [77].

Recalling that $\mathbb{E}(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$ [see Equation (2.42)] the expectation of this estimate is

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \tag{2.58}$$

indicating that the Aitken estimator is an unbiased estimate of $\boldsymbol{\beta}$.

To find the $(p \times p)$ covariance matrix of this estimate we once again make use of the result $\mathrm{Cov}(\mathbf{A}\boldsymbol{z}) = \mathbf{A}\mathrm{Cov}(\boldsymbol{z})\mathbf{A}^\top$ and find

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \left[\left(\mathbf{X}^\top\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{C}^{-1}\right]\mathrm{Cov}(\boldsymbol{y})\left[\left(\mathbf{X}^\top\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{C}^{-1}\right]^\top \\ &= \left(\mathbf{X}^\top\mathbf{C}^{-1}\mathbf{X}\right)^{-1},\end{aligned} \tag{2.59}$$

where we have used $\mathrm{Cov}(\boldsymbol{y}) = \mathbf{C}$ and the fact that the matrices $\mathbf{C}$ and $\mathbf{X}^\top\mathbf{C}^{-1}\mathbf{X}$ are symmetric.

2.4.3 Frequentist inference for NMA

In Section 2.4.2 we derived the GLS/ML estimator of the regression coefficients for a linear regression model [Equation (2.57)]. We now use this result to estimate the relative treatment effects $\boldsymbol{d}$ in our NMA model from Section 2.4.1 [Equation (2.39)]. Following this, we discuss frequentist methods of estimating the heterogeneity variance τ^2.

2.4.3.1 Estimating the mean relative treatment effects

We start from the linear regression model for a RE network meta-analysis,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{d} + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V}). \tag{2.60}$$

The within-study covariance matrix $\mathbf{V}$ (describing the statistics of sampling noise) is assumed to be known, whereas the between-study covariance matrix $\boldsymbol{\Sigma}$ depends on the unknown heterogeneity variance τ^2.

Assuming an estimate of the heterogeneity variance $\hat{\tau}^2$ (and therefore the covariance matrix $\hat{\boldsymbol{\Sigma}}$), we find the mean relative treatment effects $\boldsymbol{d}$ via the Aitken estimator in Equation (2.57),

$$\hat{\boldsymbol{d}}^{\mathrm{RE}} = (\mathbf{X}^\top(\mathbf{V} + \hat{\boldsymbol{\Sigma}})^{-1}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{V} + \hat{\boldsymbol{\Sigma}})^{-1}\boldsymbol{y}, \tag{2.61}$$

where we have labelled this explicitly as a random effects (RE) estimate, since the estimator depends on the heterogeneity τ^2 (or an estimate $\hat{\tau}^2$). It is useful to define the inverse-variance weight matrix $\mathbf{W} = (\mathbf{V} + \hat{\boldsymbol{\Sigma}})^{-1}$. We can then write

$$\hat{\boldsymbol{d}}^{\mathrm{RE}} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{y}. \tag{2.62}$$

Using Equation (2.59) the covariance matrix associated with this estimator is given by

$$\mathrm{Cov}(\hat{\boldsymbol{d}}^{\mathrm{RE}}) = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}. \tag{2.63}$$

To obtain the estimator for a fixed effects (FE) model, $\hat{\boldsymbol{d}}^{\mathrm{FE}}$, we simply set $\tau^2 = 0$. We then have $\boldsymbol{\Sigma} = 0$ and hence $\mathbf{W} = \mathbf{V}^{-1}$.

Special case: Pairwise meta-analysis. In a pairwise meta-analysis of $N = 2$ treatments, each trial provides an estimate y_i of the same relative treatment effect (we write d for its true value). The design matrix $\mathbf{X}$ is then an $M \times 1$ matrix, and all entries are equal to one. The covariance matrix $\mathbf{C}$ is an $M \times M$ diagonal matrix with elements equal to $\sigma_i^2 + \tau^2$ and the within-study variances σ_i^2 are assumed known. The RE model is then $y_i \sim \mathcal{N}(d, \sigma_i^2 + \tau^2)$ for $i = 1, \ldots, M$. The weight matrix is also diagonal and, for a given estimate of heterogeneity $\hat{\tau}^2$, its elements are equal to $w_i = (\sigma_i^2 + \hat{\tau}^2)^{-1}$. We then find that the Aitken estimator of the relative treatment effect reduces to

$$\hat{d}^{\mathrm{RE}} = \frac{\sum_{i=1}^M w_i y_i}{\sum_{i=1}^M w_i} = \frac{\sum_{i=1}^M (\sigma_i^2 + \hat{\tau}^2)^{-1} y_i}{\sum_{i=1}^M (\sigma_i^2 + \hat{\tau}^2)^{-1}}. \tag{2.64}$$

The variance of this estimate is

$$\mathrm{Var}(\hat{d}^{\mathrm{RE}}) = \frac{1}{\sum_{i=1}^M w_i} = \frac{1}{\sum_{i=1}^M (\sigma_i^2 + \hat{\tau}^2)^{-1}}. \tag{2.65}$$

Therefore, for pairwise meta-analysis, the GLS and ML approaches recover the results for a simple weighted mean of the sample. Again, for a fixed effect model, $\hat{d}^{\mathrm{FE}}$ is obtained by setting $\tau^2 = 0$, that is, $w_i = \sigma_i^{-2}$.

2.4.3.2 Estimating the heterogeneity variance

So far, we have assumed knowledge of the heterogeneity variance τ^2. We now discuss how this is estimated. There are numerous methods for obtaining a frequentist estimate of τ^2 in pairwise meta-analysis [5, 78–82], and there is much debate over which method is most appropriate [83–85]. Some of these methods have also been extended to network meta-analysis [86–90].

The most widely used methods fall into two categories: (i) the method of moments [91], and (ii) restricted maximum likelihood (REML) approaches [92]. The former involves defining a measure of heterogeneity based on the sum of squared residuals. The latter involves modifying the likelihood function of the random effects model to

remove dependence on the relative treatment effects and then maximising this modified likelihood with respect to τ^2. We describe both approaches in turn.

Method of moments. Multiple heterogeneity estimators in pairwise meta-analysis are based on the so-called 'method-of-moments' [78, 91]. Inference of the regression coefficients in pairwise meta-analysis involves a weighted mean of the sample [Equation (2.64)]. For general weights a_i $(i = 1, \ldots, M)$ associated with observations y_i we define the weighted mean,

$$\hat{y} = \frac{\sum_{i=1}^{M} a_i y_i}{\sum_{i=1}^{M} a_i}. \tag{2.66}$$

If we set $a_i = (\sigma_i^2 + \tau^2)^{-1}$ we recover the random effects estimate $\hat{d}^{\text{RE}}$ in Equation (2.64). Other choices of the weights will be discussed below.

We can then define a generalised version of the so-called 'Q statistic' [5, 78, 83] as the weighted sum of squared residuals,

$$Q = \sum_{i=1}^{M} a_i (y_i - \hat{y})^2. \tag{2.67}$$

The estimate of τ^2 is obtained by assuming that the empirical value of Q obtained via Equation (2.67) from the observed data is equal to its expectation under the random effects model [91]. That is,

$$\sum_{i=1}^{M} a_i (y_i - \hat{y})^2 = \mathbb{E}_{\text{RE}}(Q) \equiv \mathbb{E}_{\text{RE}} \left(\sum_{i=1}^{M} a_i (y_i - \hat{y})^2 \right). \tag{2.68}$$

The expectation on the right is calculated assuming that observations follow the RE model, $y_i \sim \mathcal{N}(d, \sigma_i^2 + \tau^2)$. In Section 2.10.1 of the Appendix we show that

$$\mathbb{E}_{\text{RE}}(Q) = \tau^2 \left(\sum_{i=1}^{M} a_i - \frac{\sum_{i=1}^{M} a_i^2}{\sum_{i=1}^{M} a_i} \right) + \left(\sum_{i=1}^{M} a_i \sigma_i^2 - \frac{\sum_{i=1}^{M} a_i^2 \sigma_i^2}{\sum_{i=1}^{M} a_i} \right). \tag{2.69}$$

Using this in Equation (2.68) and re-arranging for τ^2 we find

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{M} a_i \left(y_i - \frac{\sum_j a_j y_j}{\sum_j a_j} \right)^2 - \left(\sum_{i=1}^{M} a_i \sigma_i^2 - \frac{\sum_{i=1}^{M} a_i^2 \sigma_i^2}{\sum_{i=1}^{M} a_i} \right)}{\sum_{i=1}^{M} a_i - \frac{\sum_{i=1}^{M} a_i^2}{\sum_{i=1}^{M} a_i}}, \tag{2.70}$$

where we have used the definition of $\hat{y}$ in Equation (2.66). This is the general method-of-moments estimator for τ^2 in pairwise meta-analysis. In practice, the expression on the right-hand side of Equation (2.70) can come out negative. In this case one sets $\hat{\tau}^2 = 0$.

Different choices can be made for the weights a_i in Equations (2.66) and (2.67). For example, the widely used DerSimonian and Laird (DL) estimator [5] uses the fixed effect weights, $a_i = \sigma_i^{-2}$, so that $\hat{y} = \hat{d}^{\rm FE}$. Cochran's ANOVA (CA) estimator [93] uses equal weights $a_i = 1/M$ while the Paule Mandel (PM) estimator [82] uses the random effects weights $a_i = (\sigma_i^2 + \tau^2)^{-1}$. The DL and CA estimators lead to a closed form solution for the estimate of τ^2 [in the sense that the right-hand side of Equation (2.70) becomes independent of τ^2]. This is not the case for the PM estimator since these weights depend on τ^2. Equation (2.70) must then be solved numerically.

Extending the DL estimator to the case of network meta-analysis is straightforward [87, 89, 90]. We generalise Equation (2.67) using the inverse-variance weight matrix and obtain

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{V}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{y}}). \tag{2.71}$$

We recall that $\mathbf{V}^{-1}$ represents the observed within-study variances and correlations so the expression in Equation (2.71) is analogous to using $a_i = \sigma_i^{-2}$ in the pairwise case. The vector $\hat{\boldsymbol{y}}$ is the set of network estimates of $\boldsymbol{y}$ obtained using the fixed effects weights $\mathbf{V}^{-1}$. That is

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{d}}^{\rm FE} = \mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \boldsymbol{y}, \tag{2.72}$$

where $\hat{\boldsymbol{d}}^{\rm FE}$ are the estimates of the mean relative treatment effects under the fixed effect model, obtained by setting $\mathbf{W} = \mathbf{V}^{-1}$ (i.e. $\boldsymbol{\Sigma} = 0$) in Equation (2.62). Equations (2.71) and (2.72) are the NMA analogue of Equations (2.67) and (2.66) in pairwise MA (when $a_i = \sigma_i^{-2}$).

In Section 2.10.2 of the Appendix we evaluate the expectation of Q in Equation (2.71) and find

$$\mathbb{E}_{\rm RE}(Q) = \sum_{i=1}^{M} (m_i - 1) - (N - 1) + \tau^2 {\rm tr}(\mathbf{AP}), \tag{2.73}$$

where, following Jackson et al (2016) [89], we have defined the matrix

$$\mathbf{A} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1}. \tag{2.74}$$

We have also defined the block diagonal matrix $\mathbf{P}$ such that $\boldsymbol{\Sigma} = \tau^2 \mathbf{P}$. Each $(m_i - 1) \times (m_i - 1)$ block in $\mathbf{P}$ represents a trial i, and has diagonal elements equal to 1 and off-diagonal elements equal to 1/2. All other elements of $\mathbf{P}$ are zero.

Similar to the pairwise case we equate the expectation in Equation (2.73) with the empirically observed value of Q in Equation (2.71). Re-arranging for τ^2 we find

$$\hat{\tau}^2 = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{V}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{y}}) - [\sum_{i=1}^{M}(m_i - 1) - (N - 1)]}{\mathrm{tr}(\mathbf{AP})}. \tag{2.75}$$

This is the DerSimonian and Laird estimator of τ^2 in network meta-analysis. Again, if $\hat{\tau}^2$ comes out negative we set its value equal to zero.

Restricted maximum likelihood. We now explain the restricted maximum likelihood method for estimating the variance parameter. We start by discussing this method in a more general sense (for a linear regression problem) and then relate this to the NMA model.

In Section 2.4.2.3 we showed that we can obtain estimates for the regression coefficients $\boldsymbol{\beta}$ in a linear model of the form $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{C})$ by maximising the log likelihood in Equation (2.53) with respect to these parameters. This led to the Aitken estimator in Equation (2.57).

The likelihood of this model is also a function of the variance parameters $\boldsymbol{\phi}$ characterising the covariance matrix $\mathbf{C}$. The Aitken estimator assumes that these are known, which is generally not the case. One way of obtaining the variance parameters is to maximise the log likelihood simultaneously with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$.

A problem with this approach is that the resulting estimates of the variance parameters are generally biased [92]. This is because the variance estimate fails to account for the loss in degrees of freedom that results from estimating $\boldsymbol{\beta}$ [94]. This is well known, and can be demonstrated easily for a one dimensional normal distribution, see Section 2.11 of the Appendix.

The restricted maximum likelihood (REML) approach was proposed by Patterson and Thompson [92] as a method to overcome this problem. The principal idea is to carry out a linear transformation of the variables $\boldsymbol{y}$ so that the likelihood function for the transformed variables no longer depends on the parameters $\boldsymbol{\beta}$, but only on their estimates $\hat{\boldsymbol{\beta}}$ (which in turn depend on $\boldsymbol{\phi}$). This 'restricted likelihood' is then maximised with respect to the variance parameters.

The restricted likelihood is given by

$$RL(\boldsymbol{\phi}|\boldsymbol{y}, \mathbf{X}) \propto (\det \mathbf{C})^{-1/2} (\det \mathbf{X}^\top \mathbf{C}^{-1} \mathbf{X})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{C}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right). \tag{2.76}$$

We can arrive at this expression in several ways. One possible method involves evaluating the marginal likelihood of the transformed variable [see [95] for details]. Alternatively, one can use a Bayesian interpretation [96]. Here, one assumes that nothing is known about $\boldsymbol{\beta}$, assigns an improper flat prior (a prior that is not properly normalised), and integrates out $\boldsymbol{\beta}$. This directly leads to

$$RL(\boldsymbol{\phi}|\boldsymbol{y},\mathbf{X}) \propto \int_{-\infty}^{\infty} \frac{1}{(\det\ \mathbf{C})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{C}^{-1}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})\right) \mathrm{d}\boldsymbol{\beta}. \tag{2.77}$$

The Gaussian integral on the right-hand side of Equation (2.77) can then be evaluated exactly to give the result in Equation (2.76). In this context, we note that, for a given matrix $\mathbf{C}$, the integrand in Equation (2.77) is maximal at

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\top}\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{C}^{-1}\boldsymbol{y} \tag{2.78}$$

by construction [see Equation (2.57)].

From Equation (2.76), the restricted log likelihood is [94, 96]

$$\ln(RL(\boldsymbol{\phi}|\boldsymbol{y},\mathbf{X})) = -\frac{1}{2}\ln(\det\mathbf{C}) - \frac{1}{2}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\top}\mathbf{C}^{-1}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) - \frac{1}{2}\ln(\det\mathbf{X}^{\top}\mathbf{C}^{-1}\mathbf{X}) + \text{const.} \tag{2.79}$$

This is similar to the original log likelihood function in Equation (2.53) but with dependence on the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ instead of the true parameter $\boldsymbol{\beta}$ and with an additional term.

A possible iterative procedure to obtain estimates of the regression and variance parameters is now as follows: Start with an initial choice for $\hat{\boldsymbol{\phi}}$. This defines $\mathbf{C}$. Use this in Equation (2.78) to obtain $\hat{\boldsymbol{\beta}}$. Use this value for $\hat{\boldsymbol{\beta}}$ in Equation (2.79) and then maximise $\ln(RL(\boldsymbol{\phi}|\boldsymbol{y},\mathbf{X}))$ with respect to $\boldsymbol{\phi}$ (keeping $\hat{\boldsymbol{\beta}}$ constant). This delivers an updated value for $\hat{\boldsymbol{\phi}}$. Then repeat, and iterate until convergence.

In this process, the maximisation of the expression in Equation (2.79) with respect to $\boldsymbol{\phi}$ requires numerical techniques such as the Newton-Raphson method [97], Fisher's scoring algorithm [98], or the expectation-maximisation (EM) algorithm [99].

In NMA the parameter estimates $\hat{\boldsymbol{\beta}}$ are the estimates of the relative treatment effects $\hat{\boldsymbol{d}}^{\mathrm{RE}}$ in Equation (2.62). The covariance matrix $\mathbf{C}$ is given by $\mathbf{V}+\boldsymbol{\Sigma}$ where the within-study covariance matrix $\mathbf{V}$ is assumed known, and the between-study covariance matrix depends on the unknown parameter τ^2 we wish to estimate.

In pairwise meta-analysis $\hat{\boldsymbol{\beta}}$ is the estimate of the treatment effect $\hat{d}^{\mathrm{RE}}$ in Equation (2.64), $\mathbf{X}$ is an $M \times 1$ matrix of ones, and the covariance matrix $\mathbf{C}$ is diagonal with elements $\sigma_i^2 + \tau^2$ (with σ_i^2 assumed known). The restricted log likelihood in Equation (2.79) then simplifies to [84, 100]

$$\ln(RL(\tau^2|\boldsymbol{y})) = -\frac{1}{2}\sum_{i=1}^{M}\ln(\sigma_i^2+\tau^2) - \frac{1}{2}\sum_{i=1}^{M}\frac{(y_i-\hat{d}^{\mathrm{RE}})^2}{\sigma_i^2+\tau^2} - \frac{1}{2}\ln\left(\sum_{i=1}^{M}\frac{1}{\sigma_i^2+\tau^2}\right) + \text{const}, \quad (2.80)$$

where we recall that $\hat{d}^{\mathrm{RE}}$ depends on the *estimate* $\hat{\tau}^2$ [Equation (2.64)]. Following the above procedure we now maximise $\ln(RL(\tau^2|\boldsymbol{y}))$ with respect to τ^2, while keeping $\hat{d}^{\mathrm{RE}}$ fixed. This can be done analytically. Setting the partial derivative of the log likelihood with respect to τ^2 equal to zero yields the REML estimator,

$$\hat{\tau}^2_{\mathrm{REML}} = \max\left\{0, \frac{\sum_{i=1}^{M}(\sigma_i^2+\hat{\tau}^2_{\mathrm{REML}})^{-2}\left((y_i-\hat{d}^{\mathrm{RE}})^2-\sigma_i^2\right)}{\sum_{i=1}^{M}(\sigma_i^2+\hat{\tau}^2_{\mathrm{REML}})^{-2}} + \frac{1}{\sum_{i=1}^{M}(\sigma_i^2+\hat{\tau}^2_{\mathrm{REML}})^{-1}}\right\}, \quad (2.81)$$

where the truncation at zero ensures that $\hat{\tau}^2_{\mathrm{REML}}$ remains non-negative. The joint system of Equations (2.81) and (2.64) can then be solved iteratively for $\hat{\tau}^2_{\mathrm{REML}}$ and $\hat{d}^{\mathrm{RE}}$.

Of the multiple heterogeneity estimators described here, REML is usually the recommended option (see for example [83, 84]).

2.5 Reporting NMA Results

In Sections 2.3 and 2.4 we have explained how to obtain estimates for the model parameters in NMA using Bayesian and frequentist methods. We now explain how these estimates are reported and summarised for use in decision making.

2.5.1 Confidence/credible intervals in frequentist and Bayesian inference

We focus on a particular parameter x. What is reported at the end of the inference process is an estimate for the parameter along with a measure of precision. In frequentist inference the parameter estimate itself is the one discussed in Section 2.4.3.1, and an

estimate of the variance is obtained from Equation (2.63). In Bayesian inference the parameter estimate is usually the mean or median of the samples from the posterior distribution, and we also record the variance of the samples for the parameter [17].

In a frequentist setting uncertainty on a parameter is often expressed in terms of confidence intervals, for example a '95% confidence interval'. In Bayesian inference uncertainty is expressed in terms of 'credible intervals'. We note the subtle difference between these two concepts. The Bayesian interpretation is intuitive: given the observed data, there is a 95% probability that the true (unknown) parameter lies within this interval [44]. In a frequentist setting this would mean that if we were to repeat the experiment and inference many times (each time constructing a 95%-confidence interval) then 95% of these intervals would contain the true value of the parameter [44].

The ζ% confidence interval ($0 \leq \zeta \leq 100$) is constructed from the parameter estimate and its variance assuming a Gaussian distribution for the parameter with mean $\hat{x}$ and variance $\mathrm{Var}(\hat{x})$. More precisely,

$$\mathrm{CI} = \hat{x} \pm q(\zeta)\sqrt{\mathrm{Var}(\hat{x})}, \tag{2.82}$$

where $q(\zeta)$ is such that a total probability of ζ% of the Gaussian distribution is in the interval of length $2q(\zeta)\sqrt{\mathrm{Var}(\hat{x})}$ around the mean. (Scaling out the variance, this means $\int_{-q}^{q} dx\ e^{-x^2/2}/\sqrt{2\pi} = \zeta/100$.) For example, using $q = 1.96$ in Equation (2.82) indicates a 95% confidence interval. In a Bayesian setting the 95% credible interval can be obtained in a similar way from Equation (2.82), but using the mean and variance of samples drawn from the posterior. Alternatively, one can calculate the 2.5% and 97.5% quantiles of the posterior samples.

Alternative methods for constructing confidence intervals in a frequentist analysis are also possible. For example, a modification of the standard approach which relaxes the normality assumption was suggested by both Hartung and Knapp [101–103] and Sidik and Jonkman [104]. While not without its own drawbacks [105], this approach has been found to outperform the standard method in simulation studies and thus, may be a preferable option [106].

2.5.2 Forest Plots

A common way to present the relative treatment effect estimates in both frequentist and Bayesian NMA is on a forest plot [107–109]. Each basic parameter d_{T_1a} is represented by a horizontal line centred on its estimated value $\hat{d}_{T_1a}$ with length equal to its confidence (or credible) interval. For relative treatment effects measured as log odds ratios a value of zero represents no difference in treatment effect. When the outcome from the trials is the number of negative events or 'failures', a log odds ratio $\hat{d}_{T_1a} < 0$ indicates that treatment a is more effective than the baseline T_1 and vice versa. A forest plot showing the results for a frequentist analysis of the Thrombolytic drug data in Section 2.2.1.3 is shown in Figure 2.8. A 'caterpillar plot' is also sometimes used [13]. This is essentially the same as a forest plot but the relative treatment effects are sorted in order of increasing effect size [110].

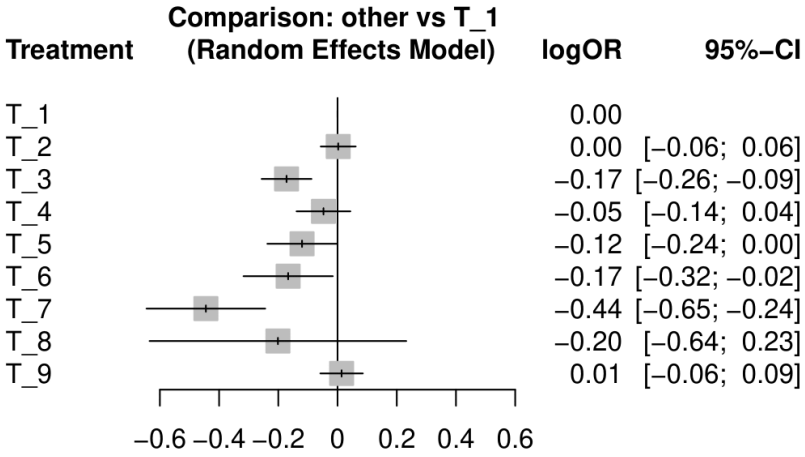


Figure 2.8: A forest plot showing the results of a frequentist (random effects) analysis of the Thrombolytic drug data set [10–12]. The network graph and treatment labels are shown in Figure 2.1. The global baseline treatment is T_1 and has a log odds ratio of 0 by definition. The outcome of interest is the number of deaths that occur within 30 or 35 days of a heart attack. Therefore a log odds ratio < 0 indicates that the treatment is more effective than the baseline T_1. The horizontal lines indicate the 95%-confidence intervals about the estimated LORs. Figure was created using the software `netmeta` [107] . (The grey boxes highlight the central value and do not convey any additional information.)

2.5.3 Ranking

The aim of a network meta-analysis is to provide clinicians with a clear statistical summary of all relevant data so that they know what the most desirable treatment options are for a particular condition. Relative treatment effect estimates and their confidence/credible intervals can be difficult to interpret and draw conclusions from, especially when many treatments have been compared [111, 112].

Ranking treatments from best to worst based on their relative effect estimates is the simplest way of summarising the results of an NMA. For example, from the estimated treatment effects $\hat{d}_{T_1a}$ in the forest plot in Figure 2.8 the treatments would be ranked from best to worst in the order $T_7, T_8, T_3 = T_6, T_5, T_4, T_1 = T_2, T_9$ (where we have written $a = b$ for treatments that are observed to be equally effective within the reported number of digits for the treatment effects).

However, this summary does not account for the level of overlap between the confidence intervals or the similarity between the point estimates. For example, in the Thrombolytic drug data set treatment T_8 is ranked second best using this method despite the fact it has a very large confidence interval that covers almost the entire width of the other intervals. On inspection of the forest plot we cannot draw any meaningful conclusion about the effectiveness of treatment T_8 but this fact is not reflected in its rank.

2.5.3.1 Rank Probability and Rankograms

A more sophisticated method of ordering the treatments is to calculate so-called rank probabilities [111]. That is, we calculate the probability that each treatment is best, second best and so on. We use the notation $P_a(r)$ to represent the probability that treatment a has rank r. These quantities are only meaningful in a Bayesian framework where probability describes the degree of belief in parameter values. In a Bayesian NMA, treatments are ranked at each iteration of the MCMC according to the values of relative treatment effect sampled at that iteration. The probability $P_a(r)$ is then estimated from the proportion of times treatment a was ranked r-th.

Although rank probabilities do not strictly make sense in a frequentist framework, so-called 're-sampling' methods have been developed to produce estimates of rank probabilities based on the results of a frequentist NMA [48, 113]. Essentially, this

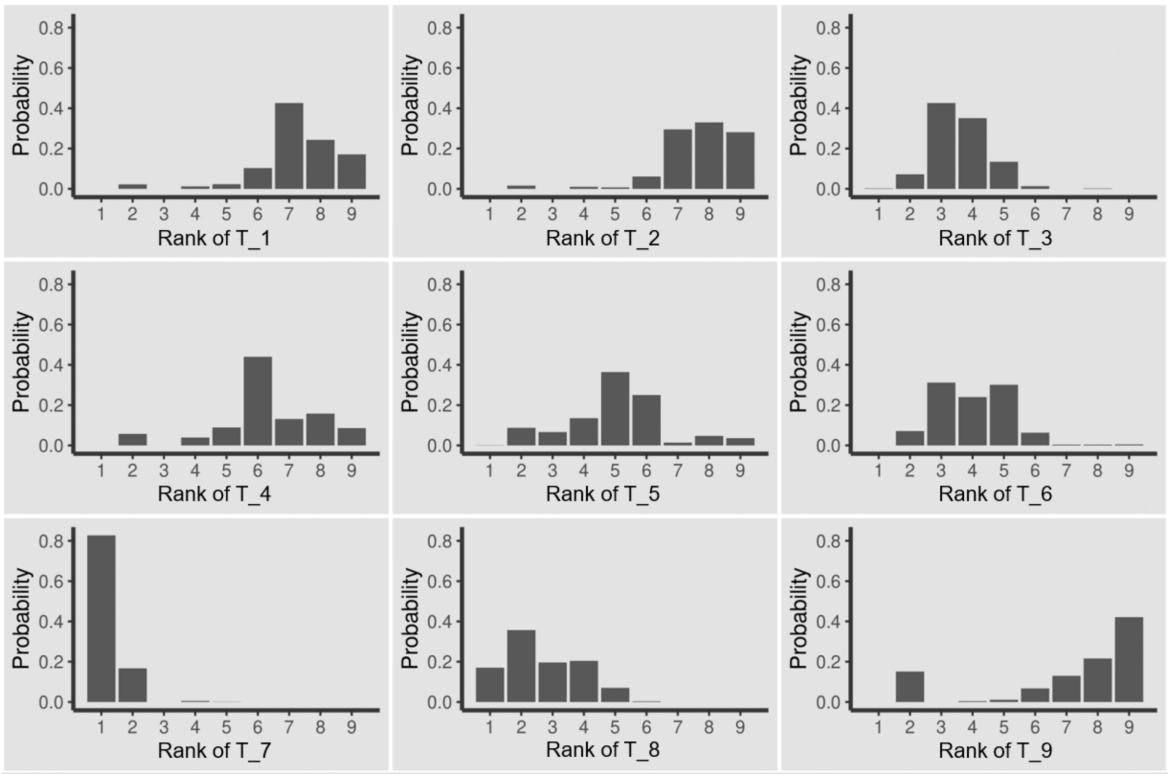


Figure 2.9: Rankograms for the Thrombolytic data set in Figure 2.1. Rank probabilities $P_a(r)$ are plotted against rank r for each treatment in the network, $a = T_1, \ldots, T_9$. Rank probabilities were obtained from frequentist re-sampling methods (based on 1000 simulations) using `netmeta` [107].

involves assuming that the distribution of the model parameters can be approximated by a normal distribution with mean and variance equal to the values estimated from frequentist methods (Section 2.4.3). Values of relative treatment effect are sampled multiple times from this approximate distribution and rank probabilities are estimated in the same way as before.

The rank probabilities can be displayed graphically using 'rankograms' [111]. For each treatment, the rank probabilities $P_a(r)$ are plotted against the rank r either as a bar chart or a line graph. The rankograms for treatments in the Thrombolytic drug data set are shown in Figure 2.9 as bar charts.

Ranking using probabilities reflects not only the point estimates of relative treatment effects but also the uncertainties on these estimates and their overlapping confidence/credible intervals. Clearly, the more overlap between intervals, the flatter the rankograms will be. For example, in the Thrombolytic data set, while treatment T_8 has the highest probability of being second best, its rankogram is relatively flat indicating the uncertainty in its rank.

2.5.3.2 SUCRA and P values

The number of rank probabilities for a network increases with N^2. Therefore rank probabilities and rankograms become increasingly difficult to interpret as the number of treatments increases [112].

Instead of rankograms, we could instead plot the cumulative probability $F_a(r)$ against rank r to obtain cumulative ranking curves. Here, $F_a(r)$ is the probability that treatment a has rank r or better,

$$F_a(r) = \sum_{s=1}^{r} P_a(s). \tag{2.83}$$

A simple summary of rank probabilities is then the area under these curves. Salanti et al (2011) [111] termed this measurement the 'surface under the cumulative ranking line' or SUCRA. The value of SUCRA for a particular treatment a is then

$$\text{SUCRA}_a = \frac{1}{N-1}\sum_{r=1}^{N-1} F_a(r) = \frac{1}{N-1}(N - \mathbb{E}(r)_a), \tag{2.84}$$

where $\mathbb{E}(r)_a$ is the mean or expected rank of treatment a,

$$\mathbb{E}(r)_a = \sum_{r=1}^{N} rP_a(r). \tag{2.85}$$

SUCRA takes values from 0 to 1, though these are often expressed as a percentage. If treatment a ranks first with probability one then it will have $\text{SUCRA}_a = 1$ (or 100%) whereas a treatment that ranks worst with probability one will have $\text{SUCRA}_a = 0$ (or 0%) [111]. SUCRA_a can be interpreted as the average proportion of treatments worse than a. SUCRA values provide a concise summary of treatment rankings that accounts for the estimated relative treatment effects, uncertainty in these estimates, and the resulting overlap in their confidence/credible intervals.

In frequentist NMA values of SUCRA can be calculated from the rank probability estimates obtained via re-sampling methods. Alternatively, Rücker and Schwarzer [114] proposed an analogous quantity called a 'P score' that does not require re-sampling. By assuming a normal distribution for the model parameters they define

$$F_{ab} = \Phi\left(\frac{\hat{d}_{ab}}{\sqrt{\text{Var}(\hat{d}_{ab})}}\right), \tag{2.86}$$

where $\Phi(.)$ is the CDF of the standard normal distribution. This is interpreted as the extent of certainty that $\hat{d}_{ab} > 0$ (i.e. that a is more effective than b). The P-score for

treatment a is then the mean of F_{ab} over treatments $b \neq a$. This is the mean extent of certainty that a is more effective than any other treatment. Rücker and Schwarzer [114] show that when the true probabilities are known, P-scores and SUCRA values are identical. In practice they give very similar results.

2.6 Existing points of contact between NMA and physics

In the previous sections we have introduced some of the essential concepts and methods for NMA. In the remainder of the paper we now discuss how physics (in particular, statistical physics) and physicists can contribute to this area.

For example, a number of analogies between meta-analysis and specific physical systems have been proposed in recent years. These analogies have provided insight, and they have helped to improve meta-analysis methodology and the visualisation of the problem. In this section we briefly outline these analogies.

Section 2.7 then describes a few examples of more general methods used in statistical physics which have been shown to be useful in a meta-analysis context. We also present a number of more speculative ideas on how knowledge from physics might be used for NMA.

2.6.1 NMA and electrical networks

Arguably the most influential of the meta-analysis analogies was developed by Rücker (2012) [49] who demonstrated the connection between NMA and electrical network theory. The starting point for this analogy is the observation that variance in meta-analysis combines in the same way as resistance in an electrical network.

2.6.1.1 Variances in NMA combine like resistance in a network

In Section 2.4.3.1 we saw that for a frequentist pairwise meta-analysis, the variance of the estimated treatment effect $\hat{d}$ is an expression for the variance of the weighted mean [Equation (2.65)] calculated in terms of the variances associated with each of the trials. Taking the reciprocal on both sides of this equation and writing v_i for the

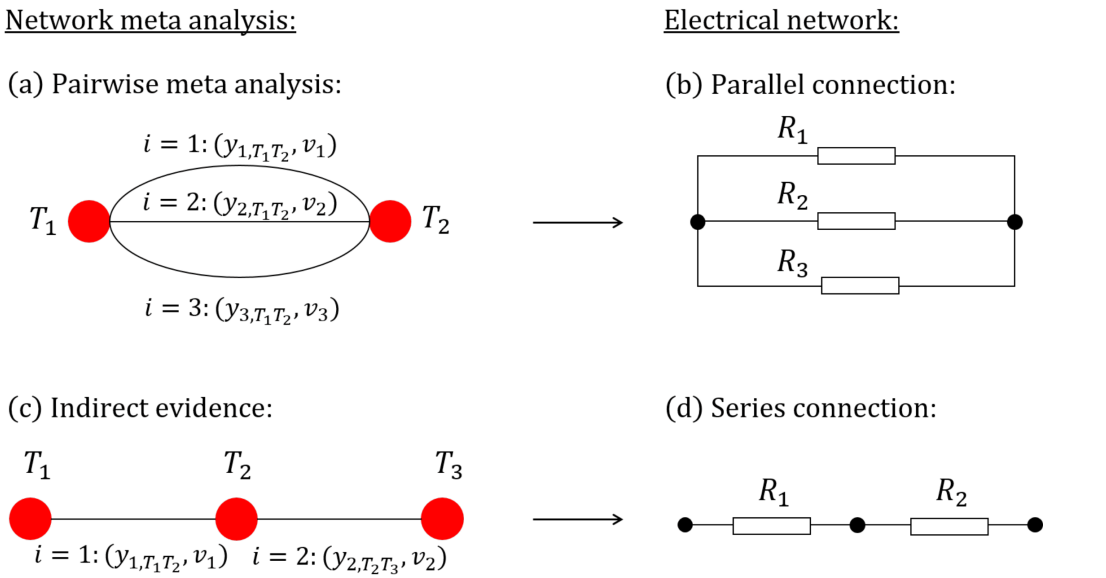


Figure 2.10: An illustration of the analogy between NMA and electrical networks. (a) A pairwise meta-analysis of three trials corresponds to (b) three resistors connected in parallel. (c) A chain of two trials connecting three treatments corresponds to (d) two resistors connected in series. We label each treatment as T_a and each resistor as R_i. Each trial i is labelled with the measurement of relative treatment effect made in that trial, $y_{i,12}$, and the associated variance, v_i.

variance associated with the measurement in trial i we find

$$\frac{1}{\mathrm{Var}(\hat{d})} = \sum_{i=1}^{M} \frac{1}{v_i}. \tag{2.87}$$

For a random effects model we have $\hat{d} = \hat{d}^{\mathrm{RE}}$ and $v_i = \sigma_i^2 + \tau^2$, whereas for a fixed effect model $\hat{d} = \hat{d}^{\mathrm{FE}}$ and $v_i = \sigma_i^2$. As illustrated in Figure 2.10 (a), a pairwise meta-analysis can be represented by a graph with two nodes (representing the two treatment options) and multiple parallel connections (edges) between the nodes (representing the individual trials comparing the treatments). The same graphical representation describes an electrical network with resistors connected in parallel [Figure 2.10 (b)]. The effective resistance, R, of a set of M parallel resistors R_i, $i = 1, \ldots, M$ is

$$\frac{1}{R} = \sum_{i=1}^{M} \frac{1}{R_i}. \tag{2.88}$$

Therefore, resistors in parallel combine like variances in a pairwise meta-analysis.

Now consider the network in Figure 2.10 (c) comprising three treatments, T_1, T_2, T_3, and two trials. The first trial $i = 1$ compares treatments T_1 and T_2 and measures a relative treatment effect y_{1,T_1T_2} with variance v_1. Trial $i = 2$ compares treatments T_2 and T_3 and measures y_{2,T_2T_3} with variance v_2. The network estimates of the relative

effect between treatments T_1 and T_2, and between T_2 and T_3, respectively, are simply the direct estimates $\hat{d}_{T_1T_2} = y_{1,T_1T_2}$ and $\hat{d}_{T_2T_3} = y_{2,T_2T_3}$. There is no direct evidence for the comparison of treatments T_1 and T_3. The network estimate of this comparison is obtained from an indirect estimate via T_2,

$$\hat{d}_{T_1T_3} = y_{1,T_1T_2} + y_{2,T_2T_3}. \tag{2.89}$$

Since the trials are independent, the variance associated with this estimate is

$$\mathrm{Var}(\hat{d}_{T_1T_3}) = v_1 + v_2. \tag{2.90}$$

As shown in Figure 2.10 (d) this set-up relates to an electrical network with resistors connected in series. The effective resistance for this network is

$$R = R_1 + R_2. \tag{2.91}$$

Therefore resistors in series combine like variances for indirect estimates in NMA.

The analogy with resistor circuits can be extended to networks of more than two trials, and those combining both parallel connections and connections in series. Through a more detailed analysis (which we do not describe here) and by comparing Ohm's law to the weighted mean expression in Equation (2.62), one can then establish a mapping between relative treatment effects and potential differences (voltages).

2.6.1.2 Analogy of the NMA problem in electric circuits

In electrical network theory, graph theoretical methods are used in different ways, for example to construct the electric potentials at the nodes from external currents, or to compute the effective resistance between two nodes from the resistors in the network. Rücker showed that a similar set of methods can be used in NMA to derive an expression for the network estimates of the relative treatment effects from the observed effects. This leads to the same results as the Aitken estimator in Equation (2.64) [49, 115].

We do not present details here, but the core of the analogy can be described as follows [49]. The NMA problem consists of finding the network estimates for the relative treatment effects using the effects observed in the trials (for a known network structure and known [inverse-variance] weights of the trials). The observed effects

will in general be inconsistent, whereas the estimates resulting from the NMA are consistent by construction. The observed effects translate to 'observed' voltages across the resistors in the network (the resistances are determined by the inverse-weights of the trials). As a consequence of the inconsistency, voltages along loops in the network will not add to zero. This means that no electric potentials can be assigned to the nodes from which the voltages would arise as potential differences. The key result is now that the problem of determining the NMA estimates for the treatment effects is equivalent to finding the set of electric potentials so that the resulting (consistent) voltages best approximate the observed (inconsistent) voltages[6]. Quality of approximation is here measured in terms of the Euclidian norm.

2.6.1.3 Reduction of multi-armed trials

One particularly useful application of the electrical network analogy is in the context of networks with multi-arm trials. Measurements of the different relative treatment effects from a multi-arm trial are correlated, and the presence of such correlations can cause complications for some NMA methodology. One therefore carries out a 'reduction' to an equivalent set of two-arm trials. We here briefly describe this, for details see [49, 115].

We focus on a single multi-arm trial with m-arms. The corresponding (sub) network involving the m treatments is then fully connected. The idea is to find a network consisting of $m(m-1)/2$ pairwise trials which is 'equivalent' to the multi-arm trial in the sense that the variances of the network estimates of relative treatment effects in the pairwise network [given by Equation (2.63)] are the same as the variances in the graph describing the multi-arm trial.

As discussed above, variances in NMA combine like resistances in electric networks, i.e. the variances of network estimates are obtained from the individual trial variances in the same way as *effective* resistance is obtained from the physical resistors in electric circuits[7]. The reduction problem for the multi-armed trial is therefore equivalent

[6]These potentials are only unique up to an overall additive constant. This reflects the fact that NMA tries to estimate relative treatment effects rather than absolute outcomes.

[7]The effective resistance between two nodes results as U/I from the current I that flows into the network if a battery of voltage U is attached to the two target nodes. Effective resistance accounts for any direct connection between the target nodes, and for all indirect connections through other nodes in the graph.

to finding the individual resistances $\{R_{ab}\}$ in an electric network given the *effective* resistances between pairs of nodes. It is well known (see e.g. [116]) that

$$R_{ab}^{\text{eff}} = L_{aa}^{+} + L_{bb}^{+} - 2L_{ab}^{+}, \tag{2.92}$$

where $\mathbf{L}$ is the graph Laplacian describing the electric network, and where $^{+}$ denotes the pseudo-inverse. The graph Laplacian is defined by the individual (physical) resistors via $L_{ab} = -R_{ab}^{-1}$ for $a \neq b$, and $L_{aa} = \sum_b R_{ab}^{-1}$.

The reduction problem therefore maps onto the problem of finding the elements of the graph Laplacian $\mathbf{L}$ in Equation (2.92) for given effective resistances $\{R_{ab}^{\text{eff}}\}$ (i.e. given variances in the multi-armed trial). The individual physical resistors (variances associated with the individual two-armed trials) can then be extracted from the off-diagonal elements of the Laplacian. Further details of the reduction method are given in Section 2.12 of the Appendix.

We perform this reduction for every multi-arm trial in a meta-analytic network, and use effect estimates from the multi-arm trials to assign estimates to the two-arm trials. This leads to a network of two-arm trials that is equivalent to the original network (i.e. it produces the same network estimates and variances). Methodology that does not allow for correlations can then be used on this new network.

Further details of the analogy between NMA and electrical networks can be found in [49, 115].

2.6.2 Random Walks

Random walks are a familiar concept to statistical physicists. Random hopping processes on networks are of particular interest in a number of areas [117–119]. In brief, a random walk on a network is a stochastic process describing a series of hops between nodes that are connected by an edge. We focus on discrete time such that each time step is associated with one hop across an edge.

There is a well-known analogy between random walks and electrical networks [120–123], briefly summarised in the next section. Using the work of Rücker [49] described in Section 2.6.1 we were able to extend this analogy to network meta-analysis [16].

2.6.2.1 Random walks and resistor networks

Each edge in an electrical resistor network has an associated resistance. Given such a network, one now constructs a random walk as follows: For a random walker currently at node a, the probability with which the walker hops from a to b in the next step is proportional to the inverse-resistance associated with the edge ab. More precisely, one defines the transition matrix elements as

$$P_{ab} = \frac{R_{ab}^{-1}}{\sum_{c \neq a} R_{ac}^{-1}}, \tag{2.93}$$

where R_{ab} is the resistance of the resistor connecting nodes a and b.

Various physical quantities in the electric network then have interpretations in the random-walk picture. A good summary can be found in [123]. For example, consider the following scenario: A battery is attached to two nodes a and b in the resistor network. We assume that the network only has one single connected component. The voltage of the battery is chosen such that one Ampère of current goes into node a from the battery (and consequently one Ampère goes from node b back into the battery). This then induces currents I_{cd} through all edges cd (the resistors) in the network. Now imagine we release a random walker at node a, and it performs the random walk defined by Equation (2.93). We stop the walk when the walker reaches node b (this will happen eventually given that the network consists of a single component). We can record the net number of times the walker will have crossed edge cd before it reaches node b (hops from d to c contribute negatively to this value). One can then show [123] that the expected net number of crossings from c to d is given by the current I_{cd} in the electric network with the battery attached to nodes a and b.

2.6.2.2 Random walks and flow of evidence in network meta-analysis

Starting from the existing analogies between electrical networks and NMA on the one hand, and electrical networks and random walks on the other, we (along with Rücker, Papakonstantinou and Nikolakopoulou) [16] proposed an analogy between NMA and random walks. In the following, we briefly summarise the main ideas.

We have seen that resistance in an electrical network is analogous to variance in an NMA. Therefore inverse-resistance is associated with inverse-variance weight. Writing w_{ab} for the weight associated with edge ab in a network meta-analysis (see

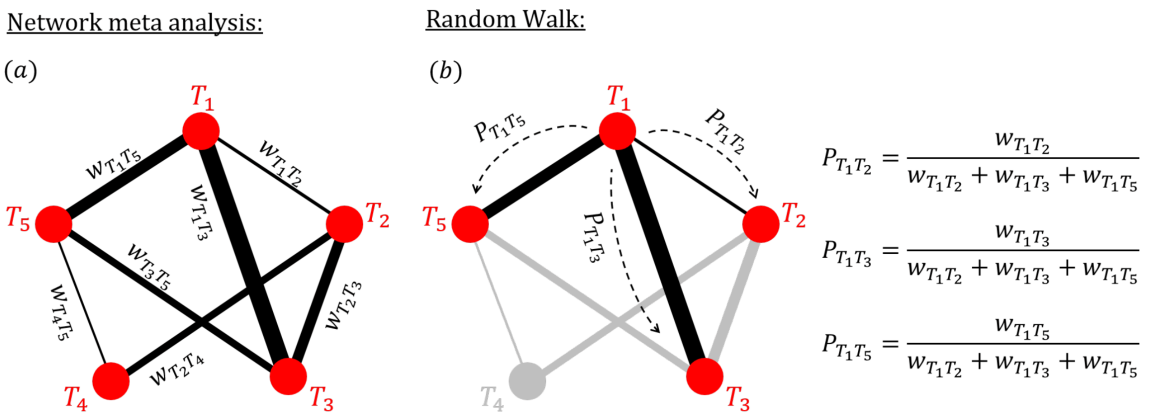


Figure 2.11: An illustration of the analogy between NMA and random walks. Panel (a) shows an NMA with five treatments $a = T_1, T_2, T_3, T_4, T_5$. Each edge is labelled with the inverse-variance weight associated with that treatment comparison, w_{ab}. Panel (b) shows the transition probabilities for a random walker on the network in (a) who is currently at node T_1. At the next time step this walker can move to node T_2, T_3 or T_5 with probabilities proportional to the edge weights.

Section 2.4.3.1), we define the transition matrix of a random walker via

$$P_{ab} = \frac{w_{ab}}{\sum_{c \neq a} w_{ac}}. \tag{2.94}$$

This is illustrated in Figure 2.11.

In [16] we found that the expected net number of times a walker crosses each edge is equal to the so-called 'evidence flow' through that edge. This is a concept introduced by König et al (2013) [124]. We do not give full definitions here. Broadly speaking the flow variable $f_{cd}^{(ab)}$ is a coefficient that describes how much the observed relative effect between treatments c and d contributes to the network estimate of the relative effect between a and b. The coefficients are related to the entries of the matrix on the right-hand side of Equation (2.62). They describe how different pieces of direct evidence in a network meta-analysis combine to give the overall network estimates of relative treatment effects. It turns out [124] that these coefficients have the properties of a flow. For example, the flow out of a node $c \neq a, b$, $\sum_x f_{cx}^{(ab)}$, equals the flow into node c, $\sum_x f_{xc}^{(ab)}$, indicating that flow is conserved at c. The flow variables are defined such that they are non-negative. A value of $f_{cd}^{(ab)} > 0$ indicates a positive flow from c to d, the transposed variable $f_{dc}^{(ab)}$ is then zero.

For a comparison ab, the evidence flow variables $\{f_{cd}^{(ab)}\}$ can be used to define a directed weighted graph. This is for a fixed choice of a and b meaning there is one directed graph for each comparison ab. Nodes in this graph again represent the

treatment options, but the weights of the (now directed) edges are given by the flow of evidence $\{f_{cd}^{(ab)}\}$.

2.6.2.3 Random walks, streams of evidence and proportion contributions

In [16] we then defined a second random walk. All walkers start at a and move on the directed graph just described. The construction is such that walkers can never return to a node they have already visited (the graph is acyclic), and all walks end at b (node b is absorbing). We can think of these walkers as collecting evidence along their way. They start at node a, hop to intermediate nodes and record differences in treatment effects (similar to differences in height in a mechanical set-up). When a walker arrives at b it reports the total difference in altitude it has experienced. Due to inconsistencies, this reported difference may be different along different paths connecting a and b. The average of what the walkers report turns out to be the network estimate of the relative treatment effect between a and b [16].

In addition, the probability of a walker taking a certain path from a to b is given by the product of the transition probabilities in the edges along that path. This expression can be used to calculate so-called 'streams of evidence' [125], and what is referred to as the 'proportion contribution matrix' [126]. In particular, we used the random walk transition probabilities to derive an analytical expression for the contribution each treatment comparison makes to each overall treatment effect estimate. As discussed in more detail in [16], this random walk approach overcomes some limitations of previous algorithms used to construct this quantity. The random-walk method has recently been implemented in the software package `netmeta` [107].

The random walk analogy is a recent addition to network meta-analysis. As a result, only a small proportion of the random walk literature has been explored in this context. We therefore feel that there is scope to extend the analogy. We hope that the introduction to NMA we give in this paper will help other statistical physicists with an interest in random walks on networks to join these efforts.

2.6.3 System of springs

Papakonstantinou et al (2021) [127] visualised the meta-analysis problem as a system of springs. In a similar vein to the electrical network analogy, they observe that when

connecting systems of springs the inverse of the spring constant combines in the same way as variance in meta-analysis.

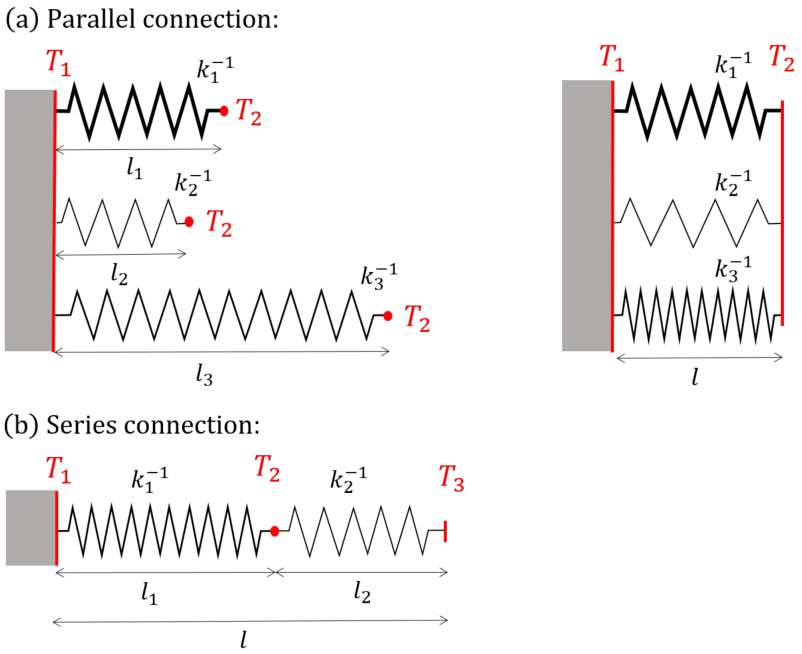


Figure 2.12: (a) An illustration of a parallel system of springs. The springs are fixed on one side corresponding to the baseline treatment. The open ends are then forced to the same length so that their natural lengths are displaced by some distance. This is equivalent to a pairwise meta-analysis. (b) A system of springs connected in series. This is equivalent to an indirect comparison in meta-analysis. We label each treatment as T_a. Each spring is labelled by its natural length, l_i, and the inverse of its spring constant, k_i^{-1}. l is the effective length of the spring system. The different thicknesses of the springs represent their different spring constants.

Hooke's law states that the force, f, needed to displace a spring by length x is $f = kx$, where k is a measure of the stiffness of the spring known as the spring constant. The potential energy stored in such a spring is

$$U = \frac{1}{2}kx^2. \tag{2.95}$$

Consider a set of M springs fixed at one end and arranged in parallel as shown in Figure 2.12 (a). Each spring has a different spring constant, k_i, and a different natural length, l_i. The open ends of the spring are then connected so that the springs are forced to assume the same length, called the 'effective length' l of the spring. Each spring has therefore been displaced by a different amount, $x_i = l_i - l$. The resulting equilibrium length of the springs, $\hat{l}$, is such that the total force on T_2 vanishes, i.e. $\sum_i k_i x_i = 0$. (Equivalently, this length minimises the energy stored in the system,

$\hat{l} = \underset{l}{\text{argmin}} \left[\sum_{i=1}^{M} k_i (l_i - l)^2\right]$.) One finds,

$$\hat{l} = \frac{\sum_{i=1}^{M} k_i l_i}{\sum_{i=1}^{M} k_i}. \tag{2.96}$$

This expression is in the form of a weighted mean. Comparing it to the estimate from a pairwise meta-analysis in Equation (2.64), one observes that we can draw an analogy between the parallel system of springs and a pairwise MA. We interpret the spring constant as inverse-variance weight. The spring constant associated with the 'effective spring' is

$$k = \sum_{i=1}^{M} k_i. \tag{2.97}$$

Comparing this to Equation (2.87), it is clear that variance in a pairwise meta-analysis combines in the same way as the inverse of the spring constant for a set of parallel springs.

One can also make this analogy for springs connected in series. As shown in Figure 2.12 (b), the effective length of a chain of springs connected together is simply the sum of their natural lengths. The effective spring constant is then

$$\frac{1}{k} = \sum_{i=1}^{M} \frac{1}{k_i}. \tag{2.98}$$

Comparing this to Equation (2.90), we observe that the inverse of the spring constants of springs connected in series combine like variances for an indirect estimate in meta-analysis.

The natural length of each spring can be interpreted as the measurement of the relative treatment effect in each trial. The displacements x_i then relate to the residuals associated with the relative treatment effects (i.e. differences between the relative effects measured in each trial and those predicted by the model). Minimising the energy is analogous to the process of minimising the sum of squared weighted residuals described in Sections 2.4.2.1 and 2.4.2.2.

This analogy provides a useful visualisation of the meta-analysis process. That is, combining the data from multiple medical trials is like minimising the energy in a system of springs. The energy stored in the final equilibrium system then represents a measure of disagreement between the different pieces of evidence. Though this analogy has been demonstrated for a set of relatively simple configurations, it has not yet been

extended to a general network meta-analysis. This would require a much more complex spring system. Further details can be found in [127].

2.6.4 Balance of torques in a mechanical system

Another visualisation of meta-analysis based on a mechanical system was proposed by Bowden and Jackson (2016) [128]. In this analogy, the process of finding the minimum sum of weighted residuals (or, equivalently, maximising the likelihood, see Sections 2.4.2.2 and 2.4.2.3) is equated to balancing torques in a system of weights. This is the same as finding the position of the centre of mass.

Figure 2.13 shows a mechanical system consisting of a bar with M objects of different masses, m_i, hanging at various positions x_i along the bar. The bar is supported by a pivot. The system of masses (or weights) is balanced when the pivot is placed such that the torques exerted by the masses balance, i.e. $\sum_{i=1}^{M} m_i g(x_i - x_{\text{pivot}}) = 0$, where g is the acceleration due to gravity. Writing the weights of the different masses as $w_i = m_i g$ the balance of torques leads to

$$x_{\text{pivot}} = x_{\text{CoM}} \equiv \frac{\sum_{i=1}^{M} w_i x_i}{\sum_{i=1}^{M} w_i}, \tag{2.99}$$

i.e. the pivot must be located at the centre of mass (in one dimension) defined by the system of weights.

As before, we compare this to the pairwise meta-analysis estimate in Equation (2.64) in order to establish the analogy. The position of each mass along the bar then represents the observed relative treatment effect in each trial and the physical weight of each object represents the inverse-variance weight associated with each observation.

The problem of finding the position of the centre of mass provides a visual representation of finding the best estimate of the relative treatment effect in a pairwise meta-analysis. Bowden and Jackson used this visualisation to create an online visualisation tool. The visualisation shows how different modelling assumptions (for example, fixed effects vs random effects) or the addition/removal of certain studies affect the position of the centre of mass. When the user makes a change to the model or the data, an animation shows the change in balance of the system. This was found to help identify the presence of small-study bias, a phenomenon where smaller studies report a systematically larger relative treatment effect than larger studies [129]. This analogy

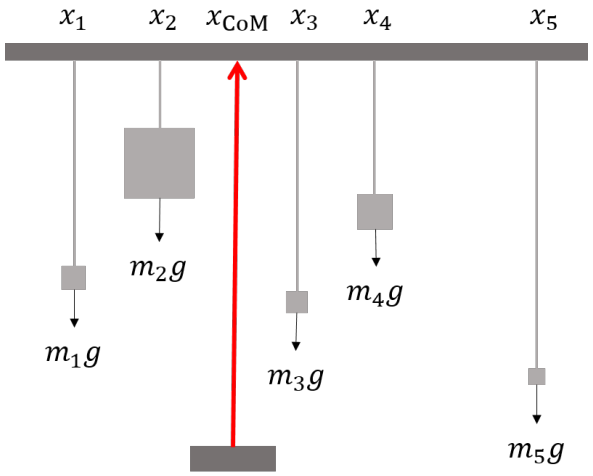


Figure 2.13: An illustration of pairwise meta-analysis as a centre of mass problem. Each mass is labelled by its position along the rod, x_i, and the force acting on it, $m_i g$, which is equal to its weight. The position of each object represents the measurement of relative treatment effect in each trial. The physical weight of each object represents the inverse-variance weight of each measurement. The centre of mass, x_{CoM}, is the position of the pivot that balances the torques. Finding this position is equivalent to finding the estimate of relative treatment effect in a pairwise meta-analysis.

has not been extended to indirect estimates or network meta-analysis. For details, see [128].

2.7 Ideas for future work: a research programme at the interface of statistical physics and network meta-analysis

In this section we now present a series of broader ideas on how statistical physics might contribute to the field of NMA in the future. Some of these have been formulated in existing literature (e.g. Markov chain methods to rank treatments in Section 2.7.1), others are more speculative.

2.7.1 Markov chain approaches to ranking

Markov chain approaches have been used as ranking methods in a variety of fields [130–132]. Most notably, Google's PageRank algorithm [130] ranks web pages in their search engine results. Broadly summarising, the algorithm is based on a Markov chain that models a 'random surfer' who, after being randomly assigned an initial webpage,

moves from page to page by randomly clicking links. The algorithm also allows for a damping mechanism that works by assigning a certain probability with which the surfer is re-set to a random page at each step (this ensures that pages with no incoming links are also visited from time to time). The stationary distribution of the Markov chain then informs the ranking of the webpages. Broadly speaking, a page with more incoming links is more likely to be visited and therefore attracts a higher probability at stationarity, resulting in a higher rank.

As discussed in Section 2.5.3 ranking treatments is an important output of NMA and has received much attention in the literature [133–140]. Chaimani et al [137] developed a ranking method for NMA based on the PageRank algorithm.

As in the random-walk framework, each state of the Markov chain is a treatment in the network. At each discrete time step the process moves from one treatment to another. Transitions represent a preference between two treatments: When the process is currently in state a, the probability of transitioning to b in the next time step is related to the probability that treatment b is more effective than treatment a [114]. The stationary distribution of the Markov chain can then be used to rank the treatments, where a higher probability of being selected indicates a more effective treatment.

The initial distribution of the Markov chain can be chosen so that it reflects clinically important factors other than the treatment effects, for example the cost or safety of the treatments [137]. Similar to the PageRank algorithm, Chaimani et al introduce a re-setting mechanism: at every time step there is a non-zero probability of hopping to a state drawn from the initial distribution. This means that the stationary distribution of the Markov chain now depends on this initial distribution as well as the transition probabilities. This allows one to incorporate information about a range of factors that influence decision making.

2.7.2 Using network theory to characterise meta-analytic graphs

A natural point of contact between NMA and statistical physics is the theory of networks. It is widely recognised that the accuracy and precision of NMA outcomes are likely to be affected by network topology. Indeed, researchers are encouraged to

provide graphical and qualitative descriptions of network geometry [141]. Based on concepts from the theory of networks, researchers have defined topological indices to describe the geometry of meta-analytic networks and related these to the outcomes of NMA [142–144]. We give a few examples in this section, recognising that this existing work is only an initial step. We think that there is significant scope to extend these activities.

In network theory, the degree of a vertex is the number of edges the vertex shares. A graph is described as 'regular' if all vertices have the same degree. From these definitions we [142] defined a measure of 'degree irregularity' of the graph of treatment options. This measure is given by

$$h^2 = \frac{1}{N}\sum_a (k_a - \bar{k})^2, \tag{2.100}$$

and quantifies the variation in the number of studies involving each treatment. We define k_a as the weighted degree of node a (here, the weight of an edge is given by the number of trials comparing the two treatments connected by the edge). The quantity $\bar{k} = \frac{1}{N}\sum_a k_a$ is the mean degree in the network. Through simulations of NMA, we found that smaller values of h^2 were associated with more precise treatment effects and smaller bias on rank probabilities.

Tonin et al (2019) [143] adapted metrics from graph theory in order to numerically describe network geometry in NMA. They performed a systematic review of 167 published NMAs and used 11 metrics to describe the topology of each network. By performing a sensitivity analysis on each metric and assessing the level of correlation between the metrics, they identified four indicators that were the most useful for describing network geometry. These are (i) Density: A measure of 'connectedness' equal to the number of edges in the network divided by the total number of edges possible for N vertices; (ii) Percentage of common comparators: The percentage of vertices with more than one connecting edge relative to the total number of vertices N; (iii) Percentage of strong edges: The percentage of edges with more than one study relative to the total number of edges; (iv) Median thickness and dispersion measure: The median number of studies per edge and the IQR (interquartile range).

Tonin et al recommended that in order to characterise the network of evidence from an NMA these measures, in addition to the number of vertices, the number of edges, and the number of trials per edge, should be reported alongside the network graph.

This existing work shows that concepts used in network theory to characterise the topology of networks can be linked to the outcome of NMA. We think that further work to explore these ideas will be very worthwhile. Significant expertise exists in the complex networks community, and a large number of tools have been developed to characterise, for example, the clustering properties of networks, as well as measures of centrality and assortativity. Techniques are available to detect communities in networks, and more general motifs. It will be exciting to see if and how these concepts can be used to study networks of medical trials. Finding good indicators of network topology for NMA performance also leads to the possibility to suggest targeted additions to an existing meta-analytic network. In other words, these methods from the theory of (complex) networks could be used to propose future trials to be added to a meta-analysis that promise to be the most useful for improving the overall estimate of treatment effects. This could include concepts from chemical graph theory, e.g. topological indices discussed in [145].

2.7.3 Network meta-analysis, constrained optimisation and statistical mechanics

Counting is very much at the heart of statistical physics. Boltzmann's entropy, $S = k \ln \Omega$, is defined based on the number of accessible microstates Ω of a system. Establishing the statistical physics of the microcanonical ensemble then boils down to accurately counting the number of microstates associated with a given macrostate. In the canonical and grand-canonical ensembles microstates are 'weighted' by the appropriate Boltzmann factors, and the counting is formalised in the canonical and grand-canonical partition functions. Anyone who has taken a course in statistical physics will remember the subtleties associated with determining the number of ways a given number of Fermions or Bosons can be distributed across a set of energy states.

The Boltzmann weights in the partition functions for the different ensembles can be obtained from the extremisation principles of classical thermodynamics. Depending on circumstances, a system will tend to the state of extremal entropy, Helmholtz free energy or grand potential for example. Using this, the idea of equal a priori probabilities in isolated systems, and the entropy $S = k \ln \Omega$ for the combination of the system

of interest and the relevant surrounding baths, the canonical and grand-canonical partition functions can be obtained[8].

It is no surprise that connections can be established to other areas facing counting and extremisation problems. Examples are the closely related areas of 'constrained optimisation' and 'constraint satisfaction' in mathematics and computer science. Constrained optimisation problems involve finding the minimum (or maximum) of a function $f(\boldsymbol{x})$, subject to constraints on the variable $\boldsymbol{x} = (x_1, \ldots, x_n)$. These constraints can come as a set of equations connecting the x_i, as inequalities or as a combination of equalities and inequalities. One example is graph partitioning, i.e. the problem of partitioning the set of nodes of a network into a given number of subsets while minimising the number of edges between these subsets. Another instance of constrained optimisation is the 'knapsack problem' (see e.g. [148]).

Constraint satisfaction problems are problems in which the variable $\boldsymbol{x}$ must satisfy a set of constraints. Often these are formulated on graphs. A good example is the graph colouring problem (more precisely, we describe the vertex colouring problem). Assume we have a graph consisting of N nodes, connected by a set of links. The problem now consists of assigning a colour to each node so that no two neighbouring nodes have the same colour. The x_i, $i = 1, \ldots, N$ here represent the colours, and the constraints are local (each node of degree $k \geq 1$ in the network leads to a constraint involving this node and its neighbours). Other constraint satisfaction problems are the celebrated travelling salesman and random k-satisfiability problems [149–151].

The methods used to address such problems include techniques such as message-passing (including belief propagation) and the cavity method. These tools are also relevant for spin glass problems in physics, and it is therefore natural for a statistical physicist to work at the interface of optimisation and computer science. Physicists have also been instrumental in characterising the typical properties of instances of constraint satisfaction or optimisation problems, this will be discussed further below (Section 2.7.4).

It is hard to avoid seeing possible connections between NMA and these classes of problems. For example, it is natural to ask if ranking in NMA can be phrased as

[8]We highlight the alternative information-theoretic approach to statistical physics, beautifully introduced by Jaynes [146, 147].

a constrained optimisation or satisfaction problem. It is perhaps useful to think of the meta-analytic network as a bipartite graph, with one type of node representing treatment options and the other trials [e.g. Figure 2.2 (d)]. If there are N treatment nodes in the graph, then the ranking task consists of assigning the ranks $1, \ldots, N$ to those nodes, while satisfying constraints set by the trials or minimising a cost defined by the trials (each trial provides noisy information on the relative ranks of some treatments). The NMA problem hence falls into a class of broader problems: We start from a set of N objects (the treatment options), who each have an unknown quality (treatment effect). We have noisy estimates of the relative qualities of subsets of these objects (from the trials), and we now wish to estimate the intrinsic property of each object. If this is not possible, what can we say about the relative comparison between pairs of objects, or a ranking of the objects in terms of quality?

While details would have to be thought through carefully we think that it is well possible that parallels between NMA and existing optimisation or satisfaction problems can be established. There is then potential for statistical physicists to contribute via the methods mentioned above (message passing, cavity method, etc.). We think this is an exciting perspective for future work.

2.7.4 Network meta-analysis and disordered systems

Related to the previous item there may be interest in looking at the typical properties of the NMA problem and its solution. This is a common approach in constrained optimisation and satisfaction problems. Instead of looking at single instances of these problems, one assumes the network is drawn from an ensemble of graphs, and then asks what the average or typical properties of problems in this ensemble are. For example, how many solutions to the optimisation or satisfaction problem are there, what subspace do they form in the overall state space (e.g. is there one connected manifold of solutions vs fragmented clusters) and what is the typical quality of the solution (i.e. how many of the constraints can be fulfilled)?

To answer these questions, an average over assignments of the graph and constraints needs to be taken. One can then make use of tools from the physics of spin glasses and disordered systems, such as the replica method, dynamic mean field theory or cavity approaches. This allows one to characterise the energy landscape associated with these

problems, the geometry of the solution space, and most importantly the performance of algorithms to find solutions to the optimisation or satisfaction problem.

It is conceivable that a similar approach could be taken in NMA. Based on known statistical features of meta-analytic graphs [144, 152] one could define an ensemble of random NMA problems (i.e. the configuration of trials, treatment options and their connections is drawn from a distribution). One could then try to assess how features of the ensemble (e.g. connectivity, regularity etc.) affect the outcome of the NMA problems in the ensemble. This would allow one to step away from *single* instances, and instead to say more about *typical* cases.

2.7.5 Machine learning approaches to systematic reviews and Bayesian MCMC

The field of machine learning interfaces with statistical physics, and there is increasing interest by physicists in machine learning methods. There are multiple ways in which machine learning can contribute to the field of NMA, and this therefore defines another point of contact with statistical mechanics.

One way in which machine learning can be used in NMA relates to the Bayesian approach in Section 2.3. Methods from machine learning have been proposed as alternatives to MCMC in general. This includes techniques such as expectation propagation [153], variational Bayesian inference [154] and integrated nested Laplace approximations [155, 156].

A perhaps even stronger link to machine learning presents itself in the process of data acquisition for an NMA. In this paper, we have so far focused on the procedure of carrying out an NMA given a set of data from multiple trials. In reality, the process of compiling the data starts by performing a 'systematic review' of the existing trials for a particular medical problem. This is a systematic screening of the literature, using well defined procedures, followed by a process to decide which trials are adopted for the NMA. This decision making again uses well defined protocols and criteria. As part of this process a large number of journal articles must be searched, and appropriate data must be identified and extracted.

This obviously lends itself to automation, which speeds up the process, saves

resources, and removes human error and inconsistencies that arise when a team of multiple researchers collects the data. Machine learning methods have indeed been employed to automate or semi-automate this process [157–159].

Given these contact points between NMA and machine learning, we think that researchers working at the interface of statistical physics and machine learning may find interest in applying the ideas and methods they are familiar with to the field of evidence synthesis. Their experience may then lead to the development of algorithms that improve on existing methods.

2.7.6 Simulation techniques

Simulation techniques play an important role in NMA. Most evidently, Monte-Carlo sampling is necessary to carry out a Bayesian NMA (Section 2.3). Further, 'simulation studies' can provide insight into NMA [160]. These are studies in which trial data is generated synthetically (simulated). This allows one, for example, to test how NMA fares on different networks, and how different features of the graph affect NMA outcomes [142, 161, 162]. The simulation proceeds along the random effects model described in Section 2.2.4, we note again the hierarchical structure and the different levels of randomness. In simulation studies one typically has to look at many instances of networks, and average over a large number of realisations of synthetic data. Efficient simulation methods for the generation of data are therefore key.

Statistical physicists are obviously familiar with simulation methods for random processes. Acknowledging that simulation studies are an established part of statistics (and consequently, that there is significant existing expertise), this defines another point of contact, and a prospective avenue for statistical physicists to contribute to the field of NMA and evidence synthesis more generally.

2.7.7 Meta-analysis in particle physics

As a final, more speculative thought, we note that meta-analysis is used in particle physics to obtain the best estimates of particle properties such as masses, widths and lifetimes [163]. Expertise developed in this area might also be useful for meta-analysis and network meta-analysis in medical statistics.

2.8 Summary

Most of the existing NMA literature is naturally written by medical statisticians for medical statisticians, or for researchers actively using NMA tools and software packages in clinical practice. As a consequence, it is not easy to find an account of the essentials of NMA presented in a language physicists would be used to. The objective of this perspective review is to make a first step towards rectifying this. We hope the paper is a useful introduction to network meta-analysis for statistical physicists. It should be noted that the paper was also written by statistical physicists. This means that we work from a limited perspective. The selection of topics and the presentation is subject to personal bias.

Naturally, we could only include what we considered to be the most essential aspects of NMA. Topics which could have been covered in a more extensive review include inconsistency [27], individual participant data [164], multi-component interventions [165], multiple outcomes [166], bias adjustment methods [18] and goodness-of-fit assessment [17]. In making our selection of the material for this paper, we aimed to focus on the concepts, ideas and methods, which would best enable the reader to access the wider literature. We tried to write a self-contained systematic introduction which could serve as a starting point for the reader to then explore the field more effectively.

We also hope that the paper highlights the importance of the area of evidence synthesis. Our review is successful if it excites others and if it convinces the members of the community that network meta-analysis is a field in which statistical mechanics can make a difference.

2.9 Appendix A: Estimating within-study variances and correlations of observed treatment effects

In this appendix we discuss how the matrix $\mathbf{V}_i$ in Equation (2.41) [see Section 2.4.1] is estimated from trial data. This matrix describes the variance and correlations within a trial due to fluctuations arising from the finite number of subjects in the different arms of the trial (i.e. sampling errors).

2.9.1 General setup and estimating the variance of observed treatment effects

For binomial data, the observed relative treatment effects in trial i are given by the log odds ratios,

$$y_{i,1\ell} = \text{logit}(\hat{p}_{i,\ell}) - \text{logit}(\hat{p}_{i,1}) = \ln \frac{\hat{p}_{i,\ell}}{1-\hat{p}_{i,\ell}} - \ln \frac{\hat{p}_{i,1}}{1-\hat{p}_{i,1}}, \tag{2.101}$$

where $\hat{p}_{i,\ell} = r_{i,\ell}/n_{i,\ell}$ is the proportion of events in arm ℓ of trial i. The variances, $\sigma^2_{i,1\ell}$, and covariances, $\text{Cov}(y_{i,1\ell}, y_{i,1\ell'})$, associated with these observations define the covariance matrix $\mathbf{V}_i$ in Equation (2.41). These values can be estimated from the data. To do so we first work out the random sampling variance associated with the values $\hat{p}_{i,\ell}$.

The number of events measured in arm ℓ of trial i is a binomial random variable $r_{i,\ell} \sim \text{Bin}(n_{i,\ell}, p_{i,\ell})$. It has mean $\mathbb{E}(r_{i,\ell}) = n_{i,\ell}p_{i,\ell}$ and variance $\text{Var}(r_{i,\ell}) = n_{i,\ell}p_{i,\ell}(1-p_{i,\ell})$. Using $\text{Var}(bx) = b^2\text{Var}(x)$ (for random variable x and constant b), we find

$$\text{Var}(\hat{p}_{i,\ell}) = \frac{p_{i,\ell}(1-p_{i,\ell})}{n_{i,\ell}}. \tag{2.102}$$

Assuming $n_{i,\ell}$ is large, and propagating the errors for the logit function to linear order then leads to

$$\begin{aligned} \text{Var}[\text{logit}(\hat{p}_{i,\ell})] &= \left(\frac{\partial \,\text{logit}(\hat{p}_{i,\ell})}{\partial \hat{p}_{i,\ell}}\right)^2 \text{Var}(\hat{p}_{i,\ell}) \\ &= \left(\frac{1}{\hat{p}_{i,\ell}(1-\hat{p}_{i,\ell})}\right)^2 \frac{p_{i,\ell}(1-p_{i,\ell})}{n_{i,\ell}}. \end{aligned} \tag{2.103}$$

We get an estimate of this variance by setting $p_{i,\ell} = \hat{p}_{i\ell}$,

$$\widehat{\text{Var}}[\text{logit}(\hat{p}_{i,\ell})] = [n_{i,\ell}\hat{p}_{i,\ell}(1-\hat{p}_{i,\ell})]^{-1} = \left[r_{i,\ell}\left(1-\frac{r_{i,\ell}}{n_{i,\ell}}\right)\right]^{-1}. \tag{2.104}$$

Since the values $\hat{p}_{i,\ell}$ for different ℓ are independent, the variance of $y_{i,1\ell}$ is $\sigma^2_{i,1\ell} = \text{Var}(y_{i,1\ell}) = \text{Var}[\text{logit}(\hat{p}_{i,\ell})] + \text{Var}[\text{logit}(\hat{p}_{i,1})]$ which we can estimate using Equation (2.104),

$$\hat{\sigma}^2_{i,1\ell} = \left[r_{i,\ell}\left(1-\frac{r_{i,\ell}}{n_{i,\ell}}\right)\right]^{-1} + \left[r_{i,1}\left(1-\frac{r_{i,1}}{n_{i,1}}\right)\right]^{-1} \tag{2.105}$$

[5, 28, 72]. The conventional assumption is that the estimates $\hat{\sigma}^2_{i,1\ell}$ can be used in the matrix $\mathbf{V}_i$ in Equation (2.41) in place of the true variances [7, 72].

2.9.2 Estimating correlations between different observed treatment effects in the trial

The observations $y_{i,1\ell}$ of the relative treatment effects within a trial are correlated because they all involve the same common treatment arm $t_{i,1}$. To calculate the covariance between the observations we start from the general relation

$$\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\mathrm{Cov}(A, B), \tag{2.106}$$

for two random variables A and B. Setting $A = y_{i,1\ell}$ and $B = y_{i,1\ell'}$ for $\ell \neq \ell'$ we find

$$\mathrm{Cov}(y_{i,1\ell}, y_{i,1\ell'}) = \frac{1}{2}\left[\mathrm{Var}(y_{i,1\ell}) + \mathrm{Var}(y_{i,1\ell'}) - \mathrm{Var}(y_{i,1\ell} - y_{i,1\ell'})\right]. \tag{2.107}$$

We evaluate the final term on the right hand side of Equation (2.107) using Equation (2.101),

$$\begin{aligned}
\mathrm{Var}(y_{i,1\ell} - y_{i,1\ell'}) &= \mathrm{Var}\left\{[\mathrm{logit}(\hat{p}_{i,\ell}) - \mathrm{logit}(\hat{p}_{i,1})] - [\mathrm{logit}(\hat{p}_{i,\ell'}) - \mathrm{logit}(\hat{p}_{i,1})]\right\} \\
&= \mathrm{Var}\left[\mathrm{logit}(\hat{p}_{i,\ell}) - \mathrm{logit}(\hat{p}_{i,\ell'})\right] \\
&= \mathrm{Var}[\mathrm{logit}(\hat{p}_{i,\ell})] + \mathrm{Var}[\mathrm{logit}(\hat{p}_{i,\ell'})],
\end{aligned} \tag{2.108}$$

where in the last line we have used the fact that the absolute outcomes in different arms within a trial are independent. Recalling that $\mathrm{Var}(y_{i,1\ell}) = \mathrm{Var}[\mathrm{logit}(\hat{p}_{i,\ell})] + \mathrm{Var}[\mathrm{logit}(\hat{p}_{i,1})]$ we find [17]

$$\mathrm{Cov}(y_{i,1\ell}, y_{i,1\ell'}) = \mathrm{Var}[\mathrm{logit}(\hat{p}_{i,1})], \tag{2.109}$$

which we can estimate via Equation (2.104),

$$\widehat{\mathrm{Cov}}(y_{i,1\ell}, y_{i,1\ell'}) = \left[r_{i,1}\left(1 - \frac{r_{i,1}}{n_{i,1}}\right)\right]^{-1}. \tag{2.110}$$

Again, the convention is to assume that the estimate of the covariance in Equation (2.110) can be used in place of the true covariance [17].

2.9.3 Limitations of estimating within study variance

It is evident from Equations (2.101) and (2.105) that the values $y_{i,1\ell}$ and $\hat{\sigma}^2_{i,1\ell}$ are correlated (both expressions depend on the variables $(r_{i,\ell}, n_{i,\ell})$ and $(r_{i,1}, n_{i,1})$) [72]. This

causes a systematic relationship between the magnitude and weight of the observations which leads to a bias on the overall effect estimates [15, 72].

Estimation of $y_{i,1\ell}$ and $\sigma^2_{i,1\ell}$ from the observed data also causes problems when events are observed for either no patients in a trial arm, or for all patients in an arm. By inspection of Equations (2.101) and (2.105), we notice that if $r_{i,\ell} = 0$ or $r_{i,\ell} = n_{i,\ell}$ then $y_{i,1\ell}$ and $\hat{\sigma}^2_{i,1\ell}$ will be undefined. A common ad-hoc method to avoid this problem is to add a value of 0.5 to every $r_{i,\ell}$ and $n_{i,\ell}$. This has been found to produce biased estimates of effect size [15, 167, 168]. In fact, this limitation, along with the assumption that within-study variances are known, is the main criticism of the frequentist inverse-variance approach to NMA [169, 170]. Alternative methods such as the Mantel-Haenszel method and generalised linear mixed models (GLMM) have been recommended [14, 15].

2.10 Appendix B: Expectation of Q under the random effects model

2.10.1 Pairwise meta-analysis

In this section we evaluate the expectation of Q,

$$\mathbb{E}_{\text{RE}}(Q) = \mathbb{E}_{\text{RE}}\left(\sum_{i=1}^{M} a_i(y_i - \hat{y})^2\right), \tag{2.111}$$

where $\hat{y}$ is defined in Equation (2.66) and the observations y_i are assumed to follow the random effects (RE) model,

$$y_i \sim \mathcal{N}(d, \sigma_i^2 + \tau^2). \tag{2.112}$$

Using $\mathbb{E}_{\text{RE}}(y_i - \hat{y}) = 0$, we find

$$\mathbb{E}_{\text{RE}}\left(\sum_{i=1}^{M} a_i(y_i - \hat{y})^2\right) = \sum_{i=1}^{M} a_i \text{Var}\left(y_i - \hat{y}\right). \tag{2.113}$$

We can then obtain the variance of $y_i - \hat{y}$ using

$$\text{Var}\left(y_i - \hat{y}\right) = \text{Var}(y_i) + \text{Var}(\hat{y}) - 2\text{Cov}(y_i, \hat{y}). \tag{2.114}$$

From the RE model in Equation (2.112) we know $\text{Var}(y_i) = \sigma_i^2 + \tau^2$. We now wish to obtain the second and third terms of Equation (2.114) in terms of $\text{Var}(y_i)$. To do so we

use standard properties of variances and covariances, and the fact that $\mathrm{Cov}(y_i, y_j) = 0$ for $i \neq j$ (for a pairwise meta-analysis, the observations y_i are independent).

To calculate $\mathrm{Var}(\hat{y})$ we use Equation (2.66) in the main paper and find

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{M} a_i y_i}{\sum_{i=1}^{M} a_i}\right) = \left(\frac{1}{\sum_{i=1}^{M} a_i}\right)^2 \sum_{i=1}^{M} a_i^2 \mathrm{Var}(y_i). \tag{2.115}$$

For $\mathrm{Cov}(y_i, \hat{y})$ we have

$$\begin{aligned}
\mathrm{Cov}(y_i, \hat{y}) &= \mathrm{Cov}\left(y_i, \frac{\sum_{i=1}^{M} a_i y_i}{\sum_{i=1}^{M} a_i}\right) \\
&= \frac{1}{\sum_{i=1}^{M} a_i}\left(a_i \mathrm{Cov}(y_i, y_i) + \sum_{j \neq i}^{M} a_j \mathrm{Cov}(y_i, y_j)\right) \\
&= \frac{1}{\sum_{i=1}^{M} a_i} a_i \mathrm{Var}(y_i).
\end{aligned} \tag{2.116}$$

Substituting these results into Equation (2.114) yields

$$\mathrm{Var}\left(y_i - \hat{y}\right) = \mathrm{Var}(y_i) + \frac{\sum_{j=1}^{M} a_j^2 \mathrm{Var}(y_j)}{\left(\sum_{j=1}^{M} a_j\right)^2} - \frac{2 a_i \mathrm{Var}(y_i)}{\sum_{j=1}^{M} a_j}. \tag{2.117}$$

Now substituting this into Equation (2.113) and using $\mathrm{Var}(y_i) = \sigma_i^2 + \tau^2$ we find the expectation of Q under the random effects model to be

$$\begin{aligned}
\mathbb{E}_{\mathrm{RE}}(Q) &= \sum_{i=1}^{M} a_i \mathrm{Var}(y_i) + \sum_{i=1}^{M} a_i \frac{\sum_{j=1}^{M} a_j^2 \mathrm{Var}(y_j)}{\left(\sum_{j=1}^{M} a_j\right)^2} - \frac{2\sum_{i=1}^{M} a_i^2 \mathrm{Var}(y_i)}{\sum_{j=1}^{M} a_j} \\
&= \sum_{i=1}^{M} a_i(\sigma_i^2 + \tau^2) - \frac{\sum_{i=1}^{M} a_i^2(\sigma_i^2 + \tau^2)}{\sum_{i=1}^{M} a_i} \\
&= \tau^2\left(\sum_{i=1}^{M} a_i - \frac{\sum_{i=1}^{M} a_i^2}{\sum_{i=1}^{M} a_i}\right) + \left(\sum_{i=1}^{M} a_i \sigma_i^2 - \frac{\sum_{i=1}^{M} a_i^2 \sigma_i^2}{\sum_{i=1}^{M} a_i}\right).
\end{aligned} \tag{2.118}$$

This is the result quoted in Equation (2.69) in the main paper.

2.10.2 Network meta-analysis

In this section we evaluate the expectation of Q under the random effects model for network meta-analysis. We now have [Equation (2.71)]

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{V}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{y}}) \tag{2.119}$$

and [Equation (2.72)]

$$\hat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \boldsymbol{y}. \tag{2.120}$$

To simplify Equation (2.119) we follow Jackson et al (2016) [89] and define the matrix

$$\mathbf{A} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1} \tag{2.121}$$

such that

$$\boldsymbol{y} - \hat{\boldsymbol{y}} = \mathbf{V}\mathbf{A}\boldsymbol{y}. \tag{2.122}$$

Therefore,

$$\begin{aligned} Q &= (\mathbf{V}\mathbf{A}\boldsymbol{y})^\top\mathbf{V}^{-1}(\mathbf{V}\mathbf{A}\boldsymbol{y}) \\ &= \boldsymbol{y}^\top\mathbf{A}\mathbf{V}\mathbf{A}\boldsymbol{y} \end{aligned} \tag{2.123}$$

since both $\mathbf{V}$ and $\mathbf{A}$ are symmetric. The former is symmetric by definition, the latter can be shown to be symmetric by taking the transpose of the right hand side of Equation (2.121). By explicit evaluation we find

$$\begin{aligned} \mathbf{A}\mathbf{V}\mathbf{A} &= \mathbf{A}[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}] \\ &= \mathbf{A} - [\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}]\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1} \\ &= \mathbf{A} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1} + \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1} \\ &= \mathbf{A} \end{aligned} \tag{2.124}$$

such that

$$Q = \boldsymbol{y}^\top\mathbf{A}\boldsymbol{y}. \tag{2.125}$$

As in the pairwise case we take the expectation of Q under the random effects model. Defining $\boldsymbol{\xi} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$, we re-write the RE model [Equation (2.39) in the main paper] as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{d} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{V} + \boldsymbol{\Sigma}). \tag{2.126}$$

This leads to

$$\begin{aligned} \mathbb{E}_{\mathrm{RE}}(Q) &= \mathbb{E}_{\mathrm{RE}}[(\mathbf{X}\boldsymbol{d} + \boldsymbol{\xi})^\top\mathbf{A}(\mathbf{X}\boldsymbol{d} + \boldsymbol{\xi})] \\ &= \mathbb{E}_{\mathrm{RE}}(\boldsymbol{d}^\top\mathbf{X}^\top\mathbf{A}\mathbf{X}\boldsymbol{d} + \boldsymbol{d}^\top\mathbf{X}^\top\mathbf{A}\boldsymbol{\xi} + \boldsymbol{\xi}^\top\mathbf{A}\mathbf{X}\boldsymbol{d} + \boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}). \end{aligned} \tag{2.127}$$

By explicit evaluation we find

$$\begin{aligned} \mathbf{X}^\top\mathbf{A} &= \mathbf{X}^\top\mathbf{V}^{-1} - \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1} = 0 \\ \mathbf{A}\mathbf{X} &= \mathbf{V}^{-1}\mathbf{X} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} = 0, \end{aligned} \tag{2.128}$$

and hence,

$$\mathbb{E}_{\mathrm{RE}}(Q) = \mathbb{E}_{\mathrm{RE}}(\boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi}). \tag{2.129}$$

For any vector $\boldsymbol{z}$ with mean $\boldsymbol{\mu}_z$ and covariance matrix $\mathbf{V}_z$, one has $\mathbb{E}(\boldsymbol{z}^\top \mathbf{B} \boldsymbol{z}) = \mathrm{tr}(\mathbf{B}\mathbf{V}_z) + \boldsymbol{\mu}_z^\top \mathbf{B} \boldsymbol{\mu}_z$ where $\mathbf{B}$ is a square matrix and tr(.) indicates the trace of a matrix. This identity can be checked directly, see also [171] (Theorem 4, pg 75, Chapter 2).

The vector $\boldsymbol{\xi}$ in Equation (2.129) has mean $\mathbf{0}$ and covariance matrix $\mathbf{V}+\boldsymbol{\Sigma}$, therefore

$$\begin{aligned} \mathbb{E}_{\mathrm{RE}}(Q) &= \mathrm{tr}[\mathbf{A}(\mathbf{V} + \boldsymbol{\Sigma})] \\ &= \mathrm{tr}(\mathbf{AV}) + \tau^2 \mathrm{tr}(\mathbf{AP}). \end{aligned} \tag{2.130}$$

In the last step we have written $\boldsymbol{\Sigma} = \tau^2 \mathbf{P}$ where $\mathbf{P}$ is a block diagonal matrix. Each $(m_i - 1) \times (m_i - 1)$ block (representing trial i) has diagonal elements equal to 1 and off-diagonal elements equal to 1/2. All other elements are zero. Using the fact the trace is invariant under cyclic permutations we find

$$\begin{aligned} \mathrm{tr}(\mathbf{AV}) &= \mathrm{tr}\left[\mathbf{V}^{-1}\mathbf{V} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{V}\right] \\ &= \mathrm{tr}\left[\mathbf{I}_{\sum_i (m_i - 1)}\right] - \mathrm{tr}\left[\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\right] \\ &= \sum_{i=1}^{M}(m_i - 1) - \mathrm{tr}\left[(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X}\right] \\ &= \sum_{i=1}^{M}(m_i - 1) - \mathrm{tr}\left[\mathbf{I}_{N-1}\right] \\ &= \sum_{i=1}^{M}(m_i - 1) - (N - 1) \end{aligned} \tag{2.131}$$

where $\sum_{i=1}^{M}(m_i - 1)$ is the lateral dimension of $\mathbf{V}$ (the number of observations in $\boldsymbol{y}$) and $N-1$ is the lateral dimension of $\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X}$ (the number of mean relative treatment effects we wish to estimate). The expression in Equation (2.131) is the number of degrees of freedom associated with the regression (that is, the difference between the number of data points and the number of parameters[9]).

[9]This can be understood via a simple example: imagine calculating the mean from 10 values. Here the number of data points is 10 and the number of parameters is 1. We only need 9 of those values plus the mean to fully specify the 10 values. The number of degrees of freedom is then 9 (which is equal to the number of data points minus the number of parameters).

2.11 Appendix C: Bias in maximum likelihood variance estimation

Consider a random variable $y = (y_1, \ldots, y_N)^\top$ with normal distribution, $y \sim \mathcal{N}(\mu, \sigma^2)$. The likelihood function is then

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y_i-\mu)^2/2\sigma^2}. \tag{2.132}$$

Maximising the likelihood function with respect to μ and σ^2 leads to the expressions

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.133}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \mu)^2. \tag{2.134}$$

The result of the joint maximisation with respect to μ and σ therefore leads to the maximum likelihood estimators,

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.135}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2, \tag{2.136}$$

that is, we have to substitute the maximum likelihood estimate $\hat{y}$ into the expression for $\hat{\sigma}^2$.

To show that the expected value of the variance estimator $\hat{\sigma}^2$ is not equal to the true variance σ^2 we re-write Equation (2.136) as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} [(y_i - \mu) - (\hat{y} - \mu)]^2, \tag{2.137}$$

which, after some rearranging and using the fact that $\hat{y} - \mu = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)$, gives

$$\hat{\sigma}^2 = \left[\frac{1}{N}\sum_{i=1}^{N} (y_i - \mu)^2\right] - (\hat{y} - \mu)^2. \tag{2.138}$$

The expectation of $\hat{\sigma}^2$ is then

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} (y_i - \mu)^2\right] - \mathbb{E}\left[(\hat{y} - \mu)^2\right]. \tag{2.139}$$

The first term is the true variance σ^2. The second term is the variance of $\hat{y}$,

$$\mathbb{E}\left[(\hat{y} - \mu)^2\right] = \mathrm{Var}(\hat{y}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}(y_i) = \frac{\sigma^2}{N}. \tag{2.140}$$

Therefore,

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2, \tag{2.141}$$

indicating that the ML variance estimator is biased downwards by $\frac{\sigma^2}{N}$. For large samples $N \to \infty$, this bias becomes negligible.

2.12 Appendix D: Adjusting multi-arm trial weights using a result from electrical network theory

Here we explain the method for adjusting variances associated with measurements in a network meta-analysis in order to account for correlations introduced by multi-arm trials [49, 115]. This is based on the method described in Gutman and Xiao (2004) [116] for reconstructing the individual resistances in an electrical network from the effective resistances between pairs of nodes.

We focus on a multi-arm trial i which compares m_i treatments. This trial yields a total of $q_i = \frac{m_i(m_i-1)}{2}$ treatment effect estimates and associated variances $v_{i,ab}$. For a random effects model these variances are $v_{i,ab} = \sigma^2_{i,ab} + \hat{\tau}^2$ where $\hat{\tau}^2$ is an estimate of the between trial heterogeneity. In a fixed effect model $v_{i,ab} = \sigma^2_{i,ab}$. We write these variances in an $m_i \times m_i$ matrix, $\tilde{\mathbf{V}}_i$. We label this matrix $\tilde{\mathbf{V}}_i$ to distinguish it from the $(m_i - 1) \times (m_i - 1)$ matrix $\mathbf{V}_i$ defined in the main paper. Each row and column of $\tilde{\mathbf{V}}_i$ represents a treatment in trial i. The diagonal elements are equal to zero and each off-diagonal element is the variance associated with the comparison of the corresponding pair of treatments. For example, in a multi-arm trial i that compares treatments $\{T_1, T_2, T_3\}$, the variance matrix is

$$\tilde{\mathbf{V}}_i = \begin{pmatrix} 0 & v_{i,T_1T_2} & v_{i,T_1T_3} \\ v_{i,T_1T_2} & 0 & v_{i,T_2T_3} \\ v_{i,T_1T_3} & v_{i,T_2T_3} & 0 \end{pmatrix}. \tag{2.142}$$

The aim now is to reconstruct the (inverse-variance) weights for a set of three two-arm trials that yield network estimates of relative treatment effects whose variances are equal to the variances in $\tilde{\mathbf{V}}_i$. To this end, we use a method from electrical theory for back calculation of edge resistances given a set of effective resistances [49, 116].

We saw in Section 2.6.1 that the effective resistances in an electrical network are related to the pseudo-inverse of the Laplacian. In NMA, effective resistances are associated with the variances $\tilde{\mathbf{V}}_i$ observed in the multi-arm trial. A result from electrical network theory is that we can construct the pseudo-inverse of the Laplacian directly from the effective resistances [116]. Using this result for an m_i-armed trial with 'effective' variances $\tilde{\mathbf{V}}_i$ gives the $m_i \times m_i$ matrix

$$\mathbf{L}_i^+ = -\frac{1}{2}\left(\tilde{\mathbf{V}}_i - \frac{1}{m_i}(\tilde{\mathbf{V}}_i\mathbf{O}_i + \mathbf{O}_i\tilde{\mathbf{V}}_i) + \frac{1}{m_i^2}\mathbf{O}_i\tilde{\mathbf{V}}_i\mathbf{O}_i\right) \tag{2.143}$$

where $\mathbf{O}_i$ is an $m_i \times m_i$ matrix of ones [49]. An equivalent more compact expression (used in [115]) is

$$\mathbf{L}_i^+ = -\frac{1}{2m_i^2}\mathbf{B}_i^\top\mathbf{B}_i\tilde{\mathbf{V}}_i\mathbf{B}_i^\top\mathbf{B}_i \tag{2.144}$$

where $\mathbf{B}_i$ is the edge-incidence matrix for trial i that describes what edges (treatment comparisons) are present in the trial [115]. $\mathbf{B}_i$ has dimensions $q_i \times m_i$, where we recall $q_i = \frac{m_i(m_i-1)}{2}$. Each row of $\mathbf{B}_i$ represents a pairwise comparison in the trial, and each column represents a treatment. There is a 1 in the column corresponding to the baseline treatment for that comparison and a -1 in the column representing the treatment compared to that baseline. All other entries are zero.

Once we have the pseudo-inverse of the Laplacian we can work out the Laplacian using $\mathbf{L}_i = (\mathbf{L}_i^+)^+$ and [49, 172]

$$\mathbf{L}_i^+ = (\mathbf{L}_i - \frac{1}{m_i}\mathbf{O}_i)^{-1} + \frac{1}{m_i}\mathbf{O}_i, \tag{2.145}$$

$$(\mathbf{L}_i^+)^+ = (\mathbf{L}_i^+ - \frac{1}{m_i}\mathbf{O}_i)^{-1} + \frac{1}{m_i}\mathbf{O}_i. \tag{2.146}$$

In an electrical network the graph Laplacian is defined by the individual (physical) resistors, $L_{ab} = -R_{ab}^{-1}$ for $a \neq b$, and $L_{aa} = \sum_b R_{ab}^{-1}$. Therefore the values R_{ab} are obtained by inspection of the off diagonal elements of $\mathbf{L}$. Similarly, in the NMA context, the adjusted (inverse-variance) edge weights for multi-arm trial i can be obtained from the off diagonal elements of the Laplacian $\mathbf{L}_i$. For the three-arm trial,

$$\mathbf{L}_i = \begin{pmatrix} \tilde{w}_{i,T_1T_2} + \tilde{w}_{i,T_1T_3} & -\tilde{w}_{i,T_1T_2} & -\tilde{w}_{i,T_1T_3} \\ -\tilde{w}_{i,T_1T_2} & \tilde{w}_{i,T_1T_2} + \tilde{w}_{i,T_2T_3} & -\tilde{w}_{i,T_2T_3} \\ -\tilde{w}_{i,T_1T_3} & -\tilde{w}_{i,T_2T_3} & \tilde{w}_{i,T_1T_3} + \tilde{w}_{i,T_2T_3} \end{pmatrix}, \tag{2.147}$$

where $\tilde{w}_{i,ab}$ are the adjusted edge weights that describe the inverse-variances for a network of two-arm trials which is equivalent to the multi-arm trial.

Bibliography

[1] G. Parisi, "Statistical physics and biology", Phys. World. **6**, 42–47 (1993).

[2] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics", Rev. Mod. Phys. **81**, 591–646 (2009).

[3] D. Stauffer, "Introduction to statistical physics outside physics", Physica A **336**, Proceedings of the XVIII Max Born Symposium "Statistical Physics Outside Physics", 1–5 (2004).

[4] M. Gallegati, S. Keen, T. Lux, and P. Ormerod, "Worrying trends in econophysics", Physica A **370**, Econophysics Colloquium, 1–6 (2006).

[5] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials", Control. Clin. Trials **7**, 177–188 (1986).

[6] T. C. Smith, D. J. Spiegelhalter, and A. Thomas, "Bayesian approaches to random effects meta analysis: a comparative study", Stat. Med. **14**, 2685–2699 (1995).

[7] J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter, "A re-evaluation of random effects meta-analysis", J. R. Stat. Soc. **172**, 137–159 (2009).

[8] F. S. Tonin, I. Rotta, A. M. Mendes, and R. Pontarolo, "Network meta-analysis: a technique to gather evidence from direct and indirect comparisons", Pharm. Pract. (Granada) **15**, 943 (2017).

[9] S. Dias, A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton, *Network meta-analysis for decision making* (Wiley, Oxford, UK, 2018).

[10] A. Boland, Y. Dundar, A. Bagust, A. Haycox, R. Hill, R. Mujica Mota, T. Walley, and R. Dickson, "Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation", Health. Technol. Asses. **7**, 1–136 (2003).

[11] E. C. Keeley, J. A. Boura, and C. L. Grines, "Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials", Lancet **361**, 13–20 (2003).

[12] S. Dias, N. J. Welton, D. M. Caldwell, and A. E. Ades, "Checking consistency in mixed treatment comparison meta-analysis", Stat. Med. **29**, 932–944 (2010).

[13] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn, *WinBUGS User Manual: version 1.4*, MRC Biostatistics Unit, University of Cambridge, 2003.

[14] O. Efthimiou, G. Rücker, G. Schwarzer, J. P. T. Higgins, M. Egger, and G. Salanti, "Network meta-analysis of rare events using the Mantel-Haenszel method", Stat. Med. **38**, 2992–3012 (2019).

[15] T. Stijnen, T. H. Hamza, and P. Özdemir, "Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data", Stat. Med. **29**, 3046–3067 (2010).

[16] A. L. Davies, T. Papakonstantinou, A. Nikolakopoulou, G. Rücker, and T. Galla, "Network meta-analysis and random walks", Stat. Med., 1–24 (2022).

[17] 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*, Online, Last updated September 2016; Available from http://www.nicedsu.org.uk. Accessed March 2020, 2011.

[18] S. Dias, N. J. Welton, A. J. Sutton, and A. E. Ades, *NICE DSU Technical Support Document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment*, Online,

Last updated April 2012; Available from http://www.nicedsu.org.uk. Accessed December 2021, 2011.

[19] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, eds., *Cochrane handbook for systematic reviews of interventions*, 2nd ed. (Wiley, Chichester, UK, 2019).

[20] G. Salanti, "Indirect and mixed treatment comparison, network, or multiple treatments meta analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool", Res. Synth. Meth. **3**, 80–97 (2012).

[21] T. Greco, G. Landoni, G. Biondi-Zoccai, F. D'Ascenzo, and A. Zangrillo, "A Bayesian network meta-analysis for binary outcome: how to do it", Stat. Methods. Med. Res. **25**, 1757–1773 (2016).

[22] G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis, "Evaluation of networks of randomized trials", Stat. Methods. Med. Res. **17**, 279–301 (2008).

[23] J. P. Jansen, B. Crawford, G. Bergman, and W. Stam, "Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons", Value. Health **11**, 956–964 (2008).

[24] O. Efthimiou, T. P. Debray, G. van Valkenhoef, S. Trelle, K. Panayidou, K. G. Moons, J. B. Reitsma, A. Shang, and G. Salanti, "GetReal methods review group. GetReal in network meta-analysis: a review of the methodology", Res. Synth. Meth. **7**, 236–263 (2016).

[25] NICE, *Glossary*, Online, Available from https://www.nice.org.uk/Glossary. Accessed February 2022.

[26] T. Pisanski and M. Randić, "Bridges between geometry and graph theory", in Geometry at work: a collection of papers showing applications of geometry (2000), pp. 174–194.

[27] S. Dias, N. J. Welton, A. J. Sutton, D. M. Caldwell, G. Lu, and A. E. Ades, *NICE DSU Technical Support Document 4: inconsistency in networks of evidence based on randomised controlled trials*, Online, Last updated April 2014; Available from http://www.nicedsu.org.uk. Accessed December 2021, 2011.

[28] T. H. Hamza, H. C. van Houwelingen, and T. Stijnen, "The binomial distribution of meta-analysis was preferred to model within-study variability", J. Clin. Epidemiol. **61**, 41–51 (2008).

[29] G. Rücker, G. Schwarzer, J. Carpenter, and I. Olkin, "Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells", Stat. Med. **28**, 721–738 (2009).

[30] J. J. Deeks, "Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes", Stat. Med. **21**, 1575–1600 (2002).

[31] S. Dias, A. J. Sutton, A. E. Ades, and N. J. Welton, "Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials", Med. Decis. Making. **33**, 607–617 (2013).

[32] D. M. Eddy, V. Hasselblad, and R. Shachter, *Meta-analysis by the confidence profile method* (Academic Press, London, UK, 1992).

[33] G. Lu and A. E. Ades, "Assessing evidence inconsistency in mixed treatment comparisons", J. Am. Stat. Assoc. **101**, 447–459 (2006).

[34] H. Hong, H. Fu, K. L. Price, and B. P. Carlin, "Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment", Stat. Med. **34**, 2794–2819 (2015).

[35] S. Dias and A. E. Ades, "Absolute or relative effects? Arm-based synthesis of trial data", Res. Synth. Meth. **7**, 23–28 (2016).

[36] J. P. T. Higgins and A. Whitehead, "Borrowing strength from external trials in a meta-analysis", Stat. Med. **15**, 2733–2749 (1996).

[37] T. Lumley, "Network meta-analysis for indirect treatment comparisons", Stat. Med. **21**, 2313–2324 (2002).

[38] G. Lu and A. E. Ades, "Combination of direct and indirect evidence in mixed treatment comparisons", Stat. Med. **23**, 3105–3124 (2004).

[39] S. E. Seide, K. Jensen, and M. Kieser, "Simulation and data-generation of random-effects network meta-analysis of binary outcome", Stat. Med. **38**, 3288–3303 (2019).

[40] D. R. Cox, *Principles of statistical inference* (Cambridge University Press, Cambridge, UK, 2006).

[41] D. J. Bartholomew, "A comparison of some Bayesian and frequentist inferences", Biometrika **52**, 19–35 (1965).

[42] E. Wagenmakers, M. Lee, T. Lodewyckx, and G. J. Iverson, "Bayesian versus frequentist inference", in *Bayesian evaluation of informative hypotheses*, edited by H. Hoijtink, I. Klugkist, and P. A. Boelen (Springer, New York, NY, USA, 2008), pp. 181–207.

[43] F. J. Samaniego, *A comparison of the Bayesian and frequentist approaches to estimation* (Springer, New York, NY, USA, 2010).

[44] L. Hespanhol, C. S. Vallio, and B. T. Saragiotto, "Understanding and interpreting confidence and credible intervals around effect estimates", Braz. J. Phys. Ther. **23**, 290–301 (2019).

[45] M. E. Glickman and D. A. van-Dyk, "Basic Bayesian methods", in *Topics in biostatistics: methods in molecular biology*, Vol. 404, edited by W. T. Abrosius (Humana Press Inc, Totowa, NJ, USA, 2007) Chap. 16, pp. 319–338.

[46] S. Greenland, "Bayesian perspectives for epidemiological research: I. Foundations and basic methods", Int. J. Epidemiol. **35**, 765–775 (2006).

[47] I. R. White, R. M. Turner, A. Karahalios, and G. Salanti, "A comparison of arm-based and contrast-based models for network meta-analysis", Stat. Med. **38**, 5197–5213 (2019).

[48] I. R. White, J. K. Barrett, D. Jackson, and J. P. T. Higgins, "Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression", Res. Synth. Meth. **3**, 111–125 (2012).

[49] G. Rücker, "Network meta-analysis, electrical networks and graph theory", Res. Synth. Meth. **3**, 312–324 (2012).

[50] A. J. Franchini, S. Dias, A. E. Ades, J. P. Jansen, and N. J. Welton, "Accounting for correlation in network meta-analysis with multi-arm trials", Res. Synth. Meth. **3**, 142–160 (2012).

[51] D. Basu, "On the elimination of nuisance parameters", J. Am. Stat. Assoc. **72**, 355–366 (1977).

[52] W. DuMouchel, *Hierarchical Bayes Linear Models for Meta Analysis. Technical report number 27.* National Institute of Statistical Sciences, Available from https://www.niss.org/research/technical-reports. Accessed December 2021, 1994.

[53] G. Lu and A. E. Ades, "Modeling between trial variance structure in mixed treatment comparisons", Biostatistics **10**, 792–805 (2009).

[54] K. J. Rosenberger, A. Xing, M. Murad, H. Chu, and L. Lin, "Prior choices of between-study heterogeneity in contemporary Bayesian network meta-analyses: an empirical study", J. Gen. Intern. Med. **36**, 1049–1057 (2021).

[55] A. Gelman, "Prior distributions for variance parameters in hierarchical models", Bayesian Anal. **1**, 515–533 (1006).

[56] R. M. Turner, J. Davey, M. J. Clarke, S. G. Thompson, and J. P. T. Higgins, "Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews", Int. J. Epidemiol. **41**, 818–827 (2012).

[57] K. M. Rhodes, R. M. Turner, and J. P. T. Higgins, "Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data", J. Clin. Epidemiol. **68**, 52–60 (2015).

[58] R. M. Turner, C. P. Domínguez-Islas, D. Jackson, K. M. Rhodes, and I. R. White, "Incorporating external evidence on between-trial heterogeneity in network meta-analysis", Stat. Med. **38**, 1321–1335 (2019).

[59] M. Plummer, *JAGS version 4.3.0 user manual*, Online, Available from https://martynplummer.wordpress.com/. Accessed December 2021, 2017.

[60] Stan Development Team, *Stan user's guide version 2.23*, Online, Available from https://mc-stan.org/. Accessed December 2021, 2019.

[61] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", Biometrika **57**, 97–109 (1970).

[62] C. J. Geyer, "Chapter 1: Introduction to Markov chain Monte Carlo", in *Handbook of Markov chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X. Meng (CRC Press, Boca Raton, FL, USA, 2011), pp. 3–48.

[63] C. P. Robert, "The Metropolis–Hastings algorithm", in *Wiley StatsRef: statistics reference online*, edited by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels (American Cancer Society, 2015), pp. 1–15.

[64] C. P. Robert and G. Casella, *Monte Carlo statistical methods* (Springer, Science and Business Media, New York, NY, USA, 2004).

[65] R. Toral and P. Colet, *Stochastic numerical methods: an introduction for students and scientists* (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2014).

[66] A. Gelman, G. O. Roberts, and W. R. Gilks, "Efficient Metropolis jumping rules", in *Bayesian statistics 5*, edited by J. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford University Press, Oxford, UK, 1996), pp. 599–607.

[67] S. M. Lynch, *Introduction to applied Bayesian statistics and estimation for social scientists* (Springer, New York, NY, USA, 2007), pp. 77–130.

[68] G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various Metropolis-Hastings algorithms", Stat. Sci. **16**, 351–367 (2001).

[69] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed. (CRC Press, Boca Raton, FL, USA, 2013).

[70] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences", Stat. Sci. **7**, 457–511 (1992).

[71] S. Brooks and A. Gelman, "General methods for monitoring convergence of iterative simulations", J. Comput. Graph. Stat. **7**, 434–455 (1998).

[72] B. Chang, C. Waternaux, and S. Lipsitz, "Meta-analysis of binary data: which within study variance estimate to use?", Stat. Med. **20**, 1947–1956 (2001).

[73] R. D. Riley, "Multivariate meta-analysis: the effect of ignoring with-study correlation", J. R. Stat. Soc. A. Stat. **172**, 789–811 (2012).

[74] T. Amemiya, "Generalised least squares theory", in *Advanced econometrics* (Havard University Press, Cambridge, MA, USA, 1985) Chap. 6.

[75] A. Charnes, E. L. Frome, and P. L. Yu, "The equivalence of generalized least squares and maximum likelihood estimates in the exponential family", J. Am. Stat. Assoc. **71**, 169–171 (1976).

[76] G. R. Dolby, "Generalized least squares and maximum likelihood estimation of non-linear functional relationships", J. Roy. Stat. Soc. B. Met. **34**, 393–400 (1972).

[77] A. C. Aitken, "On least squares and linear combination of observations", P. Roy. Soc. Edinb. B. **55**, 42–48 (1936).

[78] R. DerSimonian and R. Kacker, "Random-effects model for meta-analysis of clinical trials: an update", Contemp. Clin. Trials. **28**, 105–114 (2007).

[79] J. Hartung and K. H. Makambi, "Reducing the number of unjustified significant results in meta-analysis", Commun. Stat. Simul. Comput. **32**, 1179–1190 (2003).

[80] K. Sidik and J. N. Jonkman, "Simple heterogeneity variance estimation for meta-analysis", J. R. Stat. Soc. C-Appl. **54**, 367–384 (2005).

[81] A. L. Rukhin, "Estimating heterogeneity variance in meta-analysis", J. R. Stat. Soc. B. **75**, 451–469 (2013).

[82] R. C. Paule and J. Mandel, "Consensus values and weighting factors", J. Res. Natl. Bur. Stand. **87**, 377–385 (1982).

[83] D. Langan, J. P. T. Higgins, D. Jackson, J. Bowden, A. A. Veroniki, E. Kontopantelis, W. Viechtbauer, and M. Simmonds, "A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses", Res. Synth. Meth. **10**, 83–98 (2019).

[84] A. A. Veroniki, D. Jackson, W. Viechtbauer, R. Bender, J. Bowden, G. Knapp, O. Kuss, J. P. T. Higgins, D. Langan, and G. Salanti, "Methods to estimate the between-study variance and its uncertainty in meta-analysis", Res. Synth. Meth. **7**, 55–79 (2016).

[85] M. Petropoulou and D. Mavridis, "A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study", Stat. Med. **36**, 4266–4280 (2017).

[86] D. Jackson, A. A. Veroniki, M. Law, A. C. Tricco, and R. Baker, "Paule-Mandel estimators for network meta-analysis with random inconsistency effects", Res. Synth. Meth. **8**, 416–434 (2017).

[87] D. Jackson, I. R. White, and T. G. Simon, "Extending Dersimonian and Laird's methodology to perform multivariate random effects meta-analyses", Stat. Med. **29**, 1282–1297 (2010).

[88] M. Law, D. Jackson, R. Turner, K. Rhodes, and W. Viechtbauer, "Two new methods to fit models for network meta-analysis with random inconsistency effects", BMC Med. Res. Methodol. **16**, 1–14 (2016).

[89] D. Jackson, M. Law, J. K. Barrett, R. Turner, J. P. T. Higgins, G. Salanti, and I. R. White, "Extending Dersimonian and Laird's methodology to perform network meta-analyses with random inconsistency effects", Stat. Med. **35**, 819–839 (2016).

[90] D. Jackson, I. R. White, and R. D. Riley, "Quantifying the impact of between-study heterogeneity in multivariate meta-analyses", Stat. Med. **31**, 3805–3820 (2012).

[91] R. N. Kacker, "Combining information from inter-laboratory evaluations using a random effects model", Metrologia **41**, 132–136 (2004).

[92] H. D. Patterson and R. Thompson, "Recovery of inter-block information when block sizes are unequal", Biometrika **58**, 545–554 (1971).

[93] W. G. Cochran, "The combination of estimates from different experiments", Biometrics **10**, 101–129 (1954).

[94] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems", J. Am. Stat. Assoc. **72**, 320–338 (1977).

[95] J. Wakefield, *Lecture notes BioStat 571: introduction and motivation, revision of estimation methods, linear mixed effects models, likelihood inference*, Online, Available from http://courses.washington.edu/b571/lectures.html. Accessed February 2022, 2009.

[96] D. A. Harville, "Bayesian inference for variance components using only error contrasts", Biometrika **61**, 383–385 (1974).

[97] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C. The art of scientific computing*, 2nd ed. (Cambridge University Press, Cambridge, UK, 1992).

[98] N. T. Longford, "A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects", Biometrika **74**, 817–827 (1987).

[99] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Roy. Stat. Soc. B. Met. **39**, 1–38 (1977).

[100] W. Viechtbauer, "Bias and efficiency of meta-analytic variance estimators in the random-effects model", J. Educ. Behav. Stat. **30**, 261–293 (2005).

[101] J. Hartung, "An alternative method for meta-analysis", Biometrical J. **41**, 901–916 (1999).

[102] J. Hartung and G. Knapp, "On tests of the overall treatment effect in meta-analysis with normally distributed responses", Stat. Med. **20**, 1771–1782 (2001).

[103] J. Hartung and G. Knapp, "A refined method for the meta-analysis of controlled clinical trials with binary outcome", Stat. Med. **20**, 3875–3889 (2001).

[104] K. Sidik and J. N. Jonkman, "A simple confidence interval for meta-analysis", Stat. Med. **21**, 3153–3159 (2002).

[105] D. Jackson, M. Law, G. Rücker, and G. Schwarzer, "The Hartung-Knapp modification for random-effects meta-analysis: a useful refinement but are there any residual concerns?", Stat. Med. **36**, 3923–3934 (2017).

[106] R. C. M. van Aert and D. Jackson, "A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression", Res. Synth. Meth. **10**, 515–527 (2019).

[107] G. Rücker, U. Krahn, J. König, O. Efthimiou, A. Davies, T. Papakonstantinou, and G. Schwarzer, *Netmeta: network meta-analysis using frequentist methods*, R package version 2.0-0. https://CRAN.R-project.org/package=netmeta, R Foundation for Statistical Computing (Vienna, Austria, 2021).

[108] I. R. White, "Network meta-analysis", Stata J. **15**, 951–985 (2015).

[109] N. Hawkins, D. A. Scott, B. S. Woods, and N. Thatcher, "No study left behind: a network meta-analysis in non–small-cell lung cancer demonstrating the importance of considering all relevant data", Value. Health. **12**, 996–1003 (2009).

[110] J. Hurley, "Forest plots or caterpillar plots? [letter to the editor]", J. Clin. Epidemiol. **121**, 110 (2020).

[111] G. Salanti, A. E. Ades, and J. P. A. Ioannidis, "Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial", J. Clin. Epidemiol. **64**, 163–171 (2011).

[112] L. Mbuagbaw, B. Rochwerg, R. Jaeschke, D. Heels-Andsell, W. Alhazzani, L. Thabane, and G. H. Guyatt, "Approaches to interpreting and choosing the best treatments in network meta-analyses", Syst. Rev. **6**, 79 (2017).

[113] I. R. White, "Multivariate random-effects meta-regression: updates to mvmeta", Stata J. **11**, 255–70 (2011).

[114] G. Rücker and G. Schwarzer, "Ranking treatments in frequentist network meta-analysis works without resampling methods", BMC Med. Res. Methodol. **15**, 58 (2015).

[115] G. Rücker and G. Schwarzer, "Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis", Stat. Med. **33**, 4353–4369 (2014).

[116] I. Gutman and W. Xiao, "Generalized inverse of the Laplacian matrix and some applications", Bull. Acad. Serbe. Sci. Cl. Sci. Math. Nat. Sci. Nat. **129**, 15–23 (2004).

[117] L. Lovász, *Random walks on graphs: a survey*, Online, YALE/DCS/TR-1029. Available from http://www.cs.yale.edu/publications/techreports/tr1029.pdf. Accessed March 2021, 1994.

[118] J. D. Noh and H. Rieger, "Random walks on complex networks", Phys. Rev. Lett. **92**, 118701 (2004).

[119] N. Masuda, M. A. Porter, and R. Lambiotte, "Random walks and diffusion on networks", Phys. Rep. **716-717**, 1–58 (2017).

[120] S. Kakutani, "Markov processes and the Dirichlet problem", Proc. Jap. Acad. **21**, 227–233 (1945).

[121] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Markov chains*, University Series in Higher Mathematics (Van Nostrand, New York, NY, UK, 1966).

[122] F. P. Kelly, *Reversibility and stochastic networks*, Probability and Statistics Series (Wiley, Chichester, UK, 1979).

[123] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, arXiv:math/0001057, 2000.

[124] J. König, U. Krahn, and H. Binder, "Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons", Stat. Med. **32**, 5414–5429 (2013).

[125] T. Papakonstantinou, A. Nikolakopoulou, G. Rücker, A. Chaimani, G. Schwarzer, M. Egger, and G. Salanti, "Estimating the contribution of studies in network meta-analysis: paths, flows and streams", F1000Res. **7**, 610 (2018).

[126] G. Salanti, C. Del Giovane, A. Chaimani, D. M. Caldwell, and J. P. T. Higgins, "Evaluating the quality of evidence from a network meta-analysis", PLOS ONE **9**, e99682 (2014).

[127] T. Papakonstantinou, A. Nikolakopoulou, M. Egger, and G. Salanti, "Meta-analysis as a system of springs", Res. Synth. Meth. **12**, 20–28 (2021).

[128] J. Bowden and C. Jackson, "Weighing evidence "steampunk" style via the meta-analyser", Am. Stat. **70**, 385–394 (2016).

[129] J. A. C. Sterne, D. Gavaghan, and M. Egger, "Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature", J. Clin. Epidemiol. **53**, 1119–1129 (2000).

[130] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", Comput. Networks ISDN **30**, 107–117 (1998).

[131] C. Daniłowicz and J. Baliński, "Document ranking based upon Markov chains", Inform. Process. Manag. **37**, 623–637 (2001).

[132] J. Blanchet, G. Gallego, and V. Goyal, "A Markov chain approximation to choice modeling", Oper. Res. **64**, 886–905 (2016).

[133] L. Trinquart, N. Attiche, A. Bafeta, R. Porcher, and P. Ravaud, "Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomised trials", Ann. Intern. Med. **164**, 666–673 (2016).

[134] A. A. Veroniki, S. E. Straus, A. Fyraridis, and A. C. Tricco, "The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes", J. Clin. Epidemiol. **76**, 193–199 (2016).

[135] A. A. Veroniki, S. E. Straus, G. Rücker, and A. C. Tricco, "Is providing uncertainty intervals in treatment ranking helpful in network meta-analysis?", J. Clin. Epidemiol. **100**, 122–129 (2018).

[136] C. H. Daly, B. Neupane, J. Beyene, L. Thabane, S. E. Straus, and J. S. Hamid, "Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa", BMJ Open **9**, e024625 (2019).

[137] A. Chaimani, R. Porcher, É. Sbidian, and D. Mavridis, "A Markov chain approach for ranking treatments in network meta-analysis", Stat. Med. **40**, 451–464 (2021).

[138] V. Chiocchia, A. Nikolakopoulou, T. Papakonstantinou, M. Egger, and G. Salanti, "Agreement between ranking metrics in network meta-analysis: an empirical study", BMJ Open **10**, e037744 (2020).

[139] D. Mavridis, R. Porcher, A. Nikolakopoulou, and G. Salanti, "Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes", Biometrical J. **62**, 375–385 (2020).

[140] A. Nikolakopoulou, D. Mavridis, V. Chiocchia, T. Papakonstantinou, T. A. Furukawa, and G. Salanti, "Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric", Res. Synth. Meth. **12**, 161–175 (2021).

[141] B. Hutton, G. Salanti, D. M. Caldwell, A. Chaimani, C. H. Schmid, C. Cameron, J. P. A. Ioannidis, S. E. Straus, K. Thorlund, J. P. Jansen, C. Mulrow, F. Catalá-López, P. C. Gøtzsche, K. Dickersin, I. Boutron, D. G. Altman, and D. Moher, "The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations", Ann. Intern. Med. **162**, 777–784 (2015).

[142] A. L. Davies and T. Galla, "Degree irregularity and rank probability bias in network meta-analysis", Res. Synth. Meth. **12**, 316–332 (2021).

[143] F. S. Tonin, H. H. Borba, A. M. Mendes, A. Wiens, F. Fernandez-Llimos, and R. Pontarolo, "Description of network meta-analysis geometry: a metrics design study", PLOS ONE **14**, e0212650 (2019).

[144] G. Salanti, F. K. Kavvoura, and J. P. A. Ioannidis, "Exploring the geometry of treatment networks", Ann. Intern. Med. **148**, 544–553 (2008).

[145] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Methods and Principles in Medicinal Chemistry (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2000).

[146] E. T. Jaynes, "Information theory and statistical mechanics", Phys. Rev. **106**, 620–630 (1957).

[147] E. T. Jaynes, "Information theory and statistical mechanics II", Phys. Rev. **108**, 171–190 (1957).

[148] D. Pisinger, "Where are the hard knapsack problems?", Comput. Oper. Res. **32**, 2271–2284 (2005).

[149] R. Monasson and R. Zecchina, "Statistical mechanics of the random k-satisfiability model", Phys. Rev. E **56**, 1357–1370 (1997).

[150] M. Mézard, G. Parisi, and R. Zecchina, "Analytic and algorithmic solution of random satisfiability problems", Science **297**, 812–815 (2002).

[151] A. K. Hartmann and M. Weight, *Phase transitions in combinatorial optimisation problems* (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2005).

[152] A. Nikolakopoulou, A. Chaimani, A. A. Veroniki, H. S. Vasiliadis, C. H. Schmid, and G. Salanti, "Characteristics of networks of interventions: a description of a database of 186 published networks", PLOS ONE **9**, e86754 (2014).

[153] T. P. Minka, "Expectation propagation for approximate Bayesian inference", in Proceedings of the 17th conference in uncertainty in artificial intelligence, edited by J. Breese and D. Koller (2001), pp. 362–369.

[154] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes", in Proceedings of the 15th conference in uncertainty in artificial intelligence, edited by K. B. Laskey and H. Prade (1999), pp. 21–30.

[155] R. Sauter and L. Held, "Network meta-analysis with integrated nested Laplace approximations", Biometrical J. **57**, 1038–1050 (2015).

[156] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)", J. Roy. Stat. Soc. B. **71**, 319–392 (2009).

[157] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: a practical guide to using machine learning tools in research synthesis", Syst. Rev. **8**, 163 (2019).

[158] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace, "Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide", Res. Synth. Meth. **9**, 602–614 (2018).

[159] T. Lange, G. Schwarzer, T. Datzmann, and H. Binder, "Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies", Res. Synth. Meth. **12**, 506–515 (2021).

[160] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods", Stat. Med. **38**, 2074–2102 (2019).

[161] T. Kibret, D. Richer, and J. Bayene, "Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study", Clin. Epidemiol. **6**, 451–460 (2014).

[162] S. E. Seide, K. Jensen, and M. Kieser, "A comparison of Bayesian and frequentist methods in random-effects network meta-analysis of binary data", Res. Synth. Meth. **11**, 363–378 (2020).

[163] R. D. Baker and D. Jackson, "Meta-analysis inside and outside particle physics: two traditions that should converge?", Res. Synth. Meth. **4**, 109–124 (2013).

[164] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting", BMJ **340**, c221 (2010).

[165] G. Rücker, M. Petropoulou, and G. Schwarzer, "Network meta-analysis of multicomponent interventions", Biometrical J. **62**, 808–821 (2020).

[166] R. D. Riley, D. Jackson, G. Salanti, D. L. Burke, M. Price, J. Kirkham, and I. R. White, "Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples", BMJ **358**, j3932 (2017).

[167] M. J. Bradburn, J. J. Deeks, J. A. Berlin, and A. R. Localio, "Much ado about nothing: a comparison of the performance of meta analytical methods with rare events", Stat. Med. **26**, 53–77 (2007).

[168] A. J. Sweeting, P. C. Lambert, and A. J. Sutton, "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data", Stat. Med. **23**, 1351–1375 (2004).

[169] D. C. Hoaglin, "We know less than we should about methods of meta-analysis", Res. Synth. Meth. **6**, 287–289 (2015).

[170] D. C. Hoaglin, "Misunderstandings about Q and 'Cochran's Q test' in meta-analysis", Stat. Med. **35**, 485–495 (2016).

[171] S. R. Searle and M. H. J. Gruber, *Linear models*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, USA, 2016).

[172] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation", IEEE. T. Knowl. Data. En. **19**, 455–369 (2007).

Chapter 3

Degree irregularity and rank probability bias in network meta-analysis

Preface

The contents of this chapter constitute a manuscript published by Research Synthesis Methods[1]. What was originally published as online Supplementary Material has been split into an Appendix and a Supplementary Material chapter (Chapter 8). The former appears immediately after the main text. The latter contains a large number of figures and is given at the end of the document so as not to interrupt the flow of the text. In the following, we refer to Chapter 8 as the 'Supplementary Material'. The manuscript was authored by Annabel L Davies[2] and Tobias Galla[2,3].

ALD designed the study, contributed to discussions guiding the work, performed all of the simulations, wrote the first draft of the manuscript, produced all of the figures and edited the manuscript. TG designed the study, contributed to discussions guiding the work and edited the manuscript.

[1]A. L. Davies and T. Galla, "Degree irregularity and rank probability bias in network meta-analysis", Res. Synth. Meth. **12**, 316-332 (2021). 10.1002/jrsm.1454

[2]Theoretical Physics, School of Physics and Astronomy, The University of Manchester, Manchester, M13 9PL, United Kingdom.

[3]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

Abstract

Network meta-analysis (NMA) is a statistical technique for the comparison of treatment options. Outcomes of Bayesian NMA include estimates of treatment effects, and the probabilities that each treatment is ranked best, second best and so on. How exactly network topology affects the accuracy and precision of these outcomes is not fully understood. Here we carry out a simulation study and find that disparity in the number of trials involving different treatments leads to a systematic bias in estimated rank probabilities. This bias is associated with an increased variation in the precision of treatment effect estimates. Using ideas from the theory of complex networks, we define a measure of 'degree irregularity' to quantify asymmetry in the number of studies involving each treatment. Our simulations indicate that more regular networks have more precise treatment effect estimates and smaller bias of rank probabilities. Conversely, these topological effects are not observed for the accuracy of treatment effect estimates. This reinforces the importance of taking into account multiple measures, rather than making decisions based on a single metric. We also find that degree regularity is a better indicator for the accuracy and precision of parameter estimates in NMA than both the total number of studies in a network and the disparity in the number of trials per comparison. These results have implications for planning future trials. We demonstrate that choosing trials which reduce the network's irregularity can improve the precision and accuracy of parameter estimates from NMA.

3.1 Introduction

Meta-analysis is an important statistical technique used to combine the results of multiple randomised controlled trials. Often, individual trials have small sample sizes and involve subjects taken from a reduced population. Because of this, it is desirable to systematically integrate results from different trials that address the same clinical question. Over the last four decades meta-analysis has therefore become invaluable for the comparison of treatment options [1].

Conventional meta-analysis focuses on pairwise comparisons of treatments. More recently however, network meta-analysis (NMA) has emerged as a technique for making inferences about multiple competing treatments. NMA allows one to combine data from multiple trials even when different trials test different sets of treatment options. The term 'network meta-analysis' derives from a graphical representation of the treatments and trials. The nodes of the network graph are the different treatment options and the connecting edges represent comparisons made between the treatments in the trials. NMA combines both direct and indirect evidence for the assessment of treatments. This makes it possible to compare treatments that have not been tested together in any trial [1–4].

NMA has undergone substantial development over recent years. At the same time, it is also recognised that further research is required to fully understand its limitations and to improve the method [5, 6]. In this context, simulation studies are frequently used to evaluate the performance of NMA and the factors affecting its accuracy [7]. This approach involves setting up a model (e.g. a fixed-effect or random-effect model [2, 8]) with parameters whose numerical values can be fixed at the beginning. The model is used to produce synthetic trial data [9]. Estimates of the model parameters are then obtained by feeding this data into the NMA method. These estimates can then be compared against the known model parameters.

A key strength of this approach is the ability to systematically vary the parameters of the model. For example, different relative treatment effects can be explored, or the structure of the network of treatments and trials can be changed. This allows one to systematically investigate the performance of NMA in a range of different scenarios, and to determine the nature of any inaccuracies or biases. To do this, however, Bayesian NMAs must be carried out for many realisations of the synthetic trial data. The overall computational effort can be considerable because the NMA method relies on extensive Markov Chain Monte Carlo sampling [10, 11].

The primary outcomes of an NMA are estimates of relative treatment effects, and the corresponding credible intervals. This allows one to rank treatments based on these relative effects, providing a convenient summary for clinical decision making. However, simply ranking treatments as best, second best and so on can be misleading as it does not take into account the level of overlap between credible intervals [12].

As a consequence, a number of other metrics have been developed to compare treatments. One such metric focuses on so-called 'rank probabilities'. These quantify the degree of certainty with which each treatment is believed to be the most effective, second most effective, etc., based on the available trial data. This is a natural object in a Bayesian setting [13], but can also be achieved in frequentist NMA using resampling [14]. Results are often reported in terms of so-called SUCRA values ('surface under the cumulative ranking curve'). These values condense rank probabilities into a numerical summary [13], and reflect both the magnitude and the uncertainty of treatment effect estimates [6, 15, 16]. SUCRA endpoints can also be obtained from frequentist approaches without the need for resampling [16].

Ranking methods have attracted considerable interest in recent years [15, 17–21]. It is generally recognised that the accuracy of ranking statistics and the treatment effect estimates are likely to be affected by the topology of the network of treatments and trials [5, 18, 22–24]. PRISMA guidelines ('Preferred Reporting Items for Systematic Reviews and Meta-Analyses') therefore recommend that authors provide graphical and qualitative descriptions of network geometry [6].

A previous simulation study found that the probability of being ranked first is overestimated for the treatment that is tested in the fewest studies in a given network, and underestimated for the treatment included in the most studies [25]. It has been suggested that this is due to differences in the precision of treatment effect estimates [16]. Overall it is generally accepted that reporting only the probability of being best can lead to erroneous conclusions [15, 18, 26]. Indeed, the current advice from the PRISMA guidelines is to report the probability that each treatment has each rank [6].

In practice however, the most common ranking statistic in NMAs continues to be the probability that each treatment is ranked best [27]. Previous research on the utility of rank probabilities has also focused almost exclusively on the probability of being ranked best. As a result there is very limited evidence on the validity of reporting the full set of rank probabilities or the SUCRA values. Furthermore, due to a lack of appropriate data-generating models and the high computing power required to carry out Bayesian NMAs, simulation studies have been limited to fixed-effects models or networks of two-arm trials only [25, 28]. Some progress been made in relating characteristics of network geometry to the outcome of NMA [22, 23, 29]. However, it

is largely unexplored how exactly these metrics relate to the performance of ranking statistics and treatment effect estimates [5].

The purpose of our work is to study how the structure of the network affects the probability that each treatment is ranked first, second and so on. In particular we go beyond the probability of being ranked best. We also investigate the mechanisms by which the network affects rank probabilities. Building on recent advances in data-generating methods [30], our simulation studies include random-effects models and networks of multi-arm trials. In order to characterise network geometry we introduce a measure of asymmetry in the number of studies per treatment which we call 'degree irregularity'. Similar quantities have been used to characterise networks in other fields [31–33]. The network is said to be regular if all treatments are tested in the same number of studies, and it becomes increasingly more irregular the more this number varies across treatments. Through simulations of multiple network geometries we investigate how this metric affects the precision and accuracy of the treatment effect estimates, and the quality of rank probability estimates and SUCRA values. These results provide a simple method for the identification of additional trials as potential candidates to complement an existing network of evidence.

The remainder of this paper is set out as follows: In Section 3.2 we present the relevant background information and methods. We begin by outlining the random-effects model for a network of multi-arm trials and the Bayesian approach to NMA. Following this we define the key outcomes of NMA and the relevant ranking statistics. The design of the simulation and networks are described along with details of the data-generating models. We also introduce treatment-focused and network-level quantities that allow us to compare within and between networks how well the NMA process retrieves model parameters used to generate synthetic input data. Section 3.3 contains our main results. First, we present within-network and between-network comparisons for networks with equally effective treatments and two-arm trials. We then test how well these results generalise to scenarios in which the true treatment effects vary across treatments, and we study networks involving multi-arm trials. We also compare the results from three different data-generating models. In Section 3.4 we summarise and discuss our main findings. We provide an example that demonstrates how our results can be used to inform the choice of future trials.

3.2 Methods

3.2.1 General setup: network of trials

We consider a collection of N treatments, which we label $a = T_1, T_2, \ldots, T_N$. The network contains M trials, denoted $i = 1, \ldots, M$. Each trial compares a subset of treatments, $A_i \subset \{T_1, \ldots, T_N\}$; $m_i = |A_i|$ is the number of treatments in trial i. We use the notation $t_{i,\ell}$ to label the treatments compared in trial i, where $\ell = 1, \ldots, m_i$. Each $t_{i,\ell}$ is therefore a treatment from the set $\{T_1, \ldots, T_N\}$.

As an example, consider a network of smoking cessation data reported by Hasselblad [34]. Four treatments are compared: T_1 = no contact (control), T_2 = self-help, T_3 = individual counselling and T_4 = group counselling. The network is shown in Figure 3.1 and consists of 24 trials. Trial $i = 2$ in Hasselblad [34] compares $m_2 = 3$ treatments such that $t_{2,1} = T_1$ (no contact), $t_{2,2} = T_3$ (individual counselling) and $t_{2,3} = T_4$ (group counselling). Trial $i = 6$, on the other hand, compares $m_6 = 2$ treatments; $t_{6,1} = T_2$ (self-help) and $t_{6,2} = T_3$ (individual counselling). In Figure 3.1(a), these two trials are highlighted by dashed and dotted lines respectively. Figure 3.1(b) depicts the whole network as a weighted graph [35] where for each pair of treatments, the thickness of the link is proportional to the number of trials comparing these two treatments. The diameter of each node is proportional to the number of participants that have received the treatment represented by the node.

The treatments in trial i are referred to as the arms of the trial. For a given trial i, the treatment in arm ℓ is administered to $n_{i,\ell}$ patients. We assume a binary outcome, i.e. the application of the treatment to a particular patient either produces an 'event', or it does not. The number of resulting events, $r_{i,\ell}$, is then recorded for each trial and arm. This means that trial i is defined by the treatments it compares, $t_{i,1}, \ldots, t_{i,m_i}$, and by the number of patients in each arm, $n_{i,\ell}$. The trial reports dichotomous data of the form $r_{i,1}, ..., , r_{i,m_i}$.

The model assumes that the application of the treatment in arm ℓ of trial i generates events with probability $p_{i,\ell}$ independently for each of the $n_{i,\ell}$ patients at the end of this trial arm [36]. As a consequence of this setup, each $r_{i,\ell}$ is a binomial random variable,

$$r_{i,\ell} \sim \text{Bin}(n_{i,\ell}, p_{i,\ell}), \tag{3.1}$$

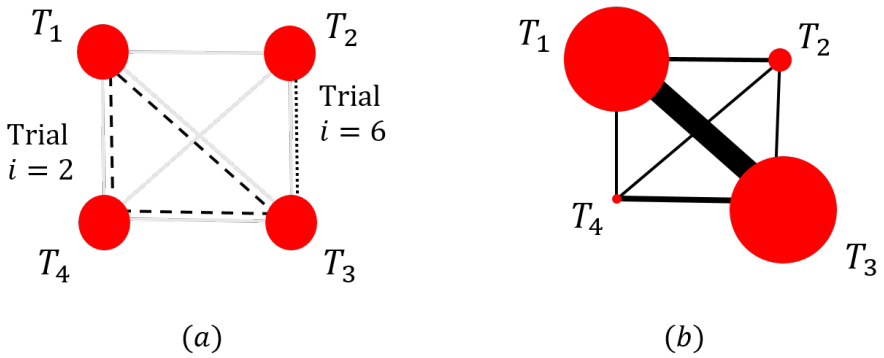


Figure 3.1: Graphical representation of the network of treatments and trials for smoking cessation [34].The four treatments are: T_1 = no contact (control), T_2 = self-help, T_3 = individual counselling and T_4 = group counselling. In panel (a), trials $i = 2$ (a 3-arm study) and $i = 6$ (a 2-arm study) are highlighted by dashed and dotted lines respectively. The thickness of each edge in panel (b) is proportional to the number of studies that make that comparison, and the diameter each node is proportional to the number of participants who received that treatment.

for $i = 1, \ldots, M$ and $\ell = 1, \ldots m_i$.

We use a random-effects model, i.e. $p_{i,\ell}$ may be different from $p_{i',\ell'}$ in different trials ($i \neq i'$), even if $t_{i,\ell} = t_{i',\ell'}$. That is to say, the effectiveness of any fixed treatment $a \in \{T_1, \ldots, T_N\}$ may be different in different trials. Following the generalised linear model framework, the probabilities $p_{i,\ell}$ are described on the logit scale [37].

Our analysis focuses on relative treatment effects. To this end, we refer to treatment $t_{i,\ell=1}$ as the 'baseline' treatment of trial i. We write $\mu_i = \text{logit } p_{i,1}$ for the absolute outcome of this trial-specific baseline. For $\ell \neq 1$ we then define the relative treatment effect $\delta_{i,1\ell}$,

$$\delta_{i,1\ell} \equiv \text{logit}(p_{i,\ell}) - \mu_i, \tag{3.2}$$

where $\text{logit}(p) = \ln p - \ln(1-p)$ for $0 < p < 1$. The trial-specific baseline treatment outcome, μ_i, is the log odds of the event probability in arm $\ell = 1$ of trial i, while the relative treatment effect is the log odds ratio of treatment $\ell \neq 1$ compared to the trial-specific baseline.

3.2.2 Random-effects model

The random effects model we use assumes the exchangeability of relative treatment effects [38–40]. This indicates that the relative effect of two treatments a and b is drawn from the same distribution for any trial involving these two treatments.

We assume that the relative treatment effects for a given trial i are drawn from a multivariate normal distribution

$$\begin{pmatrix} \delta_{i,12} \\ \vdots \\ \delta_{i,1m_i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} d_{t_{i,1}t_{i,2}} \\ \vdots \\ d_{t_{i,1}t_{i,m_i}} \end{pmatrix}, \boldsymbol{\Sigma_i} \right). \tag{3.3}$$

This means that the relative effect $\delta_{i,1\ell}$ of the ℓ-th treatment in trial i (compared to the baseline treatment of trial i) is drawn from a Gaussian distribution with mean $d_{t_{i,1},t_{i,\ell}}$. The latter quantity is the mean effect of treatment $t_{i,\ell}$ relative to the baseline treatment $t_{i,1}$ of trial i. That is to say, it is the average relative treatment effect one would see in a large sample of trials comparing these two treatments. We assume that these unknown mean relative treatment effects fulfil the consistency relations

$$d_{ab} = d_{ac} - d_{bc}. \tag{3.4}$$

The $(m_i-1)\times(m_i-1)$ covariance matrix $\boldsymbol{\Sigma_i}$ in Equation (3.3) describes the between-trial variance of the relative treatment effects, and their correlations. Following References [4, 41, 42], we will assume that its diagonal elements are all identical. We write τ^2 for their common value. This is the variance of each $\delta_{i,1\ell}$. We will further assume that the covariance between any two treatment effects is $\tau^2/2$ (these are the off-diagonal elements of $\boldsymbol{\Sigma_i}$). This ensures that the relative effect $\delta_{i,1\ell} - \delta_{i,1\ell'}$ between *any two* treatments $\ell \neq \ell'$ in trial i has variance τ^2.

The aim of network meta-analysis is to estimate the mean treatment effects d_{ab} for all pairs $a \neq b$, and the heterogeneity parameter, τ. Given the consistency assumption (3.4), not all d_{ab} are independent. As a consequence, we can use treatment $a = T_1$ as the overall global baseline treatment, and it is sufficient to estimate d_{T_1a} for $a = T_2, \ldots, T_N$ [43].

3.2.3 Bayesian network meta-analysis

We write $\boldsymbol{d} = (d_{T_1T_2}, d_{T_1T_3}, \ldots, d_{T_1T_N})$ for the vector of mean treatment effects relative to the global baseline. The vector $\boldsymbol{n}$ contains the numbers of patients in all arms in the network and $\boldsymbol{r}$ contains the trial outcomes. Bayesian NMA aims to construct posterior distributions for the model parameters, $(\boldsymbol{d}, \tau)$, conditional on the data, $(\boldsymbol{r}, \boldsymbol{n})$. This is

achieved using appropriate likelihood functions and prior distributions [4, 44]. We use non-informative prior distributions for the model parameters. Specifically, we assume independent univariate Gaussian distributions $\mathcal{N}(0, 10^8)$ for each of the parameters μ_i and d_{ab}. The prior for τ is assumed to be a uniform distribution over the interval from 0 to 5 [2, 45].

For this setup, the posterior distributions of the model parameters can usually not be obtained analytically. We therefore rely on MCMC methods, specifically the Metropolis-in-Gibbs algorithm [10, 11, 46, 47]. Following Kibret et al (2014) [25], we used a burn-in of 5×10^3 and a thinning factor of 10. Samples were drawn from the posterior distributions for 2×10^4 iterations after burn-in.

3.2.4 Reporting NMA outcomes

The primary outcomes from an NMA are the final estimates of the model parameters and their uncertainty (the latter is usually indicated by a 95% credible interval). In addition, Bayesian NMA allows for the calculation of rank probabilities $\{P_a(r)\}$. The quantity $P_a(r)$ is the probability that treatment a is ranked r-th. At each MCMC iteration the treatments are ranked from best (rank $r = 1$) to worst (rank $r = N$) based on the values of $d_{T_1 a}$ sampled at that iteration. The rank probabilities are then estimated from the proportion of times each treatment received each rank.

Treatment effect estimates and ranking probabilities become more difficult to interpret as the number of treatments in the network increases [15, 18]. In order to simplify this information, Salanti et al (2011) introduced a numerical summary, the so-called 'surface under the cumulative ranking' curve (SUCRA) [13]. The value of SUCRA for treatment a is defined as

$$\text{SUCRA}_a = \frac{1}{N-1} \sum_{r=1}^{N-1} F_a(r), \tag{3.5}$$

where $F_a(r)$ is the probability that treatment a has rank r or better [13, 16],

$$F_a(r) = \sum_{s=1}^{r} P_a(s). \tag{3.6}$$

We write the mean or expected rank as

$$\mathbb{E}(r)_a = \sum_{r=1}^{N} r P_a(r). \tag{3.7}$$

It is straightforward to see that [16]

$$\text{SUCRA}_a = \frac{1}{N-1}(N - \mathbb{E}(r)_a). \tag{3.8}$$

In this notation SUCRA takes values between zero and one.

3.2.5 Network design

The simulations reported in the main paper are restricted to networks with $N = 4$ treatments. This is mostly to keep the search space of possible graphs manageable, and in order to be efficient in the identification of the key factors determining the accuracy and precision of NMA parameter estimates. Simulations reported in Section 8.9.1 of the Supplementary Material demonstrate that our principal results continue to be valid for larger networks.

Figure 3.2 shows the five network geometries we have used: (a) star, (b) loop, (c) complete loop, (d) tadpole, and (e) ladder. These geometries were chosen as they are commonly observed in real-life network meta-analyses; combinations of these have been previously studied in [23, 25, 28, 43].

Within the constraints of these geometries, the number of studies per comparison was varied (i.e. the number of trials involving a particular pair of treatments). To describe the specific geometry of a network we use the vector of the number of studies per comparison, $\boldsymbol{K} = (K_{T_1T_2}, K_{T_1T_3}, K_{T_1T_4}, K_{T_2T_3}, K_{T_2T_4}, K_{T_3T_4})$, where $K_{ab} = K_{ba}$ is the number of studies that compare treatments a and b. The entries of $\boldsymbol{K}$ define the strengths (or weights) of the edges in the network of treatments.

We note, however, that the full setup of the treatment–trial network is not fully specified by $\boldsymbol{K}$ alone. This is because the same number of comparisons per pair of treatments can be achieved by different combinations of two-arm and multi-arm trials.

From $\boldsymbol{K}$ we can obtain the number of studies involving treatment a,

$$k_a = \sum_{b \neq a} K_{ab}. \tag{3.9}$$

In the theory of networks this quantity is referred to as the 'weighted degree' of node a [35]. We will occasionally use slightly more casual language, and refer to k_a as the 'number of studies per treatment'. We also define the average number of studies that a

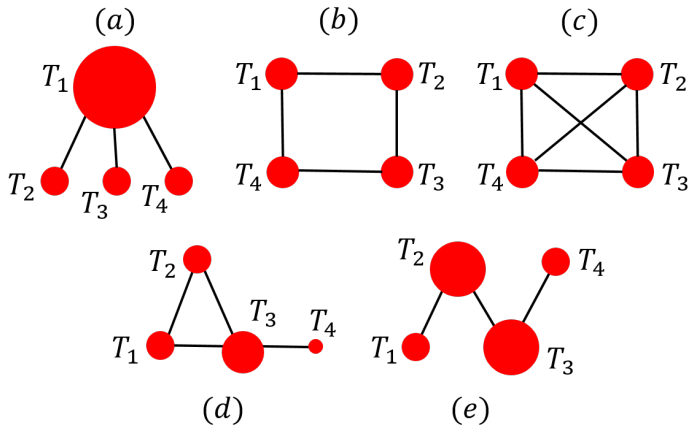


Figure 3.2: Network diagrams of the five network geometries considered in this study: (a) star (b) loop (c) complete loop (d) tadpole (e) ladder.

treatment is involved in (the 'mean degree'),

$$\bar{k} = \frac{1}{N}\sum_a k_a. \tag{3.10}$$

The overbar in our notation thus indicates an average over the N nodes of the network.

In network theory, a graph is said to be 'regular' if all nodes have the same degree [35]. With this in mind, we introduce a measure of 'degree irregularity' of the network,

$$h^2 = \frac{1}{N}\sum_a (k_a - \bar{k})^2. \tag{3.11}$$

This quantifies the variation in the number of studies per treatment. In particular, $h^2 = 0$ when all nodes are involved in the same number of trials ($k_a = \bar{k}$ for all a). When we make comparisons between networks we use the normalised network irregularity, $h^2/\bar{k}^2$. This is akin to the concept of a 'topological index' or 'molecular descriptor' in chemical graph theory [48], and we note that similar measures of graph irregularity have also been used for example in the social sciences [31].

We note that there is a direct mapping between $h^2/\bar{k}^2$ and the so-called 'probability of inter-specific encounter index' (PIE). This index is a measure of ecological diversity, and was introduced to network meta-analysis by Salanti et al (2008) [22, 23]. Further details can be found in Appendix Section 3.5.

Some of the key quantities we use in our analysis are summarised in Table 3.1.

Table 3.1: Summary of key quantities used in our analysis, and notation for different types of averages.

Variable	Definition
N	Total number of treatments in the network
M	Total number of studies in the network
d_{ab}	True mean relative treatment effect between treatments a and b
τ	Heterogeneity parameter
k_a	Number of studies involving treatment a
$\bar{k}$	Mean degree of the network (mean number of studies a treatment is involved in)
$h^2/\bar{k}^2$	Normalised degree irregularity
$\mathrm{SD}(d)_a$	Treatment-specific standard deviation of the treatment effect estimate
$\Delta P_a(r)$	Bias of the rank probability estimate
$\mathrm{SD}_{\mathrm{tot}}$	Total standard deviation of treatment effect estimates in the network
$\lvert\Delta P\rvert_{\mathrm{tot}}$	Total rank probability bias in the network
$\lvert\Delta\mathrm{SUCRA}\rvert_{\mathrm{tot}}$	Total SUCRA bias in the network
$\mathbb{E}(\cdots)_a$	Average over the distribution of ranks for treatment option a
$\langle\cdots\rangle$	Average over realisations of synthetic data
$\overline{\cdots}$	Average over nodes of the network

3.2.6 Simulation method

In our simulations we generate dichotomous trial data for a specified network geometry and for known model parameters $(\boldsymbol{d}, \tau)$. An NMA is performed for multiple independent realisations of simulated data, and the resulting estimates of the model parameters are recorded for each realisation. More specifically, we used the following numerical protocol:

(1) Define the fixed parameters of the network such as the total number of studies, M, the vector of number of studies per comparison, $\boldsymbol{K}$, the number of participants in each arm, $n_{i,\ell}$, and the true model parameter values $(\boldsymbol{d}, \tau)$.

(2) Generate and analyse independent realisations $\nu = 1, 2, ..., \Omega$ of synthetic trial outcomes. Specifically, for each ν:

   (a) For all trials i, randomly sample the $\{\delta_{i,1\ell}\}$, $\ell = 2, \ldots, m_i$, from the multivariate normal distribution in Equation (3.3).

   (b) Using the $\{\delta_{i,1\ell}\}$ and one of the three data-generating models (see Section 3.2.7), construct the probabilities $p_{i,\ell}$, $\ell = 1, \ldots, m_i$, for all trials i in the network.

We note that the values for the $p_{i,\ell}$ vary from realisation to realisation in this process.

(c) For each trial arm, generate random event data, $r_{i,\ell}$, from the binomial distribution in Equation (3.1).

(d) Use the vector of events, $\boldsymbol{r}$, and vector of participants, $\boldsymbol{n}$, to carry out a Bayesian NMA.

(e) Determine the treatment effects with respect to the baseline T_1 and use the consistency relation in Equation (3.4) to output the estimated model parameters, $\hat{d}_{ab}^{(\nu)}$, for all $a, b \in \{T_1, \ldots, T_N\}$. Also output the estimated heterogeneity parameter, $\hat{\tau}^{(\nu)}$, and the bias of rank probabilities,

$$\Delta P_a^{(\nu)}(r) = \hat{P}_a^{(\nu)}(r) - P_a^{\mathrm{bl}}(r). \tag{3.12}$$

In this equation $\hat{P}_a^{(\nu)}(r)$ is the probability that treatment a has rank r in the NMA of realisation ν. The quantity $P_a^{\mathrm{bl}}(r)$ is the 'baseline' probability for treatment a to be ranked r-th, obtained directly from the true relative treatment effects used to generate the synthetic data. Further details can be found below after Equation (3.13).

(3) Calculate the mean and standard deviations of the estimated model parameters over realisations, for example

$$\begin{aligned} \langle \hat{d}_{ab} \rangle &= \frac{1}{\Omega} \sum_{\nu=1}^{\Omega} \hat{d}_{ab}^{(\nu)}, \\ \mathrm{SD}(\hat{d}_{ab}) &= \sqrt{\frac{1}{\Omega - 1} \sum_{\nu=1}^{\Omega} (\hat{d}_{ab}^{(\nu)} - \langle \hat{d}_{ab} \rangle)^2}, \end{aligned} \tag{3.13}$$

with similar definitions for $\langle \hat{\tau} \rangle$, $\mathrm{SD}(\hat{\tau})$, $\langle \Delta P_a(r) \rangle$ and $\mathrm{SD}(\Delta P_a(r))$. In these expressions, angular brackets indicate an average over independent realisations of the simulated data.

We stress that a suitable baseline comparator is required to compute bias on rank probabilities. We use the 'baseline' probabilities $P_a^{\mathrm{bl}}(r)$. These are obtained from ranking the treatments based on the true relative treatment effects. Equivalently they are the rank probabilities one would obtain from an NMA if it were able to estimate the relative treatment effects used to generate the synthetic data with perfect accuracy and no uncertainty.

3.2.7 Data generation for simulation studies

The relative treatment effects $\delta_{i,1\ell}$ $(\ell = 2, \ldots, m_i)$ in any one trial i do not uniquely define the absolute outcomes $p_{i,\ell}$. This is because the outcome of the trial-specific baseline, $p_{i,1}$, is not determined by the $\{\delta_{i,1\ell}\}$. Equation (3.2) can be re-arranged to give

$$p_{i,\ell} = p_{i,\ell}[p_{i,1}, \delta_{i,1\ell}] = \frac{p_{i,1}\exp(\delta_{i,1\ell})}{1 + p_{i,1}\left(\exp(\delta_{i,1\ell}) - 1\right)}, \tag{3.14}$$

so that $p_{i,1}$ together with the $\{\delta_{i,1\ell}\}$ $(\ell = 2, \ldots, m_i)$ specifies all absolute outcomes in trial i.

To fully define step (2)(b) in the above algorithm it is therefore sufficient to specify the construction of $p_{i,1}$. In the context of the random-effects model and to allow for the inclusion of multi-arm trials, we use three data-generating models (DGM) based on those presented by Seide et al (2019) [30].

The first DGM, which we will call 'Euclidean', chooses the outcome for the baseline treatment to be the value that minimises the Euclidean distance of the vector $(p_{i,1}, \ldots, p_{i,m_i})$ from the vector $(1/2, \ldots, 1/2)$, i.e.

$$p_{i,1} = \min_q \left[\left(q - \frac{1}{2}\right)^2 + \sum_{\ell=2}^{m_i} \left(p_{i,\ell}\left[q, \delta_{i,1\ell}\right] - \frac{1}{2}\right)^2 \right], \tag{3.15}$$

where $p_{i,\ell}[\cdot,\cdot]$ is the expression given in Equation (3.14). This is referred to as 'DGM "Fixed" Modified' in Seide et al (2019) [30]. The other two methods are variations of the DGM "Fixed" in Seide et al (2019) [30] which we will refer to as 'Uniform' and 'Normal' respectively. The former samples $p_{i,1}$ from a uniform distribution between zero and one, whilst the latter samples it from a normal distribution $\mathcal{N}(0.5, 0.04)$, with variance $\sigma^2 = 0.04$, truncated at zero at the lower end, and at one at the upper end. To ensure our results were not due to the data-generating model, all simulations were performed using each method and the results compared.

The data generation, simulation algorithm and MCMC method used to carry out the Bayesian NMA were performed using a tailor-made `C++` code, an example of which is provided here: https://github.com/AnnieDavies/Supplementary-Material-Davies-Galla-2020.

3.2.8 Quantities indicating and characterising the accuracy and precision of estimates from NMA

In this section we introduce quantities that measure how accurate and precise the parameter estimates from the NMA are. We begin with indicators for individual treatments in the graph (as opposed to aggregate measures characterising a graph as a whole).

As a first step, we define the mean bias of the relative treatment effect,

$$\langle \Delta d_{ab} \rangle = \langle \hat{d}_{ab} \rangle - d_{ab}, \tag{3.16}$$

for each pair of treatments a and b. For each fixed treatment a we can then define the mean bias

$$\langle \Delta d \rangle_a = \frac{1}{N-1} \sum_{b \neq a} \langle \Delta d_{ab} \rangle. \tag{3.17}$$

This quantity indicates a systematic bias in the relative effect of treatment a compared to other treatments in the graph. If $\langle \Delta d \rangle_a < 0$ then the relative effect of treatment a is underestimated on average, and if $\langle \Delta d \rangle_a > 0$ it is overestimated. Similarly, the standard deviation of the treatment effect for treatment a in comparison to the other treatments in the graph is defined as

$$\mathrm{SD}(d)_a = \frac{1}{N-1} \sum_{b \neq a} \mathrm{SD}(\hat{d}_{ab}). \tag{3.18}$$

Bias of SUCRA_a values and bias of probability ranks, $\langle \Delta P_a(r) \rangle$, are specific to individual treatments by construction. The former can be written in terms of the latter,

$$\langle \Delta \mathrm{SUCRA}_a \rangle = -\frac{\sum_r r \langle \Delta P_a(r) \rangle}{N-1}. \tag{3.19}$$

We stress again that a suitable baseline comparator is required to define bias on rank probabilities.

Next we define network-level indicators allowing comparisons of the accuracy and precision of parameter estimates between networks. We introduce the total magnitude of the bias of rank probability,

$$|\Delta P|_{\mathrm{tot}} = \sum_a \sum_r |\langle \Delta P_a(r) \rangle|, \tag{3.20}$$

and the total magnitude of the bias of SUCRA,

$$|\Delta\text{SUCRA}|_{\text{tot}} = \sum_a |\langle \Delta\text{SUCRA}_a \rangle|. \tag{3.21}$$

To be able to compare numerical values for these two quantities with each other, we express these indicators as proportions of the maximum values they can take, see Appendix Section 3.6 for further details.

Finally, we introduce the total standard deviation and total bias of treatment effects,

$$\text{SD}_{\text{tot}} = \sum_a \text{SD}(d)_a, \tag{3.22}$$

$$|\Delta d|_{\text{tot}} = \sum_a |\langle \Delta d \rangle_a|. \tag{3.23}$$

Refer to Table 3.1 for a summary of some of these quantities.

3.3 Results

The set of simulated networks was chosen to cover a range of values for the degree irregularity $h^2/\bar{k}^2$. It includes all network geometries in Figure 3.2, with varying values of $\boldsymbol{K} = (K_{T_1T_2}, K_{T_1T_3}, K_{T_1T_4}, K_{T_2T_3}, K_{T_2T_4}, K_{T_3T_4})$. In all networks we used an equal number of participants per arm, $n_{i,\ell} = 25$. In the motivating simulations performed by Kibret et al (2014) [25], comparable results were found for values of $n_{i,\ell} = 25, 50, 100$. Therefore, to keep the number of simulation scenarios manageable we chose to investigate only one of these values. Results for networks with an unequal number of participants per arm are discussed in Section 3.3.7. Again, following the work of Kibret et al (2014) [25] we used $\Omega = 10^3$ independent realisations of synthetic trial outcomes for any fixed set of model parameters $(\boldsymbol{d}, \tau)$. Finally, based on typical values of τ reported in real NMAs (e.g. [49]), we chose $\tau = 0.1$ throughout. For equally effective treatments (see below), this value is somewhat arbitrary. For the values $\boldsymbol{d}$ used in Section 3.3.3, this value ensured a small overlap in the true distributions of the relative treatment effects. These values themselves were chosen based on results reported in [49] for the smoking cessation network described in Section 3.2.1. Error bars in our figures are typically smaller than the size of the markers.

In Sections 3.3.1 and 3.3.2 we first focus on networks with equally effective treatments, $\boldsymbol{d} = (d_{T_1T_2}, d_{T_1T_3}, d_{T_1T_4}) = (0, 0, 0)$. We then have $P_a^{\text{bl}}(r) = 1/4$ for all

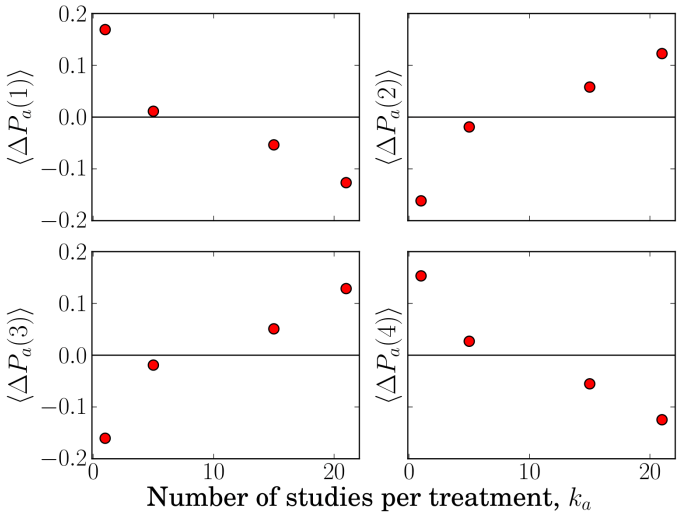


Figure 3.3: The effect of the number of studies per treatment on the bias on rank probabilities, $\Delta P_a(r)$, for $r = 1, 2, 3, 4$. These plots are for a star network with $\boldsymbol{K} = (1, 5, 15, 0, 0, 0)$.

$a \in \{T_1, \ldots, T_4\}$ and $r \in \{1, \ldots, 4\}$ (based on the true treatment effects, all four treatments have equal rank). Any systematic effect observed in the outcome of NMA is therefore a result of the structure of the network only. Networks with treatments of varying effectiveness are discussed in Section 3.3.3.

3.3.1 Comparisons within networks

In Figure 3.3 we plot bias of rank probability against number of studies per treatment, k_a, for a star network with $\boldsymbol{K} = (1, 5, 15, 0, 0, 0)$. Similar plots for other network geometries can be found in the Supplementary Material (Figures 8.1 to 8.16). This data consistently shows that the probability to be ranked best or worst, $P_a(1)$ and $P_a(4)$ respectively, is overestimated for the treatment included in the fewest studies (lowest degree k_a). The probabilities $P_a(2)$ and $P_a(3)$ are underestimated. The reverse is found for the treatment included in the most studies.

The bias of rank probability for the treatments with the most and fewest studies appears to be common in all networks. We find that the bias of the remaining two treatments can be affected by the position of their respective nodes in the network. Figure 3.4 shows the bias of $P_a(1)$ for a ladder network with $\boldsymbol{K} = (1, 0, 0, 5, 0, 15)$. In this example treatment T_2 is included in fewer studies ($k_{T_2} = 6$) than treatment T_4 ($k_{T_4} = 15$) but has more direct comparisons (it is directly compared to T_1 and T_3 whereas treatment T_4 is only directly compared to T_3). The bias on $P_a(1)$ for treatment

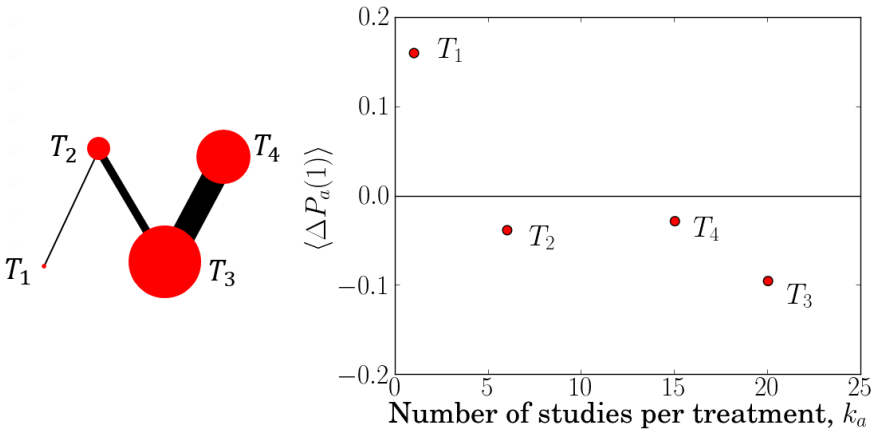


Figure 3.4: Ladder network with $\boldsymbol{K} = (1, 0, 0, 5, 0, 15)$. An example demonstrating the effect of node position on rank probability bias.

$a = T_2$ is found to be more negative than that of treatment $a = T_4$.

We conclude that disparity in the number of studies per treatment generates a trend in bias of rank probabilities. It is natural to ask if a similar trend is found for bias of treatment effect estimates. This appears not to be the case (see Figures 8.1 to 8.16 in the Supplementary Material).

Instead, the trend in $\Delta P_a(r)$ is associated with a systematic pattern in the standard deviation of treatment effect estimates, see Figure 3.5. We find that $\mathrm{SD}(d)_a$ tends to decrease with the number of studies treatment a is involved in. The standard deviation, $\mathrm{SD}(d)_a$, is particularly high for treatments with $k_a/M \lesssim 0.1$ and appears to flatten out for those included in a larger proportion of studies. We note, however, that Figure 3.5 includes data from multiple networks. On inspection of individual networks, we find a slight but consistent decrease in $\mathrm{SD}(d)_a$ as the number of studies of treatment a increases (see Figures 8.1 to 8.16 in the Supplementary Material). This data suggests that bias in the rank probabilities may originate from a variation in the uncertainties of the different treatment effect estimates. A possible mechanism for this is discussed in Section 3.4.

3.3.2 Comparisons between networks

So far we have mostly compared the outcome of NMA for different treatments within a given network. In this section we make comparisons between different networks. One main observation is a positive association between the degree irregularity of a network,

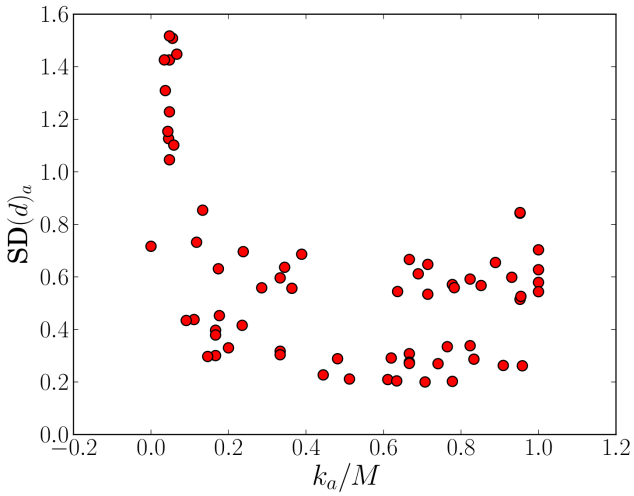


Figure 3.5: Standard deviation of treatment effect estimates. On the horizontal axis, we use the normalised number of studies, k_a/M, to capture how well connected a treatment is in the network. Each network contributes four data points, one for each treatment. The figure includes the data from all irregular networks we simulated.

$h^2/\bar{k}^2$, and the total bias on rank probabilities, $|\Delta P|_{\text{tot}}$. The data shown as red circles in Figure 3.6 demonstrate this for networks with equally effective treatments.

While we find no relationship between irregularity and total bias of relative treatment effects, $|\Delta d|_{\text{tot}}$, (see Figure 8.21 in the Supplementary Material) Figure 3.7 shows that the total standard deviation of the estimates of relative treatment effects, SD_{tot}, increases with $h^2/\bar{k}^2$. Therefore networks with a more homogeneous distribution of studies lead not only to lower bias of rank probabilities, but also to more precise estimates of relative treatment effects. This is also a possible explanation for the vertical spread in Figure 3.5. Different data points for a given value of k_a/M can be from networks with varying degrees of irregularity and hence they result in different values of $\text{SD}(d)_a$.

We find that the total bias in rank probability estimates, $|\Delta P|_{\text{tot}}$, and the total standard deviation of treatment effect estimates, SD_{tot}, are not systematically affected by the total number of studies in the network (Figures 8.22 and 8.23 in the Supplementary Material). This has implications for the planning of future studies to be added to an existing network. Naively, one may assume that adding any study to an existing network will improve parameter estimates because the amount of evidence is increased. However our results suggest that, in terms of bias on rank probabilities and the precision of treatment effect estimates, this is only true if the addition of the study

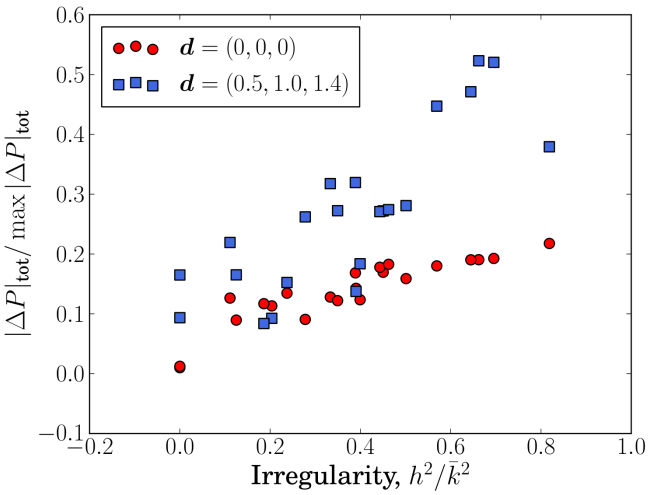


Figure 3.6: The effect of degree irregularity on a network's total rank probability bias for networks with equally effective treatments and non-equally effective treatments.

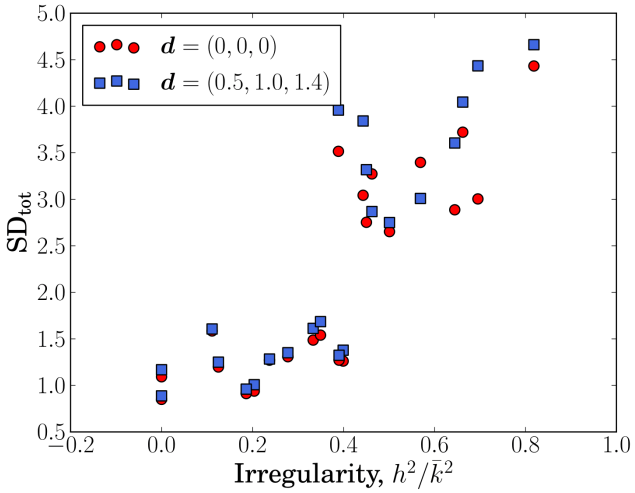


Figure 3.7: The effect of degree irregularity on a network's total standard deviation, SD_{tot}, for networks with equally effective treatments and non-equally effective treatments.

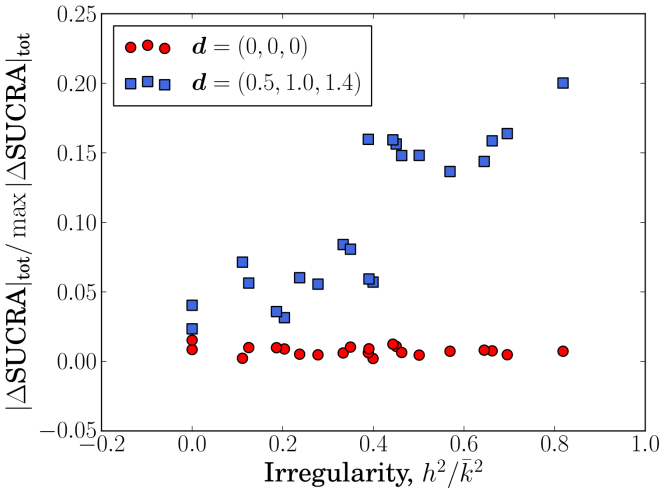


Figure 3.8: The effect of degree irregularity on a network's total SUCRA bias for networks with equally effective treatments and non-equally effective treatments.

reduces the degree irregularity, $h^2/\bar{k}^2$, of the network.

The data shown as red circles in Figure 3.8 demonstrate that, for networks with equally effective treatments, network irregularity has no effect on the total bias of SUCRA values across the network. Comparing the data in Figures 3.6 and 3.8 we find that the bias of SUCRA is approximately ten times smaller than that of the rank probabilities. This is consistent with the data in Figure 3.3 which shows that the biases of $P_a(2)$ and $P_a(3)$ are almost the exact negative of the biases of $P_a(1)$ and $P_a(4)$. These biases cancel in the calculation of SUCRA in Equation (3.5). The same reasoning also explains why, when making within-network comparisons, the number of studies per treatment has no effect on the bias of SUCRA_a (see Figure 8.17 in the Supplementary Material).

3.3.3 Treatments of varying effectiveness

The data presented so far is for networks with equally effective treatments, $\boldsymbol{d} = (0, 0, 0)$. In order to test the robustness of our findings, we now focus on a case in which the four treatments have different effectiveness. Specifically, we choose $\boldsymbol{d} = (0.5, 1.0, 1.4)$, and study the same network geometries as before. Treatment T_1 is now the most effective, followed by T_2, then T_3 and treatment T_4 is the least effective. Therefore the baseline rank probabilities are $P^{\text{bl}}_{T_1}(1) = P^{\text{bl}}_{T_2}(2) = P^{\text{bl}}_{T_3}(3) = P^{\text{bl}}_{T_4}(4) = 1$ and all other $P^{\text{bl}}_a(r)$ are zero. In order to understand this, recall that rank probabilities capture uncertainty in

the treatment effect estimates. In simulations we know the true treatment effects with certainty. When no two treatments are equally effective, the baseline probabilities then take values of zero and one.

As shown in Figure 3.7, the relationship between $\mathrm{SD}_{\mathrm{tot}}$ and degree irregularity is the same as in the case of equally effective treatments ($\boldsymbol{d} = (0, 0, 0)$); the numerical values of $\mathrm{SD}_{\mathrm{tot}}$ are also found to be largely similar, there is no systematic increase or reduction in the standard deviation.

The qualitative effect of degree irregularity, $h^2/\bar{k}^2$, on the total magnitude of the bias of rank probabilities, $|\Delta P|_{\mathrm{tot}}$, is similar compared to the case of equally effective treatments (Figure 3.6). For $h^2/\bar{k}^2 \gtrsim 0.2$ we find that the bias is larger for treatments with varying effectiveness than for equally effective treatments.

In our analysis of the case $\boldsymbol{d} = (0.5, 1.0, 1.4)$ we find that the standard deviations, $\mathrm{SD}(d_{ab})$, range from approximately 0.1 to 1.0. This means that there is significant overlap in the distributions of the estimated treatment effects. As a consequence of this, the treatments appear to be similarly effective on average. For treatments with varying effectiveness, the baseline rank probabilities (constructed from the relative treatment effects used to generate the synthetic data) take values of either zero or one. For networks with four equally effective treatments, all $P_a^{\mathrm{bl}}(r)$ are equal to 0.25. It is therefore natural that the bias of rank probabilities is greater for networks of treatments with different effectiveness, at least when the magnitude of $\mathrm{SD}(d_{ab})$ is of the same order or larger than the disparity in true treatment effects.

Figure 3.8 shows that, in contrast to the results for networks with equally effective treatments, $|\Delta\mathrm{SUCRA}|_{\mathrm{tot}}$ increases with $h^2/\bar{k}^2$ for $\boldsymbol{d} = (0.5, 1.0, 1.4)$. On inspection of the biases of rank probabilities within a given network (Figure 8.20 in the Supplement) we find that the relationship between rank probability bias and the number of studies per treatment is affected by the efficacy of the treatments that have been compared. Unlike in Figure 3.3 (where $\boldsymbol{d} = (0, 0, 0)$), the biases on $P_a(2)$ and $P_a(3)$ are not equal to $-P_a(1)$ and $-P_a(4)$ so there is no net cancellation of biases in the calculation of SUCRA_a. The total bias on rank probabilities increases for more irregular networks (higher values of $h^2/\bar{k}^2$), and as a consequence the bias on SUCRA also increases with the irregularity of the graph.

The data in Figures 3.6 to 3.8 indicate that reducing the network's irregularity

improves the precision of treatment effect estimates and reduces bias on ranking statistics in the case of treatments with varying degrees of effectiveness. Our conclusions regarding the use of network regularity for the planning of future studies are therefore also valid in this more realistic scenario.

3.3.4 Multi-arm trials

The results presented so far are for networks made up exclusively of two-arm trials. However, approximately 85% of network meta-analyses in the literature contain multi-arm trials [50]. We therefore test if our findings generalise to networks including multi-arm trials. We focus on complete-loop networks (Figure 3.2(c)) as this allows us to introduce three-arm and four-arm trials without changing the overall shape of the network (a full loop remains a full loop if further trials are added to it). The networks simulated in this section are designed specifically to cover a wide range of degree irregularities. We note that including more multi-arm trials in a network will, in general, reduce its irregularity.

For a given value of the degree irregularity, $h^2/\bar{k}^2$, we generated synthetic trial data on complete-loop networks with different combinations of two-arm, three-arm and four-arm trials. We focus on the case of equally effective treatments, and report the outcome at network-level. Results for comparisons of different treatments within networks are provided in the Supplementary Material (Figures 8.25 to 8.47). In all cases, the relationship between bias of rank probability and the number of studies per treatment follows the same pattern as in Figure 3.3.

We show $|\Delta P|_{\text{tot}}$, SD_{tot} and $|\Delta\text{SUCRA}|_{\text{tot}}$ as a function of network irregularity in Figures 3.9 to 3.11 respectively. The data from networks involving multi-arm trials is indicated by blue squares; we include the data for networks of two-arm trials (red circles) to allow comparison. As for the case of two-arm trials, the total magnitude of the bias of rank probabilities and the total standard deviation of treatment effects increase with $h^2/\bar{k}^2$, while the bias on SUCRA is largely unaffected by network irregularity. When networks are sufficiently irregular, the presence of multi-arm trials appears to reduce SD_{tot} with respect to networks consisting only of two-arm trials (Figure 3.10).

These results show that our findings concerning both within-network and between-network comparisons can be generalised to networks containing multi-arm trials.

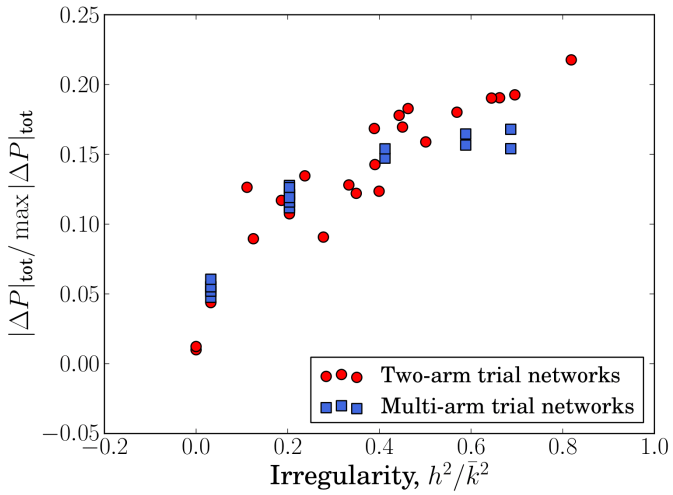


Figure 3.9: The effect of degree irregularity on a network's total rank probability bias. Data from networks with multi-arm trials is shown as blue squares.

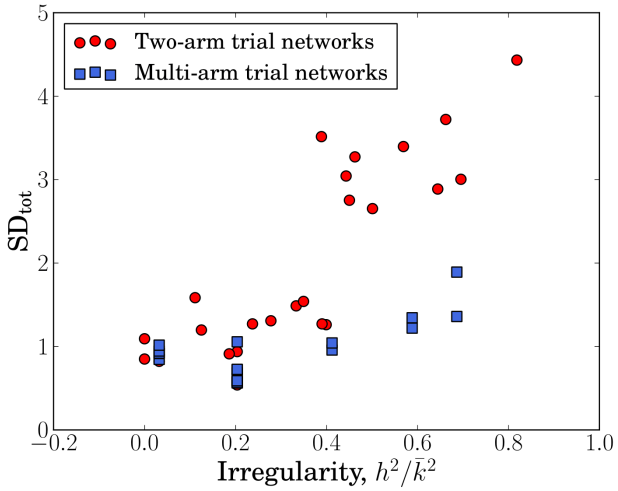


Figure 3.10: The effect of degree irregularity on a network's total standard deviation of treatment effect estimates. Data from networks with multi-arm trials is shown as blue squares.

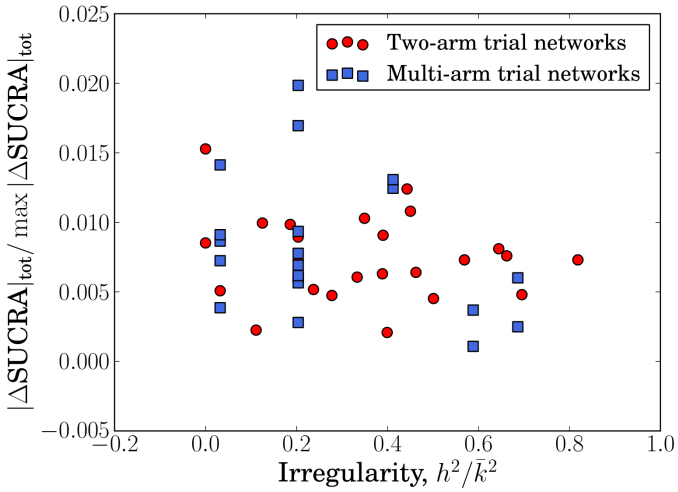


Figure 3.11: The effect of degree irregularity on a network's total bias on SUCRA values. Data from networks with multi-arm trials is shown as blue squares.

3.3.5 Data-generating models

All data so far was produced using the data-generating model 'Normal' (see Section 3.2.7). We also carried out a similar analysis using data from the 'Euclidean' and 'Uniform' methods. The only difference we observe is in the magnitude of the standard deviation of treatment effects. While the relationship between network irregularity, $h^2/\bar{k}^2$, and total standard deviation, SD_{tot}, was not affected by the choice of DGM, the values of SD_{tot} were lowest for the 'Euclidean' method and highest for the 'Uniform' method. This is not surprising as the 'Euclidean' method restricts the range of absolute outcomes that can be sampled and thus reduces variation in the event rates. The 'Uniform' method is the least restrictive in this sense. All other results were consistent between the three DGMs (see Figures 8.48 to 8.50 in the Supplementary Material). This demonstrates that the effects we observe are due to the network geometry, and are not specific to any data-generating model.

3.3.6 Bias of the heterogeneity parameter, τ

The data in Figure 3.12 shows that bias of the heterogeneity parameter, τ, decreases with the total number of studies in the network. This is the case irrespective of whether the treatments have uniform or varying effects ($\boldsymbol{d} = \mathbf{0}$ or $\boldsymbol{d} \neq \mathbf{0}$), and for networks with two-arm and multi-arm trials. The bias of τ is not affected by the network's irregularity $h^2/\bar{k}^2$ (see Figure 8.51 in the Supplementary Material). Therefore adding

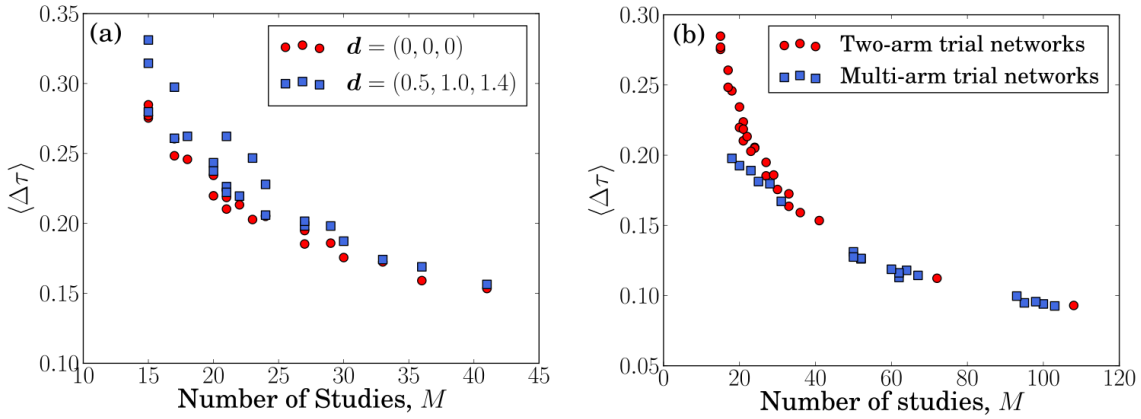


Figure 3.12: The effect of the total number of studies in a network on the accuracy of the estimate for the heterogeneity parameter τ. Panel (a) is for networks made up exclusively of two-arm trials and compares networks with equally effective and non-equally effective treatments. Panel (b) includes networks with $\boldsymbol{d} = (0, 0, 0)$ only and networks with multi-arm trials are shown as blue squares.

any trial to the network improves the accuracy of the estimate of τ.

The true value of τ in Figure 3.12 is 0.1; we note that τ is considerably overestimated in all cases we tested. To understand this, it is useful to recall that τ characterises the variation of the relative effects between any two treatments across trials in the random-effects model. Additional randomness originates from the sampling of event numbers in each trial arm. This is the case both in real-world trial data and in simulation studies (in the latter the sampling is from the binomial distributions for the respective trial arms). Some of this sampling noise may be attributed to between-trial variability by the NMA method, leading to an overestimation of τ.

A more likely cause of this bias, however, is our choice of prior distribution for τ. Naïvely, we used a uniform distribution between 0 and 5 (since this is commonly used in practice [2, 45]). However, compared to our choice of $\tau = 0.1$, the upper limit of 5 is huge. In networks with few studies, the prior distribution may dominate the posterior therefore leading to a large upward bias on τ (as observed). We note that, although other choices of the heterogeneity prior would have likely improved our estimates, this result is unlikely to affect our other findings. In particular, we have demonstrated that bias on the rank probabilities originates from *disparity* in the precision of treatment effects. While the estimate of τ may affect the absolute values of precision, the differences in precision between treatments are driven by other factors, namely the network structure.

3.3.7 Robustness tests

The simulations reported so far are for networks with four treatments and an equal number of participants per arm, $n_{i,\ell} = 25$. In order to ensure the robustness of our results we first simulated networks of ten treatments with varying irregularity. In agreement with our results for smaller networks, networks with smaller degree irregularity are again found to have more precise treatment effect estimates and smaller bias of rank probabilities (details can be found in Section 8.9.1 of the Supplementary Material). Next, we generated networks of trials in which the number of participants per arm are random numbers sampled from a flat distribution between 20 and 100. Figures 8.53 and 8.54 in the Supplementary Material show that the effect of degree irregularity is not impacted.

3.4 Summary and Discussion

3.4.1 Variation of treatment effect uncertainty is associated with biased rank probabilities

We have carried out simulation studies of network meta-analysis in random-effects models. These simulations reveal that disparity in the number of studies different treatments are involved in can lead to variation between the standard deviations of effect estimates. This in turn appears to generate a systematic bias in estimated rank probabilities. In line with previous simulations of NMA for fixed-effects models [25], the probability of a treatment being ranked best is overestimated for treatments included in the fewest number of studies, and underestimated for treatments which are part of a large number of studies. In addition, our study of networks with four treatments found the same trend for the probability of being ranked last. The probability of being ranked second and third best is subject to a bias in the opposite direction. These trends correspond to an increased standard deviation of treatment effect estimates for treatments compared in a smaller number of studies.

A general connection between standard deviation of effect estimates and bias of rank probabilities has previously been recognised in Rücker et al (2015) [16]. Our work establishes further details of the mechanics leading to biased rank probabilities.

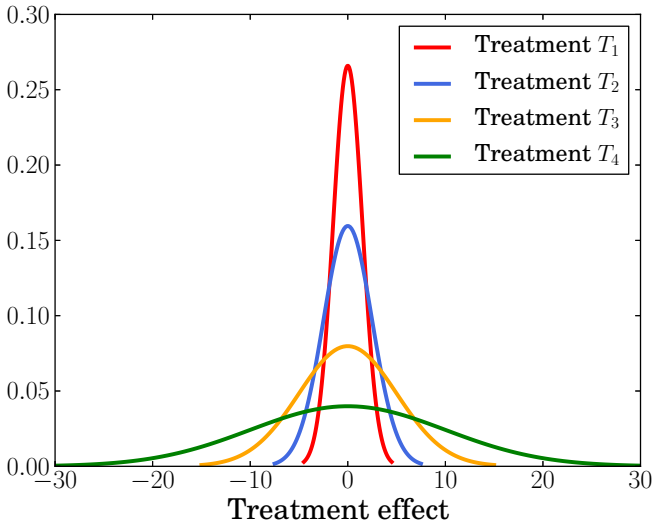


Figure 3.13: Illustrative example of posterior distributions of treatment effect estimates for four treatments in a network meta-analysis. The posterior distributions have the same mean value but varying standard deviation. Treatment T_1 has the most narrow distribution, followed by T_2, T_3 and T_4 which has the widest distribution.

We illustrate this in Figure 3.13, where we show a fictitious example of posterior distributions for the effectiveness of four different treatments. The four distributions have equal mean values, but varying standard deviations. The distribution of treatment T_4, which has the largest standard deviation, has higher density than the other treatments at very large and very small values of the treatment effect. This means that although the most probable value of the effect of treatment T_4 is the same as for the other treatments, T_4 is more likely than the other treatments to have an effect that is the largest or the smallest. Therefore treatment T_4 has the highest probability of being ranked best and the highest probability of being ranked worst. Conversely, treatment T_1 has the lowest standard deviation. Therefore, it is less likely to have extreme values of treatment effect and thus has a higher probability of being ranked second or third. Rücker et al (2015) [16] used a similar explanation to demonstrate that the probability that one treatment is better than another can be misleading when the posterior distributions of their effects have considerable overlap.

This stylised example demonstrates that biased rank probabilities can result if the uncertainty on some treatment effects is larger than on others. This effect is also to be expected when the distributions of treatment effects have different means, provided the differences in these means are small compared to their standard deviations.

Our analysis shows that the posterior distributions of treatment effect estimates

are the most narrow for treatments included in the most studies and widest for those that have been studied the least. As a consequence, biases of rank probabilities may arise if different treatments are involved in disparate numbers of studies, i.e. for large irregularity of the network.

The systematic variation in the width of posterior distributions does not, however, lead to systematic bias of SUCRA values. Since SUCRA involves a sum over the rank probabilities, the bias on these estimates is cancelled out. For networks with equally effective treatments, this cancellation is almost exact. As a result we observe very small total SUCRA bias for all values of network irregularity. In the more realistic scenario of treatments with varying efficacy, our simulations show that total bias of SUCRA does increase with irregularity, though this bias is consistently lower than that of the rank probabilities. SUCRA values therefore appear to be the more reliable ranking statistic.

Current advice in existing guidelines [6] is to report rank probabilities and treatment effects. In addition to known limitations on reporting the probability of being best, our study highlights that the full set of rank probabilities can be biased in irregular networks. It is interesting to observe that our simulations find no trend in the bias of treatment effect estimates in any of the simulated scenarios. This reinforces the importance of taking into account multiple measures, rather than making decisions based on a single metric. The weight given to different metrics may vary depending on circumstances. If the network is irregular rank probabilities become less reliable, and more weight could be given to effect estimates along with measures such as SUCRA.

The simulations presented in this paper are an explicit demonstration of biases that can occur in the comparison of multiple treatments. We have also suggested how they might originate from the structure of the network of treatments and trials. Understanding the origins of bias, we think, is vital for interpreting ranking statistics in network meta-analyses, and contributes to our understanding of the NMA method and its limitations. However, it is perhaps also interesting to consider the interpretation of 'bias' in the context of a Bayesian analysis. In particular, Bayesian probabilities reflect our subjective beliefs based on the available data and any prior evidence. Therefore it is not immediately clear how exactly such a belief can be quantified in terms of 'bias'. In our work we defined bias to be the difference between the estimated rank

probabilities and what we called 'baseline probabilities'. These latter values reflect the rank probabilities one would obtain from an NMA if it were able to estimate the relative treatment effects with perfect accuracy and no uncertainty. Although such a result is never attainable in practice, deviations from these values are still informative. For example, we have demonstrated that the conclusions drawn from estimated rank probabilities do not necessarily reflect the data. In particular, our results show that a high probability of being best or worst may actually reflect imprecision in estimated treatment effects rather than the magnitude of the effects themselves. Therefore, our definition of 'bias' is still a useful concept to consider when interpreting the results of an NMA.

We note that the definition of degree irregularity in Equation (3.11) does not explicitly account for variation in the number of participants across arms. It would be interesting to investigate if and how the definition of degree irregularity could be modified to take into account varying numbers of participants. For example, one could modify Equations (3.10) and (3.11) so that in the calculation of the mean degree and h^2, treatments are weighted by the number of participants rather than the number of trials. Further work could then investigate how quantities such as rank probabilities and SUCRA values are affected by variation in participant numbers.

3.4.2 Planning future studies to reduce the irregularity of the network

Planning future clinical trials based on existing evidence from network meta-analysis can reduce the resources and number of participants required to obtain results of a given precision [51–53]. While the design of future trials based on pairwise meta-analysis has received significant attention [54–56], methods using the outcome of network meta-analysis are less developed. Current approaches in this area [52, 57] are computationally intensive and become increasingly laborious as the network becomes more complex. Our results show that the degree irregularity of a network, $h^2/\bar{k}^2$, can provide guidance on the choice of future trials without the need for extensive simulations. The degree of a treatment in the graph is the number of trials it is involved in, and the irregularity of a network describes how this degree varies across treatments. This is easily obtained

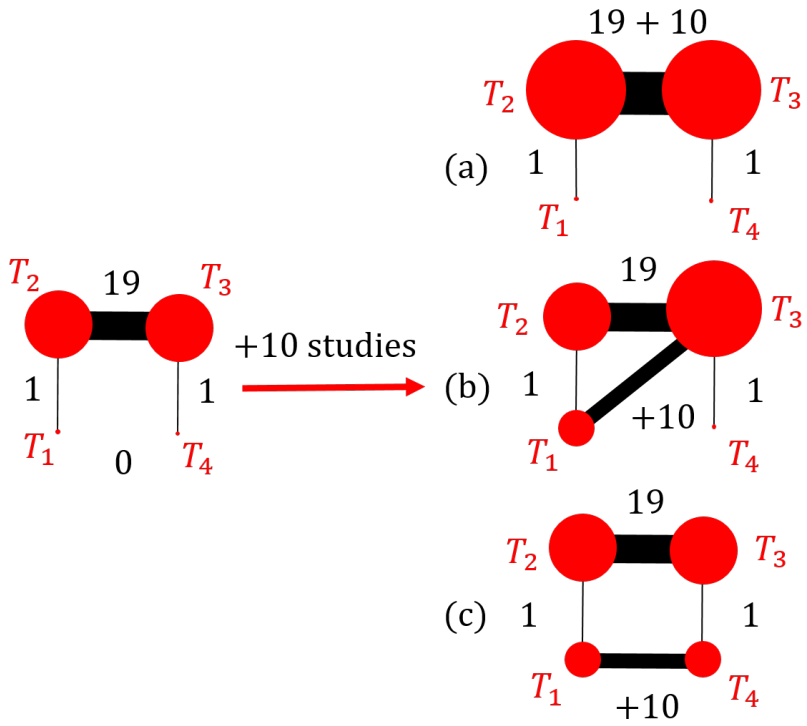


Figure 3.14: Example of three different geometries that can be created by adding a fixed number of studies to an existing network.

from the network.

Irregularity is a better indicator of the accuracy and precision of parameter estimates from a network meta-analysis than the total number of studies in the network. As we have shown, networks with a more homogeneous distribution of studies between treatments have more precise treatment effect estimates and smaller bias of rank probabilities.

Degree irregularity is therefore a useful metric for working out which comparisons could be made in future studies to improve the parameter estimates of an existing NMA. For example, consider a network of four treatments with $\boldsymbol{K} = (1, 0, 0, 19, 0, 1)$ and $h^2/\bar{k}^2 = 0.82$ as shown on the left in Figure 3.14. Now imagine resources are available to add ten new two-arm studies to this network. If (a) we add all ten studies to the most connected comparison $(T_2 - T_3)$ then we obtain $\boldsymbol{K} = (1, 0, 0, 29, 0, 1)$, and the network's irregularity increases to 0.88. We may be more inclined to populate a comparison that currently has no direct evidence such as $T_1 - T_3$ [(b) in Figure 3.14] or $T_1 - T_4$ [(c)]. The former leads to $\boldsymbol{K} = (1, 10, 0, 19, 0, 1)$ and reduces $h^2/\bar{k}^2$ to 0.48, while the latter has $\boldsymbol{K} = (1, 0, 10, 19, 0, 1)$ and reduces $h^2/\bar{k}^2$ to 0.08. These three possible 'future' networks are shown on the right-hand side in Figure 3.14.

By simulating the original network and the three 'future' networks whilst keeping all other network characteristics constant, we compare how adding the extra ten studies affects the results. Table 3.2 summarises the total standard deviation and total rank probability bias of these four networks.

Table 3.2: Degree irregularity, precision and accuracy of NMA parameter estimates for the networks in Figure 3.14.

Network	M	$h^2/\bar{k}^2$	$\mathrm{SD}_{\mathrm{tot}}$	$\lvert\Delta P\rvert_{\mathrm{tot}}$
Original:	21	0.82	4.44	1.74
(a):	31	0.88	4.36	1.73
(b):	31	0.48	3.24	1.47
(c):	31	0.08	1.68	0.16

For network (a) the precision and accuracy of the NMA parameter estimates (as measured by $\mathrm{SD}_{\mathrm{tot}}$ and $|\Delta P|_{\mathrm{tot}}$ respectively) are approximately the same as for the original network, whereas (b) and (c) show a considerable reduction in $\mathrm{SD}_{\mathrm{tot}}$ and $|\Delta P|_{\mathrm{tot}}$. The improvement in both quantities for network (c) is markedly greater than in network (b) even though in both cases the ten new studies were added to a comparison with no existing direct evidence.

This example demonstrates that equality in the number of studies per treatment is more important than equality in the number of studies per comparison. A similar example involving a network of ten treatments is discussed in Section 8.9.1 in the Supplementary Material.

Choosing future studies that reduce degree irregularity may therefore help to improve the precision of treatment effect estimates and the accuracy of rank probabilities, though this is, of course, subject to constraints in practice. For example a treatment may appear in comparatively few trials for good reasons, such as cost, negative side effects, or the treatment is outdated. In other scenarios one treatment option outranks all others by a very large margin [58, 59]. It is then largely irrelevant whether the SUCRA value of this treatment option can be made more precise as this would unlikely change the overall ranking of the treatment. Nevertheless, we think it is useful to be aware how degree regularity affects the precision and accuracy of NMA. This metric can then be used as one contributing factor in the planning of future trials, along with practical and clinical considerations.

As a final remark, we note that we have focused on network meta-analyses in a Bayesian setting. While quantities such as rank probabilities and SUCRA are more natural in Bayesian analysis, similar quantities can be obtained in a frequentist framework based on resampling methods [14]. Furthermore, Rücker et al (2015) [16] proposed a frequentist analogue to SUCRA calculated without resampling. Quantitatively, this was found to be nearly identical to SUCRA derived from a Bayesian NMA. This, along with other work which has found that treatment ranking in Bayesian and frequentist NMA are consistent [60], leads us to expect that the results in this paper would generalise to frequentist network meta-analysis.

3.5 Appendix A: Degree irregularity and probability of inter-specific encounter (PIE)

In network meta-analysis (NMA) the probability of inter-specific encounter (PIE) index measures the probability that two randomly sampled treatment groups (trial arms) from the network are associated with two different treatments [22, 23]. The sampling is understood to occur *without* replacement. This diversity measure was originally introduced in ecology [61], and first applied in the context of NMA by Salanti et al. (2008) [22, 23].

The probability that the two sampled arms represent different treatments is given by

$$\mathrm{PIE} = 1 - \sum_a \left(\frac{k_a}{k_{\mathrm{tot}}}\right)\left(\frac{k_a - 1}{k_{\mathrm{tot}} - 1}\right), \tag{3.24}$$

where k_{tot} is the total number of arms in the network,

$$k_{\mathrm{tot}} = \sum_a k_a = N\bar{k}. \tag{3.25}$$

We recall that $\bar{k} = N^{-1}\sum_a k_a$ is the mean degree of a node in the weighted network, and N the total number of treatments.

In Equation (3.24), k_a/k_{tot} is the probability that a randomly picked trial arm is of type a, and $(k_a - 1)/(k_{\mathrm{tot}} - 1)$ is the probability that an arm sampled randomly from the remaining $k_{\mathrm{tot}} - 1$ arms is also of type a. Using Equation (3.25) and a modest

amount of algebra one shows that PIE can be written in the more commonly used form

$$\text{PIE} = \frac{k_{\text{tot}}}{k_{\text{tot}} - 1}\left[1 - \sum_a \left(\frac{k_a}{k_{\text{tot}}}\right)^2\right]. \tag{3.26}$$

PIE is a probability and takes values between zero and one. PIE' is defined as PIE normalised with respect to the maximum value of PIE for a given number of studies,

$$\text{PIE}' = \frac{\text{PIE}}{\max(\text{PIE})}. \tag{3.27}$$

At fixed N PIE takes its maximum value when $k_a = \bar{k}$ for all treatments a. This means that the fraction of arms associated with any one treatment is $k_a/k_{\text{tot}} = 1/N$ for all a. In this case therefore

$$\max(\text{PIE}) = \frac{k_{\text{tot}}}{k_{\text{tot}} - 1}\left(1 - \frac{1}{N}\right). \tag{3.28}$$

In order to relate h^2 to PIE' we start from Equation (3.11) in the main paper. We have

$$\begin{aligned} h^2 &= \frac{1}{N}\sum_a (k_a - \bar{k})^2 \\ &= \frac{1}{N}\left(\sum_a k_a^2 - 2\bar{k}\sum_a k_a + N\bar{k}^2\right) \\ &= \frac{1}{N}\left(\sum_a k_a^2 - N\bar{k}^2\right). \end{aligned} \tag{3.29}$$

Therefore, using Equation (3.25),

$$\frac{h^2}{\bar{k}^2} = \frac{1}{N\bar{k}^2}\sum_a k_a^2 - 1 \tag{3.30}$$

$$= \frac{N}{k_{\text{tot}}^2}\sum_a k_a^2 - 1. \tag{3.31}$$

From the definitions of PIE and max(PIE) we have

$$\begin{aligned} \text{PIE}' &= \frac{1 - \sum_a \left(\frac{k_a}{k_{\text{tot}}}\right)^2}{1 - \frac{1}{N}} \\ &= \frac{N - \frac{N}{k_{\text{tot}}^2}\sum_a k_a^2}{N - 1} \\ &= \frac{N - \left(\frac{h^2}{\bar{k}^2} + 1\right)}{N - 1}. \end{aligned} \tag{3.32}$$

Therefore

$$\text{PIE}' = 1 - \frac{1}{N-1}\frac{h^2}{\bar{k}^2}. \tag{3.33}$$

3.6 Appendix B: Maximum total bias

To compare the extent of total rank probability bias and total SUCRA bias, we express these measures as a proportion of the maximum bias that is possible to observe in each case. In this section we calculate the values of these maxima.

3.6.1 Maximum total rank probability bias

The sets of true and estimated rank probability biases, $P_a(r)$ and $\hat{P}_a(r)$, each form doubly stochastic matrices. We call these matrices $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ such that their elements are $P_{ij} = P_{T_i}(j)$ and $\hat{P}_{ij} = \hat{P}_{T_i}(j)$. The properties of a doubly stochastic matrices are

$$\sum_{i=1}^{N} P_{ij} = 1, \qquad \sum_{j=1}^{N} P_{ij} = 1, \qquad P_{ij} \geq 0. \tag{3.34}$$

That is to say, all matrix elements are positive, and all elements in any row of $\boldsymbol{P}$ sum to one, and similarly, the sum of elements in any column is one. Analogous relations hold for $\hat{\boldsymbol{P}}$.

The total rank probability bias can be written as

$$|\Delta P|_{\text{tot}} = \sum_{i=1}^{N}\sum_{j=1}^{N} |P_{ij} - \hat{P}_{ij}|. \tag{3.35}$$

We can work out the maximum of this quantity by using the triangle inequality

$$|P_{ij} - \hat{P}_{ij}| \leq P_{ij} + \hat{P}_{ij}. \tag{3.36}$$

Therefore

$$\begin{aligned} |\Delta P|_{\text{tot}} &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}(P_{ij} + \hat{P}_{ij}) \\ &= \sum_{i=1}^{N}\underbrace{\sum_{j=1}^{N} P_{ij}}_{=1} + \sum_{i=1}^{N}\underbrace{\sum_{j=1}^{N} \hat{P}_{ij}}_{=1} \\ &= N + N = 2N. \end{aligned} \tag{3.37}$$

This bound is tight, for example it is saturated if $\boldsymbol{P}$ is the identity matrix, and $\hat{\boldsymbol{P}}$ a permutation matrix mapping no number onto itself.

For $N = 4$ treatments one has

$$\max\left(|\Delta P|_{\text{tot}}\right) = 8. \tag{3.38}$$

3.6.2 Maximum total SUCRA bias

To work out the maximum value of total SUCRA bias we first write it in terms of the $\{\Delta P_a(r)\}$,

$$|\Delta\text{SUCRA}|_{\text{tot}} = \sum_a |\Delta\text{SUCRA}_a| = \sum_a \left| -\frac{\sum_r r\hat{P}_a(r) - \sum_r rP_a(r)}{N-1} \right|, \tag{3.39}$$

Therefore we have

$$|\Delta\text{SUCRA}|_{\text{tot}} = \frac{1}{N-1}\sum_a \left| -\sum_r r\Delta P_a(r) \right|. \tag{3.40}$$

Again using the doubly stochastic matrices $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ to represent the true and estimated rank probabilities, we can write

$$|\Delta\text{SUCRA}|_{\text{tot}} = \frac{1}{N-1}\sum_{i=1}^{N} |\sum_{j=1}^{N} j(P_{ij} - \hat{P}_{ij})|. \tag{3.41}$$

The minimum possible value of $\sum_{j=1}^{N} jP_{ij}$ is 1 and the maximum is N. Similarly, $\sum_{j=1}^{N} j\hat{P}_{ij}$ takes values between 1 and N. Therefore for a fixed value of i

$$|\sum_{j=1}^{N} j(P_{ij} - \hat{P}_{ij})| \leq N-1. \tag{3.42}$$

However, due to the doubly stochastic nature of $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$, equality can hold in this relation for only two values of i, namely one in which $\sum_{j=1}^{N} jP_{ij} = 1$ and $\sum_{j=1}^{N} j\hat{P}_{ij} = N$, and the other for which the reverse holds.

The next largest value that $\left|\sum_{j=1}^{N} j(P_{ij} - \hat{P}_{ij})\right|$ can take is $N-3$ (for $\sum_{j=1}^{N} jP_{ij} = N-1$ and $\sum_{j=1}^{N} j\hat{P}_{ij} = 2$ or vice versa). Following this pattern, the maximum values of $\left|\sum_{j=1}^{N} j(P_{ij} - \hat{P}_{ij})\right|$ are $N-1, N-3, N-5, N-7, ..., N-7, N-5, N-3, N-1$. For even $N = 2k$ this gives

$$\begin{aligned}
\max\left(\sum_{i=1}^{N} |\sum_{j=1}^{N} j(P_{ij} - \hat{P}_{ij})|\right) &= 2\sum_{l=0}^{k-1}(2k-1-2l) \\
&= 2\Bigg[k(2k-1) - 2\underbrace{\sum_{l=0}^{k-1} l}_{=(k-1)k/2}\Bigg] \\
&= 2k^2 = \frac{N^2}{2}.
\end{aligned} \tag{3.43}$$

For $N = 4$ treatments we have

$$\max(|\Delta\text{SUCRA}|_{\text{tot}}) = \frac{1}{N-1}\frac{N^2}{2} = \frac{8}{3}. \tag{3.44}$$

Technically, the above argument only produces a lower bound on the maximum value of $|\Delta\text{SUCRA}|_{\text{tot}}$. However, we have tested the relation in Equation (3.43) numerically using large samples of randomly generated doubly stochastic matrices. We have found no instances in which the maximum value indicated is higher than the bound in Equation (3.43).

Data Availability Statement

The data that supports the findings of this study was generated via simulations, using the algorithm described in the manuscript. An example of the type of code used in our analysis can be found here https://github.com/AnnieDavies/Supplementary-Material-Davies-Galla-2020.

Bibliography

[1] G. Salanti, "Indirect and mixed treatment comparison, network, or multiple treatments meta analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool", Res. Synth. Meth. **3**, 80–97 (2012).

[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*, Online, Last updated September 2016; Available from `http://www.nicedsu.org.uk`. Accessed March 2020, 2011.

[3] S. Dias, A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton, *Network meta-analysis for decision making* (Wiley, Oxford, UK, 2018).

[4] G. Lu and A. E. Ades, "Combination of direct and indirect evidence in mixed treatment comparisons", Stat. Med. **23**, 3105–3124 (2004).

[5] D. C. Hoaglin, N. Hawkins, J. P. Jansen, D. A. Scott, R. Itzler, J. C. Cappelleri, C. Boersma, D. Thompson, K. M. Larholt, M. Diaz, and A. Barrett, "Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices—part 2", Value. Health **14**, 429–437 (2011).

[6] B. Hutton, G. Salanti, D. M. Caldwell, A. Chaimani, C. H. Schmid, C. Cameron, J. P. A. Ioannidis, S. E. Straus, K. Thorlund, J. P. Jansen, C. Mulrow, F. Catalá-López, P. C. Gøtzsche, K. Dickersin, I. Boutron, D. G. Altman, and D. Moher, "The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations", Ann. Intern. Med. **162**, 777–784 (2015).

[7] K. Pateras, S. Nikolakopoulos, and K. Roes, "Data-generating models of dichotomous outcomes: heterogeneity in simulation studies for a random-effects meta-analysis", Stat. Med. **37**, 1115–1124 (2018).

[8] R. J. Hardy and S. G. Thompson, "A likelihood approach to meta-analysis with random effects", Stat. Med. **15**, 619–629 (1996).

[9] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods", Stat. Med. **38**, 2074–2102 (2019).

[10] C. J. Geyer, "Chapter 1: Introduction to Markov chain Monte Carlo", in *Handbook of Markov chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X. Meng (CRC Press, Boca Raton, FL, USA, 2011), pp. 3–48.

[11] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "WinBUGS - a Bayesian modelling framework: concepts, structure and extensibility", Stat. Comput. **10**, 325–337 (2000).

[12] A. Bafeta, L. Trinquart, R. Seror, and P. Ravaud, "Reporting of results from network meta-analyses: methodological systematic review", BMJ **348**, g1741 (2014).

[13] G. Salanti, A. E. Ades, and J. P. A. Ioannidis, "Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial", J. Clin. Epidemiol. **64**, 163–171 (2011).

[14] I. R. White, J. K. Barrett, D. Jackson, and J. P. T. Higgins, "Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression", Res. Synth. Meth. **3**, 111–125 (2012).

[15] L. Trinquart, N. Attiche, A. Bafeta, R. Porcher, and P. Ravaud, "Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomised trials", Ann. Intern. Med. **164**, 666–673 (2016).

[16] G. Rücker and G. Schwarzer, "Ranking treatments in frequentist network meta-analysis works without resampling methods", BMC Med. Res. Methodol. **15**, 58 (2015).

[17] A. A. Veroniki, S. E. Straus, A. Fyraridis, and A. C. Tricco, "The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes", J. Clin. Epidemiol. **76**, 193–199 (2016).

[18] A. A. Veroniki, S. E. Straus, G. Rücker, and A. C. Tricco, "Is providing uncertainty intervals in treatment ranking helpful in network meta-analysis?", J. Clin. Epidemiol. **100**, 122–129 (2018).

[19] C. H. Daly, B. Neupane, J. Beyene, L. Thabane, S. E. Straus, and J. S. Hamid, "Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa", BMJ Open **9**, e024625 (2019).

[20] A. Chaimani, R. Porcher, É. Sbidian, and D. Mavridis, "A Markov chain approach for ranking treatments in network meta-analysis", Stat. Med. **40**, 451–464 (2021).

[21] V. Chiocchia, A. Nikolakopoulou, T. Papakonstantinou, M. Egger, and G. Salanti, "Agreement between ranking metrics in network meta-analysis: an empirical study", BMJ Open **10**, e037744 (2020).

[22] G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis, "Evaluation of networks of randomized trials", Stat. Methods. Med. Res. **17**, 279–301 (2008).

[23] G. Salanti, F. K. Kavvoura, and J. P. A. Ioannidis, "Exploring the geometry of treatment networks", Ann. Intern. Med. **148**, 544–553 (2008).

[24] P. Dequen, A. J. Sutton, D. A. Scott, and K. R. Abrams, "Searching for indirect evidence and extending the network of studies for network meta-analysis: case study in venous thromboembolic events prevention following elective total knee replacement surgery", Value. Health **17**, 416–423 (2014).

[25] T. Kibret, D. Richer, and J. Bayene, "Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study", Clin. Epidemiol. **6**, 451–460 (2014).

[26] J. P. Jansen, T. Trikalinos, J. C. Cappelleri, J. Daw, S. Andes, R. Eldessouki, and G. Salanti, "Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC good practice task force report", Value. Health **17**, 157–173 (2014).

[27] M. Petropoulou, A. Nikolakopoulou, A. A. Veroniki, P. Rios, A. Vafaei, W. Zarin, M. Giannatsi, S. Sullivan, A. C. Tricco, A. Chaimani, M. Egger, and G. Salanti, "Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015", J. Clin. Epidemiol. **82**, 20–28 (2017).

[28] D. E. Jonas, T. M. Wilkins, S. Bangdiwala, C. M. Bann, L. C. Morgan, K. J. Thaler, H. R. Amick, and G. Gartlehner, "Findings of Bayesian mixed treatment comparison meta-analyses: comparison and exploration using real-world trial data and simulation (internet)", Rockville (MD): Agency for Healthcare Research and Quality, Available from: https://www.ncbi.nlm.nih.gov/books/NBK126100/. Accessed March 2020 (2013).

[29] F. S. Tonin, H. H. Borba, A. M. Mendes, A. Wiens, F. Fernandez-Llimos, and R. Pontarolo, "Description of network meta-analysis geometry: a metrics design study", PLOS ONE **14**, e0212650 (2019).

[30] S. E. Seide, K. Jensen, and M. Kieser, "Simulation and data-generation of random-effects network meta-analysis of binary outcome", Stat. Med. **38**, 3288–3303 (2019).

[31] T. A. B. Snijders, "The degree variance: an index of graph heterogeneity", Soc. Netw. **3**, 163–174 (1981).

[32] G. Rücker, C. Rücker, and I. Gutman, "On kites, comets, and stars. Sums of eigenvector coefficients in (molecular) graphs", Z. Naturforsch **57**, 143–153 (2002).

[33] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Methods and Principles in Medicinal Chemistry (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2000).

[34] V. Hasselblad, "Meta-analysis of multi-treatment studies", Med. Decis. Making **18**, 37–43 (1998).

[35] M. Newman, *Networks*, 2nd ed. (Oxford University Press, Oxford, UK, 2018).

[36] T. H. Hamza, H. C. van Houwelingen, and T. Stijnen, "The binomial distribution of meta-analysis was preferred to model within-study variability", J. Clin. Epidemiol. **61**, 41–51 (2008).

[37] P. McCullagh and J. A. Nelder, *Generalized linear models*, 2nd ed. (Chapman and Hall, Boca Raton, FL, USA, 1989).

[38] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials", Control. Clin. Trials **7**, 177–188 (1986).

[39] H. Hong, H. Fu, K. L. Price, and B. P. Carlin, "Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment", Stat. Med. **34**, 2794–2819 (2015).

[40] S. Dias and A. E. Ades, "Absolute or relative effects? Arm-based synthesis of trial data", Res. Synth. Meth. **7**, 23–28 (2016).

[41] J. P. T. Higgins and A. Whitehead, "Borrowing strength from external trials in a meta-analysis", Stat. Med. **15**, 2733–2749 (1996).

[42] T. Lumley, "Network meta-analysis for indirect treatment comparisons", Stat. Med. **21**, 2313–2324 (2002).

[43] G. Lu and A. E. Ades, "Assessing evidence inconsistency in mixed treatment comparisons", J. Am. Stat. Assoc. **101**, 447–459 (2006).

[44] T. C. Smith, D. J. Spiegelhalter, and A. Thomas, "Bayesian approaches to random effects meta analysis: a comparative study", Stat. Med. **14**, 2685–2699 (1995).

[45] T. Greco, G. Landoni, G. Biondi-Zoccai, F. D'Ascenzo, and A. Zangrillo, "A Bayesian network meta-analysis for binary outcome: how to do it", Stat. Methods. Med. Res. **25**, 1757–1773 (2016).

[46] C. P. Robert and G. Casella, "Chapter 6: Metropolis-Hastings Algorithms", in *Introducing Monte Carlo Methods with R* (Springer, New York, NY, USA, 2010), pp. 169–197.

[47] S. M. Lynch, *Introduction to applied Bayesian statistics and estimation for social scientists* (Springer, New York, NY, USA, 2007), pp. 77–130.

[48] N. Trinajstic, *Chemical graph theory*, 2nd ed. (Routledge, FL, USA, 2018).

[49] S. Dias, N. J. Welton, D. M. Caldwell, and A. E. Ades, "Checking consistency in mixed treatment comparison meta-analysis", Stat. Med. **29**, 932–944 (2010).

[50] A. Nikolakopoulou, A. Chaimani, A. A. Veroniki, H. S. Vasiliadis, C. H. Schmid, and G. Salanti, "Characteristics of networks of interventions: a description of a database of 186 published networks", PLOS ONE **9**, e86754 (2014).

[51] G. Salanti, A. Nikolakopoulou, A. Sutton, S. Reichenbach, S. Trelle, H. Naci, and M. Egger, "Planning a future randomized clinical trial based on a network of relevant past trials", Trials **19**, 365 (2018).

[52] A. Nikolakopoulou, D. Mavridis, and G. Salanti, "Planning future studies based on the precision of network meta-analysis results", Stat. Med. **35**, 978–1000 (2015).

[53] J. P. A. Ioannidis, S. Greenland, M. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz, and R. Tibshirani, "Increasing value and reducing waste in research design, conduct, and analysis", The Lancet **383**, 166–175 (2014).

[54] A. J. Sutton, N. J. Cooper, D. R. Jones, P. C. Lambert, J. R. Thompson, and K. R. Abrams, "Evidence-based sample size calculations based upon updated meta-analysis", Stat. Med. **26**, 2479–2500 (2007).

[55] A. J. Sutton, N. J. Cooper, and D. R. Jones, "Evidence synthesis as the key to more coherent and efficient research", BMC Med. Res. Methodol. **9**, 29 (2009).

[56] V. Roloff, J. P. T. Higgins, and A. J. Sutton, "Planning future studies based on the conditional power of a meta-analysis", Stat. Med. **32**, 11–24 (2013).

[57] A. Nikolakopoulou, D. Mavridis, and G. Salanti, "Using conditional power of network meta-analysis (NMA) to inform the design of future clinical trials", Biometrical J. **56**, 973–990 (2014).

[58] K. Linde, G. Rücker, K. Sigterman, S. Jamil, K. Meissner, A. Schneider, and L. Kriston, "Comparative effectiveness of psychological treatments for depressive disorders in primary care: network meta-analysis", BMC Fam. Pract. **16**, 103 (2015).

[59] K. Linde, G. Rücker, A. Schneider, and L. Kriston, "Questionable assumptions hampered interpretation of a network meta-analysis of primary care depression treatments", J. Clin. Epidemiol. **71**, 86–96 (2016).

[60] B. Sadeghirad, R. Brignardello-Petersen, B. C. Johnston, G. H. Guyatt, and J. Beyene, *Comparing Bayesian and frequentist approaches for network meta-analysis: an empirical study*, Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews, (9 Suppl 1), 2017.

[61] S. H. Hurlbert, "The nonconcept of species diversity: a critique and alternative parameters", Ecology **52**, 577–586 (1971).

Chapter 4

Network meta-analysis and random walks

Preface

The contents of this chapter constitute a manuscript published by Statistics in Medicine[1]. What was originally published as online Supplementary Material now appears as an Appendix at the end of the main text. The manuscript was authored by Annabel L Davies[2], Theodoros Papakonstantinou[3], Adriani Nikolakopoulou[3], Gerta Rücker[3] and Tobias Galla[2,4].

ALD designed the study, contributed to discussions guiding the work, carried out the mathematical calculations, performed the data analysis, wrote the first draft of the manuscript, produced all of the figures and edited the manuscript. TP, AN, GR and TG designed the study, contributed to discussions guiding the work and edited the manuscript.

[1]A. L. Davies, T. Papakonstantinou, A. Nikolakopoulou, G. Rücker and T. Galla, "Network meta-analysis and random walks", Stat. Med. 1-24 (2022). 10.1002/sim.9346

[2]Theoretical Physics, School of Physics and Astronomy, The University of Manchester, Manchester, M13 9PL, United Kingdom.

[3]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg, Germany

[4]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

Abstract

Network meta-analysis (NMA) is a central tool for evidence synthesis in clinical research. The results of an NMA depend critically on the quality of evidence being pooled. In assessing the validity of an NMA, it is therefore important to know the proportion contributions of each direct treatment comparison to each network treatment effect. The construction of proportion contributions is based on the observation that each row of the hat matrix represents a so-called 'evidence flow network' for each treatment comparison. However, the existing algorithm used to calculate these values is associated with ambiguity according to the selection of paths. In this work we present a novel analogy between NMA and random walks. We use this analogy to derive closed-form expressions for the proportion contributions. A random walk on a graph is a stochastic process that describes a succession of random 'hops' between vertices which are connected by an edge. The weight of an edge relates to the probability that the walker moves along that edge. We use the graph representation of NMA to construct the transition matrix for a random walk on the network of evidence. We show that the net number of times a walker crosses each edge of the network is related to the evidence flow network. By then defining a random walk on the directed evidence flow network, we derive analytically the matrix of proportion contributions. The random-walk approach has none of the associated ambiguity of the existing algorithm.

4.1 Introduction

Network meta-analysis (NMA) has been established as a central tool of evidence synthesis in clinical research [1–3]. Combining direct and indirect evidence from multiple randomised controlled trials, NMA makes it possible to compare interventions that have not been tested together in any trial [4–6]. The term 'network meta-analysis' derives from the fact that one can mathematically represent the collection of interventions and trials as a graph. A graph consists of a set of nodes and a set of edges connecting pairs of nodes. The nodes of an NMA graph represent the different

treatment options, and edges are comparisons made between the treatments in the trials. In line with Rücker (2012) [7] we will refer to networks of treatment options and comparisons between treatments as 'meta-analytic graphs'.

An NMA combines data from multiple trials, each comparing different combinations of treatment options. The accuracy of the conclusions from an NMA depends on potential biases associated with individual trials, and on assumptions such as between-trial homogeneity and consistency between direct and indirect evidence. In this context it is useful to study the so-called 'flow of evidence' [8] in the network. This describes the influence different network components have on the estimates of treatment effects. For example, the comparison between two particular treatments may enter as indirect evidence into the estimate of the relative effect of two different nodes in the network. Understanding how exactly evidence flows in the graph then allows one to assess the impact of potential bias originating from different pieces of evidence in the network [8–10].

Previous literature has, for example, looked at the relative influence of direct evidence compared to indirect evidence [11, 12]. Other work has been concerned with measures of network geometry, capturing the frequency with which different comparisons are represented in the trials underpinning an NMA [13, 14]. One then asks how the network structure affects NMA estimates of treatment effects, heterogeneity and rank metrics [13–18]. The 'hat matrix' in a two-step ('aggregate') NMA model [11] describes how the overall estimates of treatment effects from the network can be expressed in terms of the direct estimates obtained from the trial data. König et al (2013) [8] observed that each row of the hat matrix represents an evidence flow network for a particular treatment effect. König et al then visualised the evidence flow on weighted directed acyclic graphs in which nodes represent treatments, and edges indicate the direction and quantity of evidence flow through each direct comparison. Based on this observation, Papakonstantinou et al (2018) [9] introduced 'streams' of evidence and developed a numerical algorithm to calculate these streams. The streams of evidence are then used to derive the 'proportion contribution' of each direct comparison to each treatment effect in the graph. This allows one to quantify how limitations of individual studies impact on the estimates obtained from the network. Indeed, the algorithm in Papakonstantinou et al is implemented in software such as

CINeMA (confidence in network meta-analysis) [10] and ROB-MEN (risk-of-bias due to missing evidence in NMA) [19], used in clinical practice for the evaluation of results from an NMA.

More widely, the study of networks plays a key role in a variety of disciplines including ecology, economics, electrical engineering and sociology [20–22]. Through the representation of treatment options and comparisons in trials as a graph, one can therefore take advantage of the extensive literature in network theory, and of ideas developed in the disciplines in which networks are studied. For example one of us [7] used the graph representation of NMA to make the connection between meta-analytic and electrical networks. This allows one to demonstrate that graph theoretical tools routinely applied to electrical networks are also of use in NMA. This approach has since led to advancements in NMA methodology such as frequentist ranking methods [23] and component NMA [24]. It is also the basis for the software package `netmeta` [25].

In this paper we present a new analogy between random walks and NMA. A random walk on a graph is a stochastic process consisting of a succession of 'hops' between vertices connected by edges. Random walks are of interest for a wide range of applications, including statistical physics, biology, ecology, genetics, transport and economics (for a selection of references see [26–30]). Random walks are also a popular tool to study the properties of networks themselves [31–33].

It is well known that there is a connection between random walks and electrical networks [34–37]. In this context, edges of the electric network are conducting connections (wires). The correspondence between random walks and electrical networks can be established by asserting that the probability that a random walker currently at node a moves to node b in the next step is proportional to the conductance (inverse resistance) of the edge connecting a and b. Quantities in the electrical network such as currents along edges or electric potentials at the nodes then have an interpretation in the random-walk picture. For further details we refer to Doyle and Snell (2000) [37].

Motivated by the connection between electrical networks and NMA on the one hand, and that of electrical networks and random walks on the other, we construct a random walk on the *meta-analytic network*. We show that the random-walk picture we develop can be used to study the flow of evidence in the NMA network. In particular there is a

random-walk interpretation of the elements in the hat matrix. Further, we construct a second random-walk model, this time on the *evidence flow network*. From this we derive an analytical expression for proportion contributions which overcomes the limitations of Papakonstantinou et al [9]. In particular, the algorithm in Papakonstantinou et al selects only a subset of paths on the evidence flow network. This means that paths of evidence that potentially contribute risk of bias are missed. Furthermore, the paths identified by the algorithm are not unambiguous and instead depend on the order in which certain steps are carried out. In contrast, the random-walk approach identifies all possible paths of evidence and delivers an unambiguous analytical result for proportion contributions. In addition, unlike the method in Papakonstantinou et al, the random-walk approach is able to handle networks with multi-arm trials.

The remainder of this paper is set out as follows: We present a motivating data set in Section 4.2. In Section 4.3 we provide the relevant background information. We describe an aggregate-level frequentist NMA model and show how the associated hat matrix can be interpreted as evidence flow. In Section 4.4 we introduce the analogies between NMA, electrical networks and random walks. Using the analogies to electrical networks in both the NMA and random-walk literature, we then express the flow of evidence in an NMA in terms of properties of random walks on the aggregate network. In Section 4.5 we introduce a second random-walk model, now on the directed evidence flow network. We use this to analytically derive the matrix of proportion contributions. In Section 4.6, we apply our method to the motivating data set and demonstrate that the random-walk approach overcomes the limitations of the numerical algorithm previously proposed by Papakonstantinou et al (2018) [9]. We summarise our results in Section 4.7 and discuss potential future impact of the analogy between NMA and random walks.

4.2 Motivating Example

We use an NMA of psychological treatments for patients with depressive disorders [38] to motivate our work. The data is described in detail in Rücker and Schwarzer (2014) [39]. For convenience we will occasionally refer to this as the 'depression data set'. The NMA compares $N = 11$ treatments based on $M = 26$ randomised controlled trials. Of

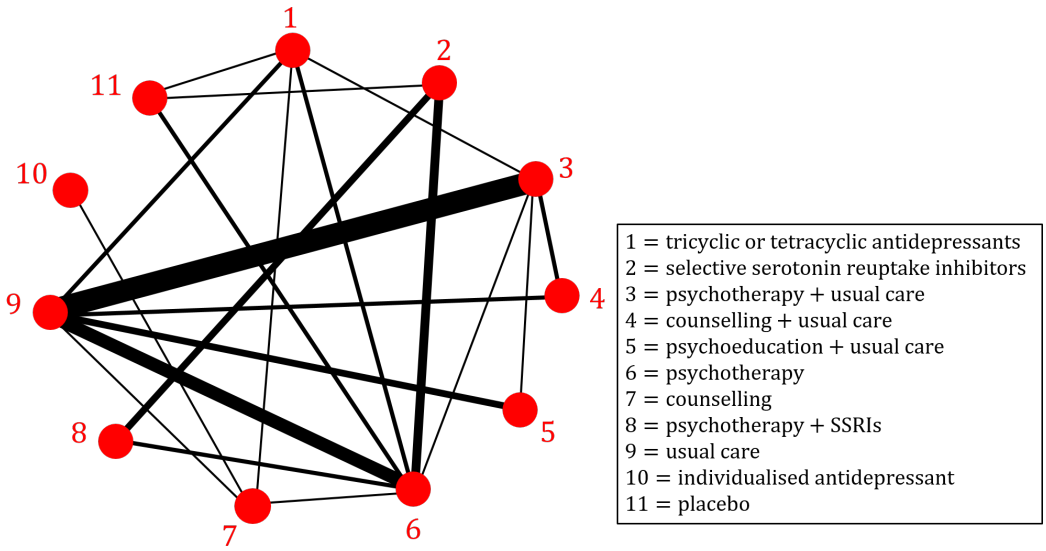


Figure 4.1: A network of psychological treatments for depression (original data from Linde et al (2013) [38]; presented in Rücker and Schwarzer (2014) [39]). We use numerical labels from 1 to 11, these are the same as in Rücker and Schwarzer (2014) [39]. Two treatments are connected by an edge if a direct comparison of the two treatments was made in at least one trial; the edge width indicates the number of trials that make the comparison. The network contains one 4-arm trial (comparing treatments 1-6-7-9), eight 3-arm trials (3-5-9, 2-6-8, 1-6-11, 1-3-9, 2-6-11, 2-6-8, 3-6-9, and 3-4-9) and 17 2-arm trials. Multi-arm trials are not explicitly indicated in the network graph. The data, including the number of trials per comparison, is described in detail in Rücker and Schwarzer (2014) [39].

these one is a four-arm trial, eight are three-arm trials and 17 contain just two arms. In total, the trials provide $K = 20$ pairs of treatments which are directly compared in at least one trial. The primary outcome of the trials was a binary variable representing patient response after treatment completion. The odds ratio (OR) was used as the measure of relative treatment effect. The graph representing this set of treatments and trials is shown in Figure 4.1. Vertices in the graph are treatments, and edges represent comparisons between pairs of treatments (two vertices are connected if they have been directly compared in at least one trial). The graph therefore has $N = 11$ vertices and $K = 20$ edges. The thickness of the edges in the figure represent the number of trials making the different comparisons.

NMA aims at estimating treatment effects for all pairs of interventions within this network. One aim of our paper is to determine the contribution (as a proportion) of each direct comparison to these estimates.

4.3 Network meta-analysis model

4.3.1 Definitions and notation

Among the multiple equivalent frequentist formulations of NMA [6, 11, 13, 39, 40] we choose a so-called 'aggregate level' (or two-step) approach [11] to the graph theoretical model developed in Rücker (2012) [7]. Rücker's original (one-step) model is implemented in the R package `netmeta` [25]. In Section 4.8.1 of the Appendix we outline how the aggregate-level graph theoretical approach relates to other frequentist NMA models.

We consider a network of N treatments, denoted $a = 1, \ldots, N$, and M studies, $i = 1, \ldots, M$. Throughout this article we will use the lower case letters a, b, c and d to refer to treatment nodes. Occasionally we also use x and y as dummy indices referring to nodes in sums or products. Study i compares a subset of n_i treatments (i.e. n_i is the number of treatments in trial i). We use a random-effects model where we focus on relative, rather than absolute effects. To this end, we write $Y_{i,ab}$ for the observed effect of treatment b in trial i relative to treatment a. We denote the variance associated with this observation by $\sigma^2_{i,ab}$. The heterogeneity, τ^2, in the network can be estimated, for example, using the method-of-moments approach [41]. The estimated heterogeneity is added to the within-trial variance estimate from each study to make the total variance $\sigma^2_{i,ab} + \tau^2$.

Trial i contributes $q_i = n_i(n_i - 1)/2$ observed relative treatment effects and associated variances. For a trial with $n_i = 2$, comparing treatments a and b, the weight assigned is given by the inverse variance, $w_{i,ab} = 1/(\sigma^2_{i,ab} + \tau^2)$. In order to account for correlations induced by multi-arm trials ($n_i \geq 3$), we use an adjustment method described in detail in References [7, 39, 42]. The method involves adjusting the variances associated with each pairwise comparison in a multi-arm trial. For multi-arm trial i this results in $q_i \geq 3$ weights, $w_{i,ab}$, where a and b run through all treatments compared in that trial. This defines a complete sub-graph of q_i two-arm trials which is equivalent to the multi-arm trial.

4.3.2 Aggregate-level description

The set of adjusted weights $\{w_{i,ab}\}$ for all trials $i = 1, \ldots, M$ defines a network of $\sum_{i=1}^{M} q_i$ two-arm trials. This network is equivalent to the original network of M (potentially

multi-arm) trials in that the resulting relative treatment effect estimates from the network of two-arm trials described by $\{w_{i,ab}\}$ are the same as those obtained from the original network [39].

We write M_{ab} for the set of trials $i \in \{1, \ldots, M\}$ comparing treatments a and b. Using the weights $\{w_{i,ab}\}$, we perform a pairwise meta-analysis across each of the K edges in the network. For the edge connecting nodes a and b, the direct estimate is calculated as the weighted mean,

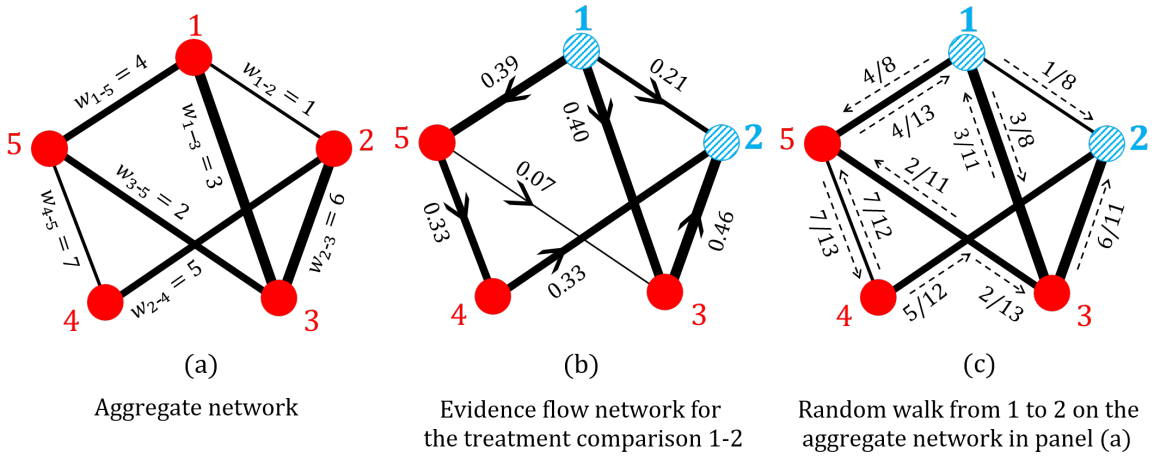
$$\hat{\theta}_{ab}^{\text{dir}} = \frac{\sum_{i \in M_{ab}} w_{i,ab} Y_{i,ab}}{\sum_{i \in M_{ab}} w_{i,ab}}. \tag{4.1}$$

This results in K *direct* estimates of the relative treatment effects, $\hat{\theta}_{ab}^{\text{dir}}$, which we collect in the vector $\hat{\boldsymbol{\theta}}^{\text{dir}}$. The weight associated with the direct estimate $\hat{\theta}_{ab}^{\text{dir}}$ (and to be used in the subsequent analysis) is given by

$$w_{ab} = \sum_{i \in M_{ab}} w_{i,ab}. \tag{4.2}$$

The direct estimates of the relative treatment effects have been termed 'aggregate' data [8, 43]. Therefore, Equations (4.1) and (4.2) describe the observations and inverse-variance weights for an aggregate-level model.

The aggregate model can be represented by an 'aggregate network' where w_{ab} is the weight associated with the edge ab. We collect the aggregate edge weights in a $K \times K$ diagonal matrix, $\boldsymbol{W} = \text{diag}(w_{ab})$. Figure 4.2 (a) shows a fictional example of an aggregate network with five treatments $a = 1, 2, 3, 4, 5$. The aggregate weight matrix for this example is $\boldsymbol{W} = \text{diag}(1, 3, 4, 6, 5, 2, 7)$. We write $\boldsymbol{B}$ for the $K \times N$ edge-incidence matrix of the aggregate network. Each column of $\boldsymbol{B}$ corresponds to a treatment in the network and each row corresponds to an edge. To construct the matrix, one of the two treatments in each edge is designated as the 'baseline treatment' for this edge without loss of generality. Entries are $+1$ in the column corresponding to the 'baseline' treatment of the comparison represented by that row, and -1 in the column corresponding to the treatment compared to that baseline. For the example in



(a) Aggregate network

(b) Evidence flow network for the treatment comparison 1-2

(c) Random walk from 1 to 2 on the aggregate network in panel (a)

Figure 4.2: (a) A fictional example of an aggregate meta-analytic network with edges weighted and labelled by their respective (inverse-variance) weights. (b) The resulting evidence flow network for the comparison 1-2 from the aggregate network in (a); the comparison 1-2 is indicated by shading these nodes with blue stripes. Edges are directed according to the sign of the corresponding element of the hat matrix, and are weighted by the absolute value of the hat matrix element. (c) The random walk on the aggregate network in (a) for a walker starting at node 1 and finishing at node 2; edges are labelled by the associated transition probabilities.

Figure 4.2 (a) the edge incidence matrix can be chosen as

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \tag{4.3}$$

where the columns represent treatments 1, 2, 3, 4, and 5, and the rows represent the edges (direct comparisons) 1-2, 1-3, 1-5, 2-3, 2-4, 3-5, and 4-5. In the following we will use this hyphenated notation when we refer to specific comparisons (e.g. 1-2 for the comparison between treatments 1 and 2). When we refer to a comparison between unspecified treatments a and b, we will then use the notation ab, to avoid confusion with 'a minus b'.

4.3.3 Hat matrix and network estimates

The *network* estimates of the relative treatment effects $\hat{\theta}_{ab}^{\text{net}}$ are obtained via

$$\hat{\boldsymbol{\theta}}^{\text{net}} = \boldsymbol{H}\hat{\boldsymbol{\theta}}^{\text{dir}}, \tag{4.4}$$

where the hat matrix associated with the aggregate model is [39]

$$\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}^{\top}\boldsymbol{W}\boldsymbol{B})^{+}\boldsymbol{B}^{\top}\boldsymbol{W}. \tag{4.5}$$

The hat matrix has dimension $K \times K$ where each row and each column correspond to one edge. We denote the element in the ab row and cd column by $H_{cd}^{(ab)}$. The matrix $\boldsymbol{L} = \boldsymbol{B}^{\top}\boldsymbol{W}\boldsymbol{B}$, with dimensions $N \times N$ and rank $N-1$, is the Laplacian of the aggregate network. The matrix $\boldsymbol{L}^{+} = (\boldsymbol{B}^{\top}\boldsymbol{W}\boldsymbol{B})^{+}$ is its pseudo-inverse [7, 42]. The hat matrix describes how the direct evidence combines to give the network estimates. Each network estimate is a weighted linear combination of direct and indirect evidence. The coefficients of the estimates $\hat{\boldsymbol{\theta}}^{\text{dir}}$ for each network treatment effect are found in the corresponding row of $\boldsymbol{H}$. The diagonal elements of $\boldsymbol{H}$ give the coefficients for the direct evidence while the off-diagonal elements indicate the contribution of indirect evidence. The larger the diagonal elements, the more weight is given to direct evidence [8]. For the example in Figure 4.2 we calculate $\boldsymbol{H}$ using Equation (4.5), $\boldsymbol{W} = \text{diag}(1, 3, 4, 6, 5, 2, 7)$ and $\boldsymbol{B}$ given in Equation (4.3). The resulting hat matrix is quoted in Equation (4.32) in the Appendix.

4.3.4 Evidence flow

König et al (2013) [8] noted that each row in the hat matrix can be interpreted as a flow network. Focusing on one row of the hat matrix, the magnitude of the flow of evidence between two nodes is given by the absolute value of the element in the corresponding column of $\boldsymbol{H}$. The direction is determined by the sign of the element of the hat matrix. For the ab-row of the hat matrix one defines evidence flows $f_{cd}^{(ab)}$ (from c to d) and $f_{dc}^{(ab)}$ (from d to c) as follows [8]:

$$\begin{aligned} &\text{if } H_{cd}^{(ab)} > 0: \quad f_{cd}^{(ab)} = H_{cd}^{(ab)}, \quad f_{dc}^{(ab)} = 0, \\ &\text{if } H_{cd}^{(ab)} < 0: \quad f_{cd}^{(ab)} = 0, \qquad f_{dc}^{(ab)} = |H_{cd}^{(ab)}|. \end{aligned} \tag{4.6}$$

Flows are non-negative, and only one of $f_{cd}^{(ab)}$ and $f_{dc}^{(ab)}$ is non-zero. We note that setting one of the coefficients for each pair of nodes to zero is a choice. Alternatively, one could have chosen conventions such that $f_{cd}^{(ab)} = -f_{dc}^{(ab)}$ for all pairs c and d. This is an equivalent re-parameterisation, but is less convenient for our subsequent workings. We comment further on this in Section 4.10 of the Appendix.

It is important to note that each comparison ab gives rise to a separate network of flows. We refer to these graphs as 'evidence flow networks'. The edges of these graphs are directional and point in the direction of positive flow. Due to the properties of the hat matrix each of these evidence flow networks is acyclic. Specifically, in the network corresponding to the comparison ab, node a only has outgoing edges, and node b only incoming edges. The flow network then has the following properties:

1. The total outflow from a is equal to one, $\sum_x f_{ax}^{(ab)} = 1$;
2. the sum of inflows to node b is also one, $\sum_x f_{xb}^{(ab)} = 1$;
3. and at every intermediate node, $c \neq a, b$, the sum of outflows equals the sum of inflows, $\sum_x f_{cx}^{(ab)} = \sum_x f_{xc}^{(ab)}$.

These properties were stated in König et al (2013) [8], and an algebraic proof for the first and the second property was given in Papakonstantinou et al (2018) [9]. We provide a heuristic argument for all three properties in Section 4.11 of the Appendix. These three properties make an interpretation as a 'flow' natural. We adopt the term 'flow of evidence' used in previous literature [8, 9], noting that it is perhaps not immediately clear what precisely 'evidence' is mathematically, and how it can flow from one node to another. The random-walk picture we develop later in this paper offers a possible interpretation, which we will discuss in Section 4.4.3 and in the Appendix (Section 4.10).

Figure 4.2 (b) shows the evidence flow network for the comparison 1-2 for the aggregate network in Figure 4.2 (a). The values of flow shown in Figure 4.2 (b) correspond to the first row of the matrix $\boldsymbol{H}$ given in Equation (4.32) in the Appendix.

4.4 NMA, electrical networks and random walks

In this section we set up the analogies between NMA, electrical networks and random walks. A summary of these analogies can be found in Table 4.1.

4.4.1 NMA and electrical networks

The connection between meta-analytic and electrical networks was first introduced by one of us [7]. In the meta-analytic network, treatments are nodes connected by edges representing pairwise comparisons. On the other hand, edges in an electrical network represent resistors that connect at the nodes. If two (or more) nodes of an electric network are connected to the poles of a battery then an electric potential (a real-valued scalar quantity) can be associated with each node in the network. The potential in turn results in voltages (=differences in electric potential) across all edges. The potential difference between two nodes connected by a path on the graph is the sum of the voltages along each edge of the path. If there are multiple paths connecting two nodes then the sum of voltages is independent of the path. Voltages along a cycle on the network sum to zero.

The voltages in turn induce currents across the edges (current=voltage divided by resistance). Currents may also flow into or out of a node from or to the external battery (often referred to simply as the 'exterior'). The sum of currents entering each node equals the sum of currents leaving that node (Kirchhoff's current law, see for example Urbano (2019) [44]).

The analogy between NMA and electric networks is based on the observation that resistances in parallel and sequential electrical circuits combine in the same way as variances of treatment effects in an NMA. Variance therefore corresponds to resistance. One can show that relative treatment effects are the analogue of voltages measured across edges, and weighted treatment effects the analogue of electrical current (see Rücker (2012) [7] for details). This allows one to use graph theoretical tools, routinely applied to electrical networks, to address questions in NMA.

In Rücker (2012) [7] no voltages or external currents are applied directly to the electric circuit representing the NMA network (i.e. there is no external battery). Instead, the starting point is given by measurements of treatment effects (voltages) across the

Table 4.1: Summary of the analogy between NMA, electrical networks and random walks (RW) on the aggregate network.

NMA	Electric circuit	RW on the aggregate network
Treatments $1, 2, \ldots, N$	Nodes $1, 2, \ldots, N$	Nodes $1, 2, \ldots, N$
Direct treatment comparisons	Edges (conducting wires)	Edges (along which a random walker can travel in both directions)
w_{ab} inverse-variance weight associated with edge ab on the aggregate network	$C_{ab} = R_{ab}^{-1}$ conductance (inverse resistance) across edge ab	$T_{ab} = C_{ab} / \sum_{c \neq a} C_{ac} = w_{ab} / \sum_{c \neq a} w_{ac}$ probability that a walker at node a hops to node b in the next step
The aggregate hat matrix element $H_{cd}^{(ab)}$ that defines the flow of evidence through the direct comparison cd for the network treatment effect ab	Flow of current through edge cd when a battery is attached across nodes a and b such that a unit current flows into a and out of b	Expected net number of times a walker starting at a and ending at b crosses the edge from c to d

edges of the network. These are understood to be the true treatment effects subject to some random additive error. It is then shown in Rücker (2012) [7] that finding the NMA estimates of treatment effects corresponds to finding the set of consistent voltages across all edges that minimises the (Euclidian) distance to these observed treatment effects.

Here, we extend this analogy and show that the elements of the hat matrix have an interpretation in the electric-circuit picture. More precisely, the elements of the row in the hat matrix corresponding to the comparison between treatments a and b can be obtained as follows: Connect a battery to nodes a and b in the electric circuit so that one unit of current flows from the exterior into node a, and out of the network (to the exterior) from node b. The external currents into/out of all other nodes are maintained at zero. This reflects the properties 1-3 of the coefficients $f_{cd}^{(ab)}$ in Section 4.3.4. These derive from the properties of the hat matrix via Equation (4.6), which in turn are a consequence of the function of the hat matrix to project onto the space of consistent relative treatment effects (see Rücker (2012) [7]). This set-up induces currents across the edges in the network. Our main result is then the following: The current along edge cd is identical to the hat matrix element $H_{cd}^{(ab)}$. A detailed mathematical proof can be found in Section 4.12 of the Appendix.

We illustrate this with a simple network of four nodes in Figure 4.3. Panel (a) shows a generic electrical circuit resulting from a meta-analytic graph with four treatment options and with direct comparisons between all pairs of treatments except treatments 1 and 4. We focus on the row in the hat matrix corresponding to the comparison between treatments 1 and 2. Using Equation (4.4) we have for this example

$$\hat{\theta}_{1\text{-}2}^{\text{net}} = H_{1\text{-}2}^{(1\text{-}2)}\hat{\theta}_{1\text{-}2}^{\text{dir}} + H_{1\text{-}3}^{(1\text{-}2)}\hat{\theta}_{1\text{-}3}^{\text{dir}} + H_{2\text{-}3}^{(1\text{-}2)}\hat{\theta}_{2\text{-}3}^{\text{dir}} + H_{2\text{-}4}^{(1\text{-}2)}\hat{\theta}_{2\text{-}4}^{\text{dir}} + H_{3\text{-}4}^{(1\text{-}2)}\hat{\theta}_{3\text{-}4}^{\text{dir}}. \tag{4.7}$$

Our result indicates that the coefficients $H_{cd}^{(1\text{-}2)}$ can be obtained from the setup shown in Figure 4.3 (b). A battery is attached to nodes 1 and 2 and the voltage of the battery is chosen such that one unit of current flows into node 1 (from the battery) and out of node 2 (into the battery). This induces currents in the five edges (resistors) of the electric circuit. These currents are the hat matrix elements in Equation (4.7). Via Equation (4.6) these then determine the flow of evidence.

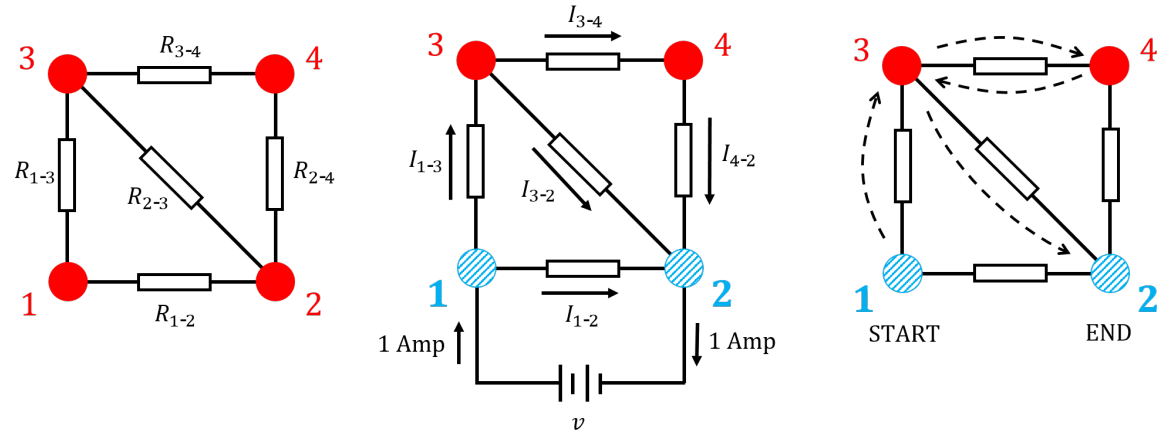


Figure 4.3: An illustration of the interpretation of current. (a) An electrical network with associated edge resistances. (b) The same network with a battery attached across the edge 1-2 such that a unit current flows into 1 and out of 2. The current in edge cd is labelled I_{cd}. Current is measured in ampéres, hence the unit current is labelled as '1 Amp'. The direction of the current induced in the edges is shown. (c) A possible path taken by a random walker starting at node 1 and stopping at node 2. The sequence of nodes visited is $1 \to 3 \to 4 \to 3 \to 2$. For this particular realisation of the random walk, the net number of times the walker crosses edges 1-3 and 3-2 is one, while all other edges are crossed net zero times. The expected net number of times the walker crosses an edge is given by the currents shown in (b) for that edge [37]. The focus on the comparison of nodes 1 and 2 in panels (b) and (c) is indicated by the blue striped pattern of these nodes.

4.4.2 Electrical networks and random walks

4.4.2.1 Definitions and notation

As illustrated in Figure 4.4, a random walk on a graph is a stochastic process consisting of a succession of 'hops' between neighbouring nodes (nodes connected by an edge). We use the word 'path' to describe the sequence of nodes visited by the walker, including repeat visits to individual nodes. We always assume that time is discrete. The walk is then a Markov process described by an $N \times N$ transition matrix, $\boldsymbol{T}$, where N is the number of nodes in the network. The element T_{ab} of this matrix is the probability that a walker, currently at node a, moves to node b in the next time step. These probabilities only depend on the current position of the walker, and not on the path taken to reach that position. One has $\sum_b T_{ab} = 1$ for all a, i.e. $\boldsymbol{T}$ is a stochastic matrix.

The connection between random walks and electrical networks has been recognised for some time [34–36] and is described extensively in Doyle and Snell (2000) [37]. Here we will only summarise the concepts and known results that are most relevant for our

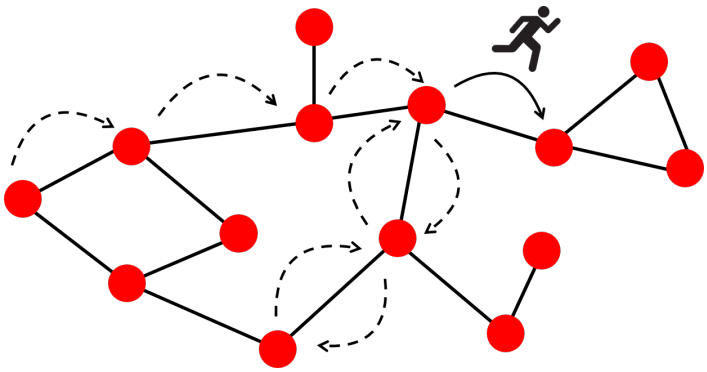


Figure 4.4: An illustration of a random walker moving on a network graph. The walker starts its journey from the far left node. The arrows show the path taken by the walker for one realisation of the random walk. The figure indicates the 'current' position of the walker as it hops between two nodes. The solid arrow indicates this transition. The dotted arrows indicate the previous transitions made between nodes by the walker.

work.

Starting from an electrical network with given resistances R_{ab} a random walk process can be constructed by defining the transition probabilities ($a \neq b$)

$$T_{ab} = \frac{R_{ab}^{-1}}{\sum_{c \neq a} R_{ac}^{-1}}. \tag{4.8}$$

This definition indicates that transitions from one node to another occur in proportion to the inverse resistance of the direct connection between the two nodes (if there is no direct connection, then no hop can occur between the two nodes). We set $T_{aa} = 0$ for all a. The denominator in Equation (4.8) ensures normalisation ($\sum_b T_{ab} = 1$).

We always assume the network does not divide into multiple disconnected components. As a result, the transition matrix defined in Equation (4.8) is such that a walker starting at any node a will eventually reach any other node $b \neq a$ with finite probability.

4.4.2.2 Interpretation of electrical current

Electrical current can be interpreted in the random-walk picture as follows [37]: When a voltage is applied between two nodes a and b such that the total current flowing into a and out of b from the exterior is 1, the current induced in each edge, cd, is equal to the expectation value for the *net* number of times a random walker, starting at a and walking until it reaches b, moves along the edge from c to d. The net number of times the walker moves from c to d is the number of crossings in the direction from c to d minus the number of crossings in the opposite direction.

To describe this mathematically we need to ensure that no more hops occur when the walker reaches the designated end point b. In other words, this node must become absorbing. This is achieved by setting the elements T_{bc} to zero for all c. For later convenience we denote the resulting modified transition matrix by $\boldsymbol{T}^{(ab)}$, recognising that the modifications made only depend on the choice of b, and not on a. Mathematically, we have $T^{(ab)}_{bc} = 0$ for all $c \neq b$, and $T^{(ab)}_{cd} = T_{cd}$ for $c \neq b$ and all d. We set $T^{(ab)}_{bb}$ to unity.

Now consider random walks starting at node a and then following the process defined by the transition matrix $\boldsymbol{T}^{(ab)}$. All walks therefore end at node b. The probability that a walker takes a particular path π connecting a and b can be written as

$$P^{(ab)}(\pi) = \prod_{\{xy \in \pi\}} T^{(ab)}_{xy}, \tag{4.9}$$

where the notation $\{xy \in \pi\}$ indicates the set of pairs of successive nodes in the path π. We note that $P^{(ab)}(\pi)$ is non-zero if and only if the path π starts at a and ends when b is reached for the first time.

The average number of net crossings from node c to node d along paths starting at a and ending at b can therefore be obtained as

$$\overline{N^{(ab)}_{cd}} = \sum_{\pi} P^{(ab)}(\pi) N_{cd}(\pi), \tag{4.10}$$

where $N_{cd}(\pi)$ is the net number of crossings from c to d along path π. We note that this quantity can be negative; this occurs if the walker makes more transitions from d to c than from c to d. The sum in Equation (4.10) extends over all paths connecting a and b.

To develop some intuition, consider again the electrical network in Figure 4.3 (a). Assume that we are interested in the scenario where the external current flows into node 1 and out of node 2, but not into or out of any of the other nodes. We then start the random-walk process at node 1, and use transition probabilities as defined in Equation (4.8) until the walker reaches node 2. In the first step, the walker either hops to node 2 (this occurs with probability $T_{1\text{-}2}$) or to node 3 (with probability $T_{1\text{-}3}$). If the walker hops to node 2, the walk stops and the path taken by the walker is $1 \to 2$. Otherwise, the walker is at node 3 and in the next step it can transition to 2, 4 or back to 1 with respective probabilities $T_{3\text{-}2}$, $T_{3\text{-}4}$ and $T_{3\text{-}1}$. This process continues until the

walker eventually reaches node 2. The current through the edge cd is then given by the expected net number of times such a walker crosses the edge from c to d before it arrives at node 2. A crossing in the direction from d to c contributes negatively to this value.

Since the random walker can move in both directions along the network edges, there are infinitely many paths the walker can take as it travels from node 1 to node 2 in this example. Figure 4.3 (c) shows one possible path, $1 \to 3 \to 4 \to 3 \to 2$. The probability the random walker takes this path is given by the product of the individual transition probabilities along the path, that is

$$P^{(1\text{-}2)}(1 \to 3 \to 4 \to 3 \to 2) = T_{1\text{-}3}T_{3\text{-}4}T_{4\text{-}3}T_{3\text{-}2}. \tag{4.11}$$

Although $P^{(ab)}(\pi)$ can be obtained relatively easily for each path π, carrying out the sum in Equation (4.10) by exhaustive enumeration of all relevant paths is not practicable. This is because there are generally infinitely many paths starting and ending at the designated nodes (due to the possibility to hop back to nodes visited earlier).

The analogy between electrical circuits and random walks [37] however can be used to calculate the expected number of net crossings through an edge analytically. This is detailed in Sections 4.13 and 4.14 of the Appendix, see in particular Equation (4.72).

The expected number of net crossings can also be obtained from simulations of the random-walk process. An ensemble of walkers is released at the starting point a. Each walker then independently hops from node to node on the network with transition rates as in Equation (4.8) until it hits the designated endpoint (node b). The process then stops. For each walker the net number of crossings from c to d can be recorded, and this is then averaged over the ensemble of walkers.

4.4.3 Random walk on a meta-analytic network

As described above, conductance (inverse resistance) in an electrical network has an analogue in terms of both NMA, and random walks. Exploiting these analogies, we now define a random-walk process on a meta-analytic network via the transition rates

$$T_{ab} = \frac{w_{ab}}{\sum_{c \neq a} w_{ac}}, \tag{4.12}$$

with weights w_{ab} associated with the edges as discussed in Section 4.3.2, see in particular Equation (4.2).

In order to study walks starting at node a and ending at b we use the matrix $\boldsymbol{T}^{(ab)}$ as defined in Section 4.4.2.2. This enforces absorption of the walker at node b when this node is reached. For the example aggregate network in Figure 4.2 (a), the transition matrix for a random walk starting at node 1 and ending at node 2 is

$$\boldsymbol{T}^{(1\text{-}2)} = \begin{pmatrix} 0 & 1/8 & 3/8 & 0 & 4/8 \\ 0 & 1 & 0 & 0 & 0 \\ 3/11 & 6/11 & 0 & 0 & 2/11 \\ 0 & 5/12 & 0 & 0 & 7/12 \\ 4/13 & 0 & 2/13 & 7/13 & 0 \end{pmatrix}. \tag{4.13}$$

Each row and column of $\boldsymbol{T}^{(1\text{-}2)}$ represents a treatment in the network, $a = 1, 2, 3, 4, 5$. Given that we focus on the comparison between treatments 1 and 2, node 1 is the start point of the walk, and node 2 is absorbing. Therefore, the row corresponding to treatment 2 contains only zeroes except for the diagonal element which is equal to one (when the walker reaches node 2 it stays there indefinitely). The entries in each row of the matrix in Equation (4.13) sum to one. The diagonal elements of $\boldsymbol{T}^{(1\text{-}2)}$ (except for the element relating to node 2) are zero. This indicates that, with the exception of the absorbing state, the random walker cannot stay at the same place at any step. Figure 4.2 (c) illustrates the dynamics of the random walk from node 1 to node 2 for this example.

In Section 4.4.1 we made the connection between the flow of electric current and the flow of evidence in an NMA. Using the interpretation of current as a random walk we can now establish the following analogy: For the comparison of treatments a and b, the hat matrix element $H_{cd}^{(ab)}$ that defines the flow of evidence through the direct comparison cd is equal to the expected *net* number of times a random walker starting at node a on the aggregate NMA network moves along the edge from c to d before it reaches node b. In other words, we equate

$$H_{cd}^{(ab)} = \overline{N_{cd}^{(ab)}}, \tag{4.14}$$

and define the flow of evidence $f_{cd}^{(ab)}$ in terms of $H_{cd}^{(ab)}$ via Equation (4.6).

The random-walk picture that we have developed provides a possible interpretation for the concept of 'flow of evidence'. Namely, it is random walkers starting at a and ending at b that 'flow' along the network based on the rules defined by the transition rates T_{cd}. A more detailed discussion of this interpretation can be found in Section 4.10 in the Appendix.

In summary, we have used existing analogies between electric circuits and random walks on the one hand, and network meta-analysis and electric circuits on the other to introduce an interpretation of the flow of evidence in network meta-analysis in terms of random walks. The analogies between all three areas are highlighted in Table 4.1.

4.5 Proportion contribution

In this section we present a random-walk interpretation and construction of the so-called 'proportion contribution matrix' [9]. As explained in more detail below, the definition of these proportion contributions originates from the fact that we can interpret each row of the hat matrix as a flow network [8, 9]. For this task, therefore, the random walk now no longer takes place on the meta-analytic network. Instead, walkers move on the evidence flow network. The entries of the proportion contribution matrix in NMA can then be obtained from this random walk.

We show that the random-walk approach overcomes the limitations of the algorithm proposed for the evaluation of proportion contributions in Papakonstantinou et al (2018) [9]. In particular, it provides an analytical expression for proportion contributions that removes ambiguity associated with the selection of paths. Furthermore, unlike the numerical algorithm of Papakonstantinou et al [9], the random-walk approach identifies all paths of evidence so that all potential sources of bias are taken into account. In Section 4.5.1 we introduce the concept of proportion contributions. In 4.5.2 we describe the algorithm in Papakonstantinou et al [9] and its limitations. We then present and discuss the random-walk approach in Section 4.5.3.

4.5.1 Background and definition

In NMA it is important to assess the influence of individual study bias on the estimates obtained from the network. To this end, the CINeMA framework and software [10, 45]

provides a user friendly system to assess confidence in the results from an NMA. One function of the software is to display the relative influence of evidence that comes from studies with high, moderate and low risk of bias on each network treatment effect. This assessment involves calculating the matrix of so-called 'proportion contributions' [9]. This matrix describes how much each direct treatment effect contributes to each network treatment effect as a relative *proportion*. The idea of the proportion contribution matrix is based on the hat matrix. The elements of the hat matrix are the coefficients of the linear relation between network estimates and direct estimates in the NMA as described in Equation (4.4). These coefficients can be positive or negative. The proportion contribution matrix uses the properties of the hat matrix and translates the elements of $\boldsymbol{H}$ to positive proportion contributions, where the total contribution is normalised to one. We now explain this in more detail using the work of Papakonstantinou et al (2018) [9].

Consider the example network in Figure 4.5 (a). This relates to an NMA of the four topical antibiotics given in the figure caption for the treatment of chronically discharging ears [46]. To keep the text concise we label the treatments 1, 2, 3 and 4. In accordance with Equation (4.4), the network estimate of comparison 1-2 is given by the linear equation (4.7), which we repeat here for clarity,

$$\hat{\theta}^{\text{net}}_{\text{1-2}} = H^{(\text{1-2})}_{\text{1-2}}\hat{\theta}^{\text{dir}}_{\text{1-2}} + H^{(\text{1-2})}_{\text{1-3}}\hat{\theta}^{\text{dir}}_{\text{1-3}} + H^{(\text{1-2})}_{\text{2-3}}\hat{\theta}^{\text{dir}}_{\text{2-3}} + H^{(\text{1-2})}_{\text{2-4}}\hat{\theta}^{\text{dir}}_{\text{2-4}} + H^{(\text{1-2})}_{\text{3-4}}\hat{\theta}^{\text{dir}}_{\text{3-4}}. \tag{4.15}$$

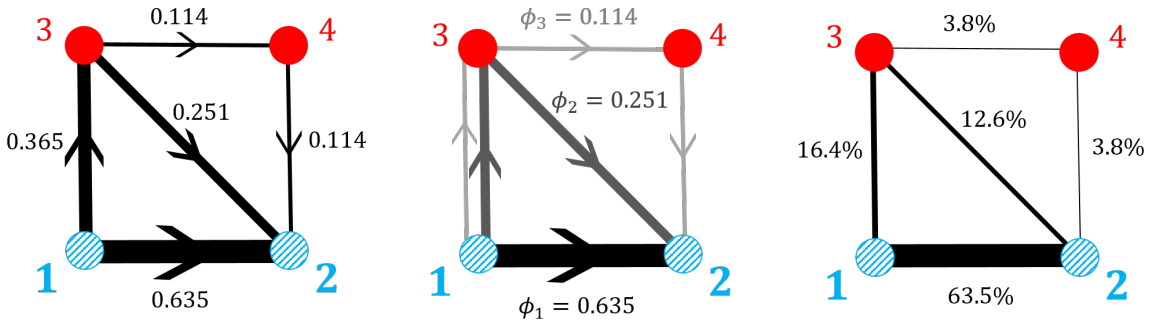
We can think of the expression on the right-hand side as a combination of different direct and indirect estimates of $\theta_{\text{1-2}}$. The direct estimate is simply $\hat{\theta}^{\text{dir}}_{\text{1-2}}$. We obtain one indirect estimate using node 3 and the consistency equation,

$$\hat{\theta}^{\text{ind}(1)}_{\text{1-2}} = \hat{\theta}^{\text{dir}}_{\text{1-3}} - \hat{\theta}^{\text{dir}}_{\text{2-3}}. \tag{4.16}$$

A second indirect estimate is found via nodes 3 and 4,

$$\hat{\theta}^{\text{ind}(2)}_{\text{1-2}} = \hat{\theta}^{\text{dir}}_{\text{1-3}} - (\hat{\theta}^{\text{dir}}_{\text{2-4}} - \hat{\theta}^{\text{dir}}_{\text{3-4}}). \tag{4.17}$$

These three ways of estimating $\theta_{\text{1-2}}$ correspond to so-called 'paths of evidence' on the evidence flow network [9]. We label these paths π_i $(i = 1, 2, 3)$. As illustrated in Figure 4.5 (b), these are $\pi_1 = 1 \to 2$, $\pi_2 = 1 \to 3 \to 2$, and $\pi_3 = 1 \to 3 \to 4 \to 2$. We can now write the network estimate $\hat{\theta}^{\text{net}}_{\text{1-2}}$ as a linear combination of the estimates $\hat{\theta}^{\text{dir}}_{\text{1-2}}$,



(a) Evidence flow network for comparison 1-2

(b) Evidence streams for comparison 1-2 (paths taken by a random walker)

(c) Proportion contributions for the network effect 1-2

Figure 4.5: Illustration of evidence flow, streams of evidence and proportion contributions for a network of topical antibiotics without steroids for chronically discharging ears presented in Macfadyen (2005) [46]. Node 1 is no treatment; 2 is quinolone antibiotic; 3 is antiseptic; and 4 is non-quinolone antibiotic. (a) The evidence flow network for comparison 1-2, based on Figure 1, panel (b) in Papakonstantinou et al (2018) [9]. The edge labels are the entries of the 1-2 row of the hat matrix, their signs are associated with the direction of the arrows. (b) The decomposition of edge flows into flow through paths of evidence as estimated by the algorithm in Papakonstantinou et al. The paths of evidence shown are equivalent to the possible paths taken by a random walker on the evidence flow network. (c) The proportion contributions (expressed as percentages) of each direct treatment effect to the network estimate of the 1-2 relative treatment effect.

$\hat{\theta}^{\text{ind}(1)}_{1\text{-}2}$, and $\hat{\theta}^{\text{ind}(2)}_{1\text{-}2}$. That is,

$$\begin{aligned}\hat{\theta}^{\text{net}}_{1\text{-}2} &= \phi_1\hat{\theta}^{\text{dir}}_{1\text{-}2} + \phi_2\hat{\theta}^{\text{ind}(1)}_{1\text{-}2} + \phi_3\hat{\theta}^{\text{ind}(2)}_{1\text{-}2} \\ &= \phi_1\hat{\theta}^{\text{dir}}_{1\text{-}2} + \phi_2(\hat{\theta}^{\text{dir}}_{1\text{-}3} - \hat{\theta}^{\text{dir}}_{2\text{-}3}) + \phi_3(\hat{\theta}^{\text{dir}}_{1\text{-}3} - \hat{\theta}^{\text{dir}}_{2\text{-}4} + \hat{\theta}^{\text{dir}}_{3\text{-}4}).\end{aligned} \tag{4.18}$$

The coefficients, ϕ_i, define the flow of evidence through each path π_i, see Papakonstantinou et al (2018) [9].

Figure 4.5 (b) shows how the flows in each edge, described by the hat matrix coefficients, are deconstructed into the flows through each path of evidence, described by the coefficients ϕ_i. In this example, only the edge 1-3 is used for more than one path. When calculating the flow through each path, the flow in edge 1-3 is 'split' between the paths $\pi_2 = 1 \to 3 \to 2$, and $\pi_3 = 1 \to 3 \to 4 \to 2$ according to the flow in the subsequent edges along those two paths.

A so-called 'stream' of evidence [9] is a pair consisting of a path and the flow associated with this path, $S_i = (\pi_i, \phi_i)$. The proportion contribution of each direct

comparison cd to the network estimate of each comparison ab, is then defined as [9]

$$p_{cd}^{(ab)} = \sum_{i:\ cd \in \pi_i} \frac{\phi_i}{|\pi_i|}, \tag{4.19}$$

where $|\pi_i|$ is the number of edges that make up the path. The sum extends over all paths in the evidence flow network for the comparison ab that contain the edge cd. We note that all such paths start at a and end at b, and, because the evidence flow network is acyclic, multiple visits to the same node do not occur.

For simple examples, such as the one in Figure 4.5, one can obtain the path flows ϕ_i by directly comparing coefficients in Equations (4.15) and (4.18). Using the properties of the hat matrix in Section 4.3.4 one can then also see that $\phi_i \geq 0$ for all i, and that $\sum_i \phi_i = 1$. This means that the proportion contributions in Equation (4.19) are also non-negative, and sum to one. Figure 4.5 (c) shows the proportion contributions, expressed as percentages, for the example in Figure 4.5 (a).

For larger, more connected networks it is not immediately clear how to obtain the ϕ_i. In particular, when there are more paths than edges, expressing the ϕ_i in terms of the coefficients of the hat matrix is non-trivial. Papakonstantinou et al [9] present an iterative algorithm to identify streams for a general evidence flow network. The starting point for this algorithm is the hat matrix. In the initial implementation of the algorithm, Papakonstantinou et al (2018) [9] used a hat matrix that did not account for correlations due to multi-arm trials (it treated each comparison in a multi-arm trial as an independent two-arm study). In this work, we instead implement the algorithm using the hat matrix of the aggregate model defined in Equation (4.5). We will now briefly describe the algorithm.

4.5.2 Existing iterative numerical algorithm to determine streams of evidence

Broadly speaking, each iteration of the algorithm consists of the following steps: (i) A path in the evidence flow network is selected. (ii) The minimum flow through the edges making up the path is identified. This is assigned as the flow associated with the path. (iii) The flow of the path is subtracted from the values of flow in the edges that make up that path. This means that the edge corresponding to the minimum flow in that path is removed from the graph. (iv) A new path is then selected from the

remaining graph. The process repeats until all the evidence flow in the edges has been assigned to a path.

Different methods for selecting the paths in step (i) give rise to multiple variants of the algorithm. For example, paths may be selected at random or in order from shortest to longest. We refer to these approaches as 'Random' and 'Shortest' respectively. The Shortest algorithm is implemented in the software `netmeta` using a breadth first search algorithm [47]. Each time the Random algorithm is run it selects the paths in a different order and, potentially, gives a different outcome. For reasons of reproducibility, this version of the algorithm is not implemented in current software. For simple networks such as the example in Figure 4.5, the order of selection does not affect the outcome. However, for more complicated networks this is not the case. In some graphs, the flow of evidence is fully exhausted before every possible path has been selected. The remaining paths can then not be associated with any flow. Critically, this approach means that many paths of evidence are not identified and their contribution (along with any potential bias) is not accounted for. The set of paths that are missed in this way can depend on the order in which paths are selected by the algorithm. Examples of this behaviour are presented in Supplementary File 3 in Papakonstantinou et al (2018) [9] and in Section 4.6.2 of this paper.

One potential remedy consists of averaging results from the Random algorithm by Papakonstantinou et al [9] over a large number of realisations. We call this method 'Average'. Provided enough realisations are generated, the Average algorithm will eventually identify every evidence path. However, because of the nature of the algorithm, the number of times a particular path is sampled by this method can depend on features of the network not directly related to the path. In step (iii) of the algorithm the edge associated with the smallest flow in a particular path is removed from the network. This means that any other path containing this edge can no longer be selected. As a result, paths that do not share edges with any other paths will be selected in every run of the algorithm, whereas paths which do share edges with other paths will be sampled less often. It is therefore not clear how to interpret average proportion contributions determined in this way. Furthermore, this approach is computationally intensive as it relies on repeating the (already iterative) algorithm many times. For this reason, this version of the algorithm is not implemented in current software.

To overcome these limitations, we develop a random-walk approach for deriving the streams of evidence. We will now describe this.

4.5.3 Random walk on the evidence flow network

To obtain the evidence streams we define a random walk on the evidence flow network for comparison ab. We denote the transition matrix for this model by $\boldsymbol{U}^{(ab)}$ to distinguish it from the random-walk on the aggregate NMA network defined in Section 4.4.3. We note that there is a different evidence flow network for each treatment comparison ab. We indicate this by the superscript (ab). Since the evidence flow network has directed edges the walker can only move in one direction along each edge (in the direction of evidence flow). Node a in the evidence flow network for comparison ab has only outgoing edges, and node b only incoming edges. We also note that the evidence flow network is acyclic [8]. This means that a walker can never visit any node more than once.

It is important to distinguish carefully between the random-walk model on the aggregate network and that on the evidence flow network. In Section 4.4.3 we defined a transition matrix for a random walker moving from node a to node b on the aggregate meta-analytic network. The walker was allowed to move in both directions along the edges of the network. We labelled this transition matrix $\boldsymbol{T}^{(ab)}$ where the superscript indicates the start and end nodes of the walk, i.e. the treatment comparison we are interested in. By analysing the average movement of the walker, we obtained the evidence flow. In this section we focus instead on a random walk on the evidence flow network, and our aim is to construct streams of evidence. The two approaches are summarised in Table 4.2.

To illustrate this, we consider the evidence flow network for comparison 1-2 in Figure 4.5 (a). We now construct a transition matrix for a random walk on this directed acyclic graph assuming that the walker starts at node 1. In contrast to random walks on the undirected meta-analytic graphs in Section 4.4.3, the walker can only move in one direction across each edge as indicated by the direction of evidence flow. If the flow $f_{cd}^{(ab)} = 0$ (because the associated hat matrix element $H_{cd}^{(ab)} \leq 0$), then no hop from c to d can occur. Each possible transition occurs with probabilities proportional to the evidence flows indicated in Figure 4.5 (a). More generally, for the evidence flow

Table 4.2: Summary of the two random-walk approaches to NMA. In one approach ('aggregate') the walker moves on the undirected aggregate network. In the second ('evidence flow'), the walker moves on the directed acyclic evidence flow network for a particular comparison of treatments. The transition matrices are denoted by $\boldsymbol{T}$ and $\boldsymbol{U}^{(ab)}$ respectively. Except for the imposition of a suitable absorbing state (see text) the transition probabilities on the aggregate network do not depend on the particular comparison that is studied. In contrast, there are separate evidence flow networks (and hence random-walk models) for each comparison ab, hence the superscript in $\boldsymbol{U}^{(ab)}$. The first column in the table indicates the sections in the text containing further definitions and details.

Section	Network	Transition probabilities	Measured quantity	Outcome
4.4.3	Aggregate	$T_{cd} = \dfrac{w_{cd}}{\sum\limits_{x \neq c} w_{cx}}$	Expected net number of times a walker crosses an edge while travelling from a to b	Flow of evidence through the edge (elements of the hat matrix in the row corresponding to comparison ab)
4.5.3	Evidence flow	$U_{cd}^{(ab)} = \begin{cases} \dfrac{H_{cd}^{(ab)}}{\sum_{x \neq c} H_{cx}^{(ab)}} & \text{if } H_{cd}^{(ab)} > 0 \\ 0 & \text{if } H_{cd}^{(ab)} < 0 \end{cases}$	Proportion of walkers taking a particular path while travelling from a to b	Evidence streams for the comparison between a and b

network of comparison ab, the elements of the transition matrix $\boldsymbol{U}^{(ab)}$ are given by

$$U_{cd}^{(ab)} = \frac{f_{cd}^{(ab)}}{\sum_{x \neq c} f_{cx}^{(ab)}} = \begin{cases} \frac{H_{cd}^{(ab)}}{\sum_{x \neq c} H_{cx}^{(ab)}} & \text{if } H_{cd}^{(ab)} > 0 \\ \\ 0 & \text{if } H_{cd}^{(ab)} < 0. \end{cases} \tag{4.20}$$

For the comparison ab, the walker remains at b indefinitely once it gets there, i.e. we have $U_{bb}^{(ab)} = 1$, and the probability of transitioning from b to any other node $c \neq b$ is $U_{bc}^{(ab)} = 0$. All other elements of the matrix $\boldsymbol{U}^{(ab)}$ are given by Equation (4.20).

For the example in Figure 4.5 (a), the transition matrix for a random walk on this graph is

$$\boldsymbol{U}^{(1\text{-}2)} = \begin{pmatrix} 0 & 0.635 & 0.365 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{0.251}{0.251+0.114} & 0 & \frac{0.114}{0.251+0.114} \\ 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0.635 & 0.365 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.688 & 0 & 0.312 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{4.21}$$

The third row of $\boldsymbol{U}^{(1\text{-}2)}$ corresponds to transitions from node 3. From Equation (4.20) and the edge flows shown in Figure 4.5 (a), we find that if the walker is at node 3, then it moves to either node 2 or node 4 with probabilities $0.251/(0.251 + 0.114)$ and $0.114/(0.251 + 0.114)$ respectively. Similar calculations are done to find the elements in the other rows. Once arrived at 2 the walker remains there indefinitely. This behaviour is described by the second row of $\boldsymbol{U}^{(1\text{-}2)}$.

The walker can take one of three paths from 1 to 2: $\pi_1 = 1 \to 2$, $\pi_2 = 1 \to 3 \to 2$, or $\pi_3 = 1 \to 3 \to 4 \to 2$. These are the same as the paths of evidence defined in Section 4.5.1 and are illustrated in Figure 4.5 (b). The probability of a walker taking a certain path is given by the product of the individual transition probabilities associated with each edge along that path (Equation (4.9)). For example, the probability that a random walker takes the path $1 \to 3 \to 2$ is $P^{(1\text{-}2)}(\pi_2) = U_{1\text{-}3}^{(1\text{-}2)} U_{3\text{-}2}^{(1\text{-}2)} = 0.365 \times 0.688$.

The probability that a walker takes a given path can also be measured from simulations of the random-walk process on the evidence flow network. To do this one simulates a large ensemble of independent walkers, and measures the proportions of walkers taking each path. We can think of this as flows of walkers through the different paths. We use this interpretation to provide a general analytical definition of the flow

of evidence through a particular path: for the evidence flow network for comparison ab, we define

$$\phi_i = P^{(ab)}(\pi_i) = \prod_{cd \in \pi_i} U_{cd}^{(ab)}. \tag{4.22}$$

With this definition we can construct decompositions such as the one in Equation (4.18) for all networks. From the ϕ_i the proportion contributions can then be calculated via Equation (4.19).

For the example in Figure 4.5 (a), Equation (4.22) leads to the streams,

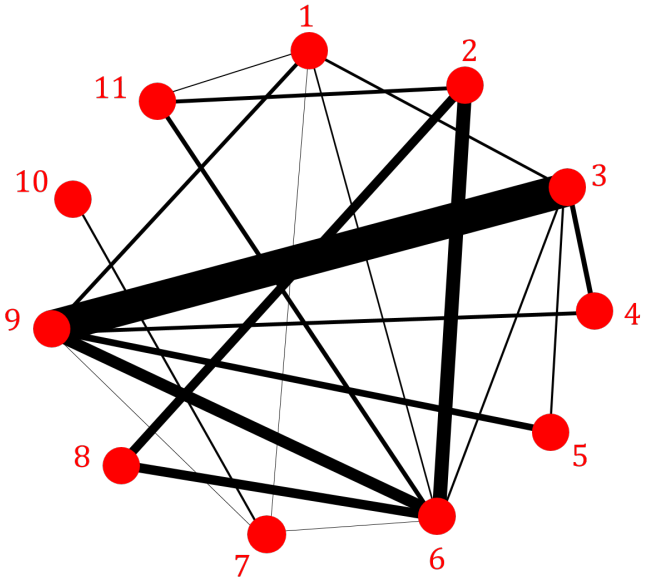
$$\begin{aligned} S_1 = (\pi_1, \phi_1): \quad & \pi_1 = 1 \to 2 && \phi_1 = U_{1\text{-}2}^{(1\text{-}2)} \\ & && = 0.635 \qquad (4.23) \\ S_2 = (\pi_2, \phi_2): \quad & \pi_2 = 1 \to 3 \to 2 && \phi_2 = U_{1\text{-}3}^{(1\text{-}2)} U_{3\text{-}2}^{(1\text{-}2)} \\ & && = 0.365 \times \frac{0.251}{0.251 + 0.114} = 0.251 \qquad (4.24) \\ S_3 = (\pi_3, \phi_3): \quad & \pi_3 = 1 \to 3 \to 4 \to 2 && \phi_3 = U_{1\text{-}3}^{(1\text{-}2)} U_{3\text{-}4}^{(1\text{-}2)} U_{4\text{-}2}^{(1\text{-}2)} \\ & && = 0.365 \times \frac{0.114}{0.251 + 0.114} \times 1 = 0.114 \qquad (4.25) \end{aligned}$$

For this simple example the random-walk approach results in the same evidence streams (and therefore proportion contributions) as the algorithm by Papakonstantinou et al, see Figure 4.5 (b).

The random-walk approach provides an analytical construction of the proportion contributions. Unlike the iterative algorithm, the outcome is unambiguous. In the following section we demonstrate how the random-walk approach can be used for the more intricate network from Section 4.2.

4.6 Application to real data set

We now apply the random-walk approach to the data set described in Section 4.2. Following Rücker and Schwarzer (2014) [39], we choose a fixed-effect model ($\tau^2 = 0$). The edge weights in the aggregate network were obtained using the methods described in Section 4.3 and are shown in Figure 4.6.



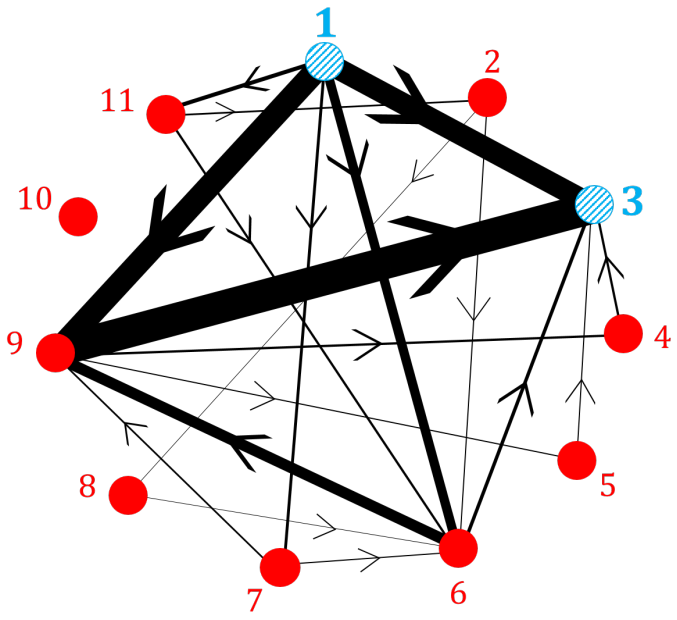
Aggregate edge weights, w_{ab}:	
1 ↔ 3:	7.605
1 ↔ 6:	4.432
1 ↔ 7:	1.785
1 ↔ 9:	9.410
1 ↔ 11:	2.322
2 ↔ 6:	36.419
2 ↔ 8:	23.576
2 ↔ 11:	10.474
3 ↔ 4:	13.559
3 ↔ 5:	5.118
3 ↔ 6:	5.187
3 ↔ 9:	87.697
4 ↔ 9:	10.533
5 ↔ 9:	16.946
6 ↔ 7:	1.620
6 ↔ 8:	23.759
6 ↔ 9:	29.714
6 ↔ 11:	12.154
7 ↔ 9:	1.713
7 ↔ 10:	5.894

Figure 4.6: The aggregate network for the depression data set in Section 4.2. Treatments 1 to 11 are defined in Figure 4.1. Here the thickness of each edge ab represents the associated weight, w_{ab}. The aggregate weights, as presented in the box, were calculated using the methods described in Section 4.3. The values are quoted to 3 decimal places.

4.6.1 Evidence flows

First, we use the random-walk approach described in Section 4.4.3 to obtain the evidence flows for a certain comparison. We focus on the comparison of treatments 1 (tricyclic or tetracyclic antidepressants) and 3 (psychotherapy + usual care). To this end, we define the transition matrix for a random walker on the *aggregate network* (Figure 4.6) starting at node 1 and ending at node 3. Using Equation (4.12) we find

$$\boldsymbol{T}^{(1\text{-}3)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{pmatrix} 0 & 0 & 0.298 & 0 & 0 & 0.173 & 0.070 & 0 & 0.368 & 0 & 0.091 \\ 0 & 0 & 0 & 0 & 0 & 0.517 & 0 & 0.335 & 0 & 0 & 0.149 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.563 & 0 & 0 & 0 & 0 & 0 & 0.437 & 0 & 0 \\ 0 & 0 & 0.232 & 0 & 0 & 0 & 0 & 0 & 0.768 & 0 & 0 \\ 0.039 & 0.321 & 0.046 & 0 & 0 & 0 & 0.014 & 0.210 & 0.262 & 0 & 0.107 \\ 0.162 & 0 & 0 & 0 & 0 & 0.147 & 0 & 0 & 0.156 & 0.535 & 0 \\ 0 & 0.498 & 0 & 0 & 0 & 0.502 & 0 & 0 & 0 & 0 & 0 \\ 0.060 & 0 & 0.562 & 0.068 & 0.109 & 0.190 & 0.011 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.093 & 0.420 & 0 & 0 & 0 & 0.487 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{4.26}$$



Flow through each edge:	
1 → 3:	0.353
1 → 6:	0.152
1 → 7:	0.044
1 → 9:	0.380
1 → 11:	0.072
2 → 6:	0.022
2 → 8:	0.007
2 → 11:	−0.030
3 → 4:	−0.036
3 → 5:	−0.024
3 → 6:	−0.062
3 → 9:	−0.526
4 → 9:	−0.036
5 → 9:	−0.024
6 → 7:	−0.016
6 → 8:	−0.007
6 → 9:	0.178
6 → 11:	−0.042
7 → 9:	0.027
7 → 10:	0

Figure 4.7: The evidence flow network for the comparison of treatments 1 and 3 in the depression data set in Section 4.2. The thickness of each edge corresponds to the expected net number of times a random walker crosses each edge of the aggregate network in Figure 4.6 as it travels from node 1 to node 3. The direction of flow is indicated by the arrow. These values are summarised in the box and quoted to 3 decimal places.

We have labelled the rows and columns according to the treatments they represent and we quote the values of the entries in the matrix to 3 decimal places. The third row of $\boldsymbol{T}^{(1\text{-}3)}$ is constructed such that once the walker reaches node 3 (the end node) it remains there indefinitely.

As described in Section 4.4.3, the evidence flow through each direct comparison for the network comparison 1-3 is obtained from the expected net number of times a walker crosses each edge as it travels from node 1 to node 3 on the aggregate network (Figure 4.6). The expected net number of times a walker crosses each edge can be estimated by simulating a large ensemble of random walkers, each moving independently as described by the transition matrix $\boldsymbol{T}^{(1\text{-}3)}$. For each walker we count the net number of times it crosses the designated edge, and we then subsequently average over all walkers. The more walkers we simulate, the more accurate our estimation.

Alternatively, we can use the analogy to electrical networks described in Section 4.4.2 to obtain an analytical result for this value in terms of electric current. These methods are described in more detail in Section 4.14 of the Appendix. We choose the analytical approach which results in the evidence flow network shown in Figure 4.7. We find that

for the comparison of treatments 1 and 3, most of the evidence flows directly from 1 to 3 or indirectly via treatment 9. Comparing Figures 4.6 and 4.7 we observe that the pairwise comparison of treatments 7 and 10 is the only piece of direct evidence that has no influence on the network comparison 1-3.

The hat matrix of the aggregate model for this data is given in Section 4.15.1 of the Appendix. The flow network obtained from the row of the hat matrix corresponding to the comparison of treatments 1 and 3 is identical to the network in Figure 4.7.

4.6.2 Proportion contributions

Next, we calculate the proportion contributions for the network comparison 1-3. To do this we first define the transition matrix for a random walker moving from node 1 to node 3 on the *evidence flow network* (Figure 4.7). From Equation (4.20) we find

$$\boldsymbol{U}^{(1\text{-}3)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{pmatrix} 0 & 0 & 0.353 & 0 & 0 & 0.152 & 0.044 & 0 & 0.380 & 0 & 0.072 \\ 0 & 0 & 0 & 0 & 0 & 0.755 & 0 & 0.245 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.259 & 0 & 0 & 0 & 0 & 0 & 0.741 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.371 & 0 & 0 & 0.629 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.899 & 0.061 & 0.040 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.415 & 0 & 0 & 0 & 0.585 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \tag{4.27}$$

where we have again labelled the rows and columns. Matrix entries are quoted to 3 decimal places. The third row indicates that once a walker reaches node 3 it remains there indefinitely. Since treatment 10 is disconnected from all other nodes in the evidence flow network (Figure 4.7), the probability of transitioning to this node from any other is zero. Similarly, if the walker starts at node 10, it remains there forever ($U^{(1\text{-}3)}_{10\text{-}10} = 1$).

The set of all possible paths that a random walker can take on the evidence flow

network can be found using a recursive algorithm [48]. The probability with which the walker takes a particular path is calculated from Equation (4.22). This is the flow of evidence through that path. For the comparison of treatments 1 and 3 in the depression data set, we find 27 distinct paths. These paths and their associated flow ϕ_i make up the evidence streams presented in Table 4.3. We find $\phi_i \geq 0$ and $\sum_i \phi_i = 1$. Using these values we can construct the network estimate $\hat{\theta}^{\text{net}}_{1\text{-}3}$ as a linear combination of direct and indirect estimates following each evidence path listed in Table 4.3. This leads to the same odds ratios as those quoted in Rücker and Schwarzer (2014) [39] up to the precision provided.

Table 4.3 also contains the streams identified by the algorithm in Papakonstantinou et al (2018) [9] (see Section 4.5.1). We present the results for three versions of the algorithm, Shortest, Random and Average. The results for the Random algorithm are obtained from one single run. Each result in the column labelled 'Average' is an average over 10^8 runs of the Random algorithm. From Table 4.3, it is clear that the streams identified by the iterative algorithm depend on the order in which paths are selected. For this example, fewer than half of the possible paths are identified by the Shortest and Random algorithms (paths not detected are indicated by the symbol '-'). Therefore, these versions of the algorithm fail to take into account multiple evidence paths that contribute to the NMA (and potentially have a high risk of bias).

Compared to the Shortest and Random versions of the algorithm, the Average algorithm produces results which are more similar to flows obtained from the random-walk approach. However, as described in Section 4.5.1, the frequency with which a path is selected across different runs depends on whether it shares edges with other paths in the network. Therefore, the results of the Average algorithm do not necessarily converge to the results from the random-walk approach even as the number of iterations becomes large.

Using Table 4.3 and Equation (4.19), we calculate the proportion contribution of each direct estimate to the network comparison of treatments 1 and 3 from the random-walk approach. These contributions are presented as percentages in the second column of Table 4.4. The direct evidence from trials comparing treatments 1 and 3 has the largest contribution followed by indirect evidence from trials comparing 3 and 9, and 1 and 9. Table 4.4 also contains the proportion contributions obtained from

Table 4.3: Evidence streams (paths and their associated flow) for the network comparison of treatments 1 and 3 in the depression data set in Section 4.2. Results obtained from the random-walk (RW) approach are presented along with the results from three versions of the algorithm in Papakonstantinou et al (2018) [9]. 'Shortest' refers to the algorithm where paths are selected from shortest to longest. 'Random' describes the variant in which paths are selected at random, and 'Average' is the average over 10^8 iterations of the Random algorithm. Values are rounded to 4 decimal places. The Shortest and Random algorithms fail to identify all possible paths, as indicated by the symbol '-'.

Stream, S_i	Path, π_i	Associated flow, ϕ_i			
		RW approach (analytical)	Algorithm		
			Shortest	Random	Average
$S_1 = (\pi_1, \phi_1)$	$1, 3$	0.3526	0.3526	0.3526	0.3526
$S_2 = (\pi_2, \phi_2)$	$1, 6, 3$	0.0394	0.0622	0.0549	0.0303
$S_3 = (\pi_3, \phi_3)$	$1, 6, 9, 3$	0.1015	0.0901	0.0974	0.1091
$S_4 = (\pi_4, \phi_4)$	$1, 6, 9, 4, 3$	0.0069	-	-	0.0082
$S_5 = (\pi_5, \phi_5)$	$1, 6, 9, 5, 3$	0.0045	-	-	0.0048
$S_6 = (\pi_6, \phi_6)$	$1, 7, 6, 3$	0.0042	-	-	0.0055
$S_7 = (\pi_7, \phi_7)$	$1, 7, 6, 9, 3$	0.0108	0.0162	-	0.0061
$S_8 = (\pi_8, \phi_8)$	$1, 7, 6, 9, 4, 3$	0.0007	-	0.0162	0.0024
$S_9 = (\pi_9, \phi_9)$	$1, 7, 6, 9, 5, 3$	0.0005	-	-	0.0021
$S_{10} = (\pi_{10}, \phi_{10})$	$1, 7, 9, 3$	0.0246	0.0274	0.0274	0.0171
$S_{11} = (\pi_{11}, \phi_{11})$	$1, 7, 9, 4, 3$	0.0017	-	-	0.0060
$S_{12} = (\pi_{12}, \phi_{12})$	$1, 7, 9, 5, 3$	0.0011	-	-	0.0043
$S_{13} = (\pi_{13}, \phi_{13})$	$1, 9, 3$	0.3414	0.3798	0.3604	0.3656
$S_{14} = (\pi_{14}, \phi_{14})$	$1, 9, 4, 3$	0.0231	-	0.0194	0.0090
$S_{15} = (\pi_{15}, \phi_{15})$	$1, 9, 5, 3$	0.0153	-	-	0.0052
$S_{16} = (\pi_{16}, \phi_{16})$	$1, 11, 2, 6, 3$	0.0058	-	-	0.0076
$S_{17} = (\pi_{17}, \phi_{17})$	$1, 11, 2, 6, 9, 3$	0.0150	-	-	0.0085
$S_{18} = (\pi_{18}, \phi_{18})$	$1, 11, 2, 6, 9, 4, 3$	0.0010	0.0062	-	0.0034
$S_{19} = (\pi_{19}, \phi_{19})$	$1, 11, 2, 6, 9, 5, 3$	0.0007	0.0163	0.0225	0.0030
$S_{20} = (\pi_{20}, \phi_{20})$	$1, 11, 2, 8, 6, 3$	0.0019	-	0.0073	0.0024
$S_{21} = (\pi_{21}, \phi_{21})$	$1, 11, 2, 8, 6, 9, 3$	0.0049	-	-	0.0027
$S_{22} = (\pi_{22}, \phi_{22})$	$1, 11, 2, 8, 6, 9, 4, 3$	0.0003	-	-	0.0012
$S_{23} = (\pi_{23}, \phi_{23})$	$1, 11, 2, 8, 6, 9, 5, 3$	0.0002	0.0073	-	0.0010
$S_{24} = (\pi_{24}, \phi_{24})$	$1, 11, 6, 3$	0.0109	-	-	0.0163
$S_{25} = (\pi_{25}, \phi_{25})$	$1, 11, 6, 9, 3$	0.0280	0.0126	0.0409	0.0170
$S_{26} = (\pi_{26}, \phi_{26})$	$1, 11, 6, 9, 4, 3$	0.0019	0.0294	-	0.0054
$S_{27} = (\pi_{27}, \phi_{27})$	$1, 11, 6, 9, 5, 3$	0.0013	-	0.0011	0.0033

Table 4.4: Proportion contributions, expressed as percentages, for the network comparison of treatments 1 and 3 in the depression data set. Results obtained from the random walk (RW) approach are presented along with the results from three versions of the algorithm in Papakonstantinou et al (2018) [9]. Shortest refers to the algorithm where paths are selected from shortest to longest. Random is when paths are selected at random. Average is the average over 10^8 iterations of the Random algorithm. Values are rounded to 1 decimal place.

Direct evidence, *ab*	Proportion contribution, $p_{ab}^{(1\text{-}3)}$			
	RW approach	Algorithm		
		Shortest	Random	Average
1-3	35.3%	35.3%	35.3%	35.3%
1-6	5.6 %	6.1 %	6.0%	5.5%
1-7	1.3 %	1.3%	1.2%	1.3%
1-9	18.3 %	19.0%	18.7%	18.8%
1-11	1.7%	1.4%	1.6%	1.7%
2-6	0.5%	0.4%	0.4%	0.5%
2-8	0.1%	0.1%	0.1%	0.1%
2-11	0.6%	0.5%	0.5%	0.6%
3-4	1.1%	0.7%	1.0%	0.9%
3-5	0.7%	0.4%	0.4%	0.6%
3-6	2.7%	3.1%	2.9%	2.5%
3-9	22.6%	23.6%	23.2%	23.3%
4-9	1.1%	0.7%	1.0%	0.9%
5-9	0.7%	0.4%	0.4%	0.6%
6-7	0.4%	0.4%	0.3%	0.4%
6-8	0.1 %	0.1%	0.1%	0.1%
6-9	5.1%	4.8%	5.0%	5.2%
6-11	1.1%	0.9%	1.0%	1.1%
7-9	0.9%	0.9%	0.9%	0.8%
7-10	0%	0%	0%	0%

the three versions of the algorithm (Shortest, Random and Average). As before, these results depend on the order in which paths are selected.

4.7 Summary and Discussion

4.7.1 The analogy between random walks and evidence flow, and the role of the graph theoretical model

In this paper, we have presented a novel analogy between NMA and random walks. Edge weights from the aggregate graph theoretical NMA model define a transition matrix for a random walk on the network of evidence. The walker moves around on the aggregate network along edges corresponding to direct evidence. The movement of the

random walker contains information about the propagation of evidence through the network. In particular, we have shown that the expected net number of times a walker crosses an edge can be interpreted as the evidence flow through the direct comparison represented by that edge. Therefore, we can obtain the elements of the hat matrix of the aggregate model from the random-walk process on the aggregate network.

The flow of evidence defined by König et al (2013) [8] is based on a two-step version of the standard frequentist NMA model (see Section 4.8 of the Appendix). In the first step, the direct estimates are obtained by pooling evidence from trials making the same comparisons. For two-arm trials, a pairwise meta-analysis is performed. For multi-arm trials that compare a particular subset of treatments, an NMA is performed on the sub-graph described by the multi-arm trial design. The direct estimates are therefore separated into evidence that comes from two-arm trials and evidence from multi-arm trials. This is reflected in the hat matrix of this model. Consequently, in König et al's evidence flow networks, the flow through multi-arm trials is displayed separately. This is an interesting feature but, as the authors note, it is only feasible for simple networks [8].

In our definition of evidence flow, we have instead used a two-step version of the so-called *graph theoretical* model [7]. We make use of the fact that the adjusted weights describe a network of two-arm trials which is equivalent to the network of multi-arm trials. The direct estimates are then obtained from pairwise meta-analyses using the adjusted edge weights. The elements in the row of the hat matrix for a particular comparison then assign a single value of flow to each direct treatment comparison in the network. The flow through an edge therefore represents the combined contribution from all studies, two-arm and multi-arm, that make that comparison. While this means that the specific contribution of multi-arm studies is not displayed, our approach makes it easier to display evidence flow networks for graphs with a large number nodes, edges and multi-arm trials of varying designs. In addition, it is this property of the aggregate level graph theoretical approach that means we are able to make the analogy to random walks in the general case (i.e. networks including multi-armed trials).

As explained in Section 4.8 of the Appendix, the standard NMA model, the graph theoretical model and the aggregate level versions of both these models, all yield the same network treatment effect estimates [8, 39]. For networks containing exclusively

two-arm trials, the hat matrices of the two aggregate level models are the same. Therefore, for these networks, the evidence flow networks we define are the same as those in König et al.

The graph theoretical approach provides a straightforward visualisation of the flow of evidence for each treatment comparison. Random effects models and networks with multi-arm trials can be accounted for with no extra complications. For networks with both of these characteristics, heterogeneity needs to be combined with the original observed variances (i.e. one needs to use $\sigma^2_{i,ab} + \tau^2$ instead of $\sigma^2_{i,ab}$) before adjusting the weights to deal with multi-arm trials [39, 42].

4.7.2 The random walk derivation of evidence streams overcomes the limitations of previous algorithms

We have shown that the random-walk analogy for NMA leads to an analytical derivation of evidence streams. In doing so, we defined a second transition matrix, this time for a random walker moving on the evidence flow network. For each comparison of treatments ab there is one separate evidence flow network. The network is directed and it has no cycles. Walkers can only move in one direction along each edge, according to the direction of flow. All paths on this graph start at a and end at b. As the walker travels from a to b it moves along paths of direct and indirect evidence. Imagining a large number of independent random walkers undergoing this process, we interpret the proportion of walkers flowing through a particular path as the flow of evidence through that path, i.e. the flow of evidence through a path is the probability of a walker taking that path. This can be expressed analytically as the product of the transition probabilities along the edges that make up the path.

The analytical definition of evidence streams leads directly to an analytical derivation of the so-called proportion contributions defined in Papakonstantinou et al (2018) [9]. The result is unambiguous in contrast with previously proposed algorithms whose output depends on the order in which paths are selected. Furthermore, individual runs of the algorithm in Papakonstantinou et al can fail to identify all paths of evidence on the evidence flow network. This means that in the calculation of proportion contributions, multiple paths of evidence and their potential bias are not taken into account. Running

the algorithm many times and subsequently performing an average, we are eventually able to identify every path of evidence. However, the frequency with which a given path is selected depends strongly on the number of other paths with which it shares edges. As a result, the average flow obtained in this way does not accurately reflect the contribution of each path. The random-walk approach overcomes these limitations. All possible paths of evidence are identified and they are each assigned a value of flow that reflects the properties of the hat matrix. Therefore, all possible sources of bias are taken into account in the calculation of the proportion contributions.

In our application to real data, we observe that the differences between the proportion contributions obtained from the random walk approach and those obtained via the Average algorithm are relatively small. We would expect larger differences between the two approaches when the network contains fewer independent paths, i.e. when many pairs of paths have shared edges. This increases the bias in path selection in the Average algorithm. Potential characteristics that may lead to this scenario include networks that are highly connected, and networks that contain 'central' nodes or edges.

For multi-arm trials, the method presented in Papakonstantinou et al (2018) [9] naïvely treats each pairwise comparison in a multi-arm trial as an independent two-arm study. This does not account for correlations due to multi-arm trials. By instead using the adjusted weights from the graph theoretical model, we are able to define a network of two-arm trials that is equivalent to the original network of multi-arm trials. Therefore, an additional advantage of the methods presented in this paper, is that networks with multi-arm trials are handled more appropriately.

The CINeMA software currently relies on the algorithm in Papakonstantinou et al to calculate the relative contribution of studies with high, moderate and low risk of bias to each network treatment effect. Similarly, ROB-MEN (risk of bias due to missing evidence in network meta-analysis [19]) also uses the contribution matrix. Due to the advantages of the random-walk approach in deriving evidence streams we expect that applications such as these would benefit in terms of accuracy from the implementation of the method described in this paper. The recently updated PRISMA guidelines [49] require systematic reviewers to assess their body of evidence for risk of bias. The results of our paper mean that existing software tools to help researchers make this assessment can now be made more reliable. To this end, we have

implemented the aggregate hat matrix in `netmeta` [25], along with the random-walk approach to proportion contributions, see Section 4.16 in the Appendix for details.

The work described in this article focusses on improving the construction of the proportion contributions defined in Papakonstantinou et al (2018) [9]. However, other attempts to quantify the influence of different pieces of evidence in NMA have also been proposed. In particular, Rücker et al (2020) [50] developed a method for defining the 'statistical importance' of an individual study for a particular NMA estimate. In this approach, the importance of a study is based on the loss of precision (increase in variance) in a network estimate when that study is removed from the network. These values are uniquely defined and have an intuitive interpretation but cannot be expressed as a relative 'proportion'. The advantage of our approach is that, along with being uniquely defined and having an interpretation in terms of a random walk, contributions can be expressed as proportions that sum to one. As a result, our approach can be used in software such as CINeMA to illustrate the relative contributions of studies with varying levels of bias.

4.7.3 Potential future impact

We believe that the analogy between NMA and random walks is interesting and that it provides new insight into NMA methodology. In our work we have explored the applications of only a small subset of the random-walk literature; there is, therefore, scope for the impact of this analogy to be investigated further. We hope that by presenting this analogy, more ideas will be shared between the two disciplines and additional practical applications of the random walk-approach will be developed in the future.

For example, we have looked at the interpretation of the number of times a walker crosses each edge in the network. However, there is potentially also interest in investigating the number of times the walker visits each node. The random walk transition probabilities are proportional to the respective edge weights. Therefore, a walker is more likely to travel across an edge corresponding to a more precise treatment effect estimate. The expected number of times a walker visits a certain node will depend on how many connections the node has, and the weight (i.e. the inverse variance) associated with each of these connections. A node corresponding to a

treatment that is involved in many direct comparisons will be visited more often than a node corresponding to a treatment with comparatively few connections. Furthermore, the larger the weight associated with the edges connected to a certain node, the more often the random walker will visit that node. Potentially, this value provides a measure of vertex centrality that accounts for both connectivity and the precision of treatment effect estimates. There may also be interest in measuring random walk variation. The variability in the information gathered along different paths traversed by a walker moving on the evidence flow network may indicate inconsistency between paths of indirect evidence. Finally, we may also be able to use the random-walk analogy in the methodology for planning future studies based on an NMA. By considering a random walk on the network with the addition of the proposed study, it may be possible to work out how much the addition of that study will contribute to the overall results.

In summary, by using the analogy to electrical networks as an intermediate step, we have made a novel connection between NMA and random walks. The interdisciplinary analogy provides new insight into NMA methodology. In particular, the analogy leads to an analytical derivation of the proportion contribution matrix without the ambiguity of existing numerical algorithms. Our approach can therefore be used to reliably quantify the contribution of individual study limitations to the resulting network treatment effects. We hope that this paper will provide a starting point for future developments of NMA methodology that can benefit from ideas in the random-walk literature.

Data Availability Statement

The data, results and associated codes used in this work can be found in the GitHub repository here https://github.com/AnnieDavies/NMA_and_RW.

4.8 Appendix A: Frequentist NMA

4.8.1 Standard and graph theoretical approaches ('reduce dimensions' vs. 'reduce weights')

4.8.1.1 Standard frequentist NMA

The standard frequentist approach to NMA is a regression analysis [6, 13, 40]. The method relies on a design matrix $\boldsymbol{X}$ which is constructed to have full rank. Each n_i-arm trial contributes $n_i - 1$ independent observations from which we aim to estimate $N - 1$ independent network treatment effects. Therefore, the matrix $\boldsymbol{X}$ has dimensions $\sum_i (n_i - 1) \times (N - 1)$. The 'global baseline' treatment is chosen as treatment 1. Each column of $\boldsymbol{X}$ then refers to a treatment $\in \{2, \ldots, N\}$. The rows represent the comparisons to the *trial-specific* baseline in each study. For a given row, the entry in the column corresponding to the treatment that is compared with the trial-specific baseline treatment is $+1$. If the trial specific baseline treatment is not the global baseline treatment, there is a -1 in the column corresponding to the trial-specific baseline. All other elements in the row are zero.

The so-called 'information matrix' is defined as $\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}$ where $\boldsymbol{V}$ is the block-diagonal variance-covariance matrix. Each trial contributes an $(n_i - 1) \times (n_i - 1)$ block to $\boldsymbol{V}$ with observed variances on the diagonal and covariances (due to multi-arm trials) off the diagonal. The inverse of this matrix, $\boldsymbol{V}^{-1}$, is distinct from the matrix $\boldsymbol{W}$ in the main text. The latter is a diagonal $K \times K$ matrix that contains the weight associated with each edge in the network after the adjustment for multi-arm trials and the aggregation of direct estimates.

The hat matrix of the standard model is [8, 39]

$$\boldsymbol{H}^{(\text{standard})} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1}. \tag{4.28}$$

4.8.1.2 Graph theoretical approach

Rücker introduced an alternative *graph theoretical* approach to NMA based on electrical network theory [7]. This model is formulated around an edge-vertex incidence matrix $\boldsymbol{B}_0$ with dimensions $\sum_i \frac{n_i(n_i-1)}{2} \times N$, where $\sum_i \frac{n_i(n_i-1)}{2}$ is the total number of pairwise comparisons in the network. We write $\boldsymbol{B}_0$ for this matrix to distinguish it from the

(similar) matrix $\boldsymbol{B}$ in the aggregate model described in Section 4.3.2 of the main paper. Each n_i-arm study contributes $\frac{n_i(n_i-1)}{2}$ rows to $\boldsymbol{B}_0$. Each column represents a treatment $\in \{1, \ldots, N\}$. Unlike the design matrix, $\boldsymbol{B}_0$ does not have full rank. Indeed, the elements in each row of $\boldsymbol{B}_0$ sum to zero [7, 39]. Entries of $\boldsymbol{B}_0$ are $+1$ in the column corresponding to the 'baseline' treatment of the comparison represented by that row, and -1 in the column corresponding to the treatment compared to that baseline.

We write $\boldsymbol{W}_0$ for the weight matrix of this model. Again, this is distinct from the matrix $\boldsymbol{W}$ in the main paper. $\boldsymbol{W}_0$ has dimensions $\left(\sum_i \frac{n_i(n_i-1)}{2}\right) \times \left(\sum_i \frac{n_i(n_i-1)}{2}\right)$ and contains on its diagonal the adjusted weights, $w_{i,ab}$, defined in the main paper. We obtain the adjusted weights from a method described in References [7, 39, 42] which accounts for the correlations introduced by multi-arm trials. An important result of this method is that the adjusted weights describe the weights associated with a network of two-arm trials that is equivalent to the original network of multi-arm trials in the sense that the resulting relative treatment effect estimates from the network of two-arm trials are the same as those from the original network. By using these weights, we can therefore apply any NMA methodology that is only valid for networks of two-arm trials. The hat matrix of this model is,

$$\boldsymbol{H}^{(\text{graph})} = \boldsymbol{B}_0(\boldsymbol{B}_0^\top \boldsymbol{W}_0 \boldsymbol{B}_0)^+ \boldsymbol{B}_0^\top \boldsymbol{W}_0. \tag{4.29}$$

4.8.1.3 'Reduce dimension' vs 'reduce weights'

The design matrix $\boldsymbol{X}$ contains the same information about the structure of the network as $\boldsymbol{B}_0$ but has lower dimensions and full rank. For this reason Rücker and Schwarzer (2014) [39] termed the standard model the 'reduce dimension' approach. The alternative (graph theoretical) method relies on reducing the weights associated with observations from multi-arm trials. Therefore, this was termed the 'reduce weights' approach [39]. In Rücker and Schwarzer (2014) the authors proved that, although their respective hat matrices are different, the two approaches give rise to the same network treatment effect estimates and are, therefore, equivalent.

4.8.2 Two-step models and evidence flow

The concept of evidence flow was introduced by König et al (2013) [8]. Their approach was based on a two-step, or 'aggregate', version of the *reduce dimensions* (standard) model [11, 43]:

Step 1. In the first step, evidence from all trials making the same comparisons is pooled. For two-arm trials, a pairwise meta-analysis is performed. For multi-arm trials with a particular design, an NMA is performed on the sub-graph described by the multi-arm design. The results from this first step define the direct evidence.

Step 2. In step two, the direct estimates are used as observations in a linear regression model.

The hat matrix associated with this model defines the evidence flow. Since the direct evidence is separated into evidence from two-arm trials and evidence from multi-arm trials, König et al display the flow through multi-arm trials separately on the evidence flow networks. The authors note that, with this approach, there is no unique way to represent evidence flow through multi-arm trials. Furthermore, explicitly showing multi-arm trials on evidence flow networks becomes increasingly difficult for large, highly connected networks.

In the main paper, we instead describe a two-step (aggregate) version of the *reduce weights* (=graph theoretical) approach. The fact that the reduce weights model defines a matrix of two-arm trials that is equivalent to the matrix of multi-arm trials makes the two-step approach simpler. In the first step, we perform a pairwise meta-analysis across each edge using the adjusted weights. In the second step, we combine this aggregate (direct) data in a network meta-analysis. This approach yields exactly the same relative treatment effect estimates as the one-step reduce weights approach and, consequently, the reduce-dimensions approach. This equivalence also holds true for random effects models. One then needs to account for heterogeneity, i.e. $\sigma^2_{i,ab}$ is replaced by $\sigma^2_{i,ab} + \tau^2$, before using the adjustment method [7, 39, 42] to obtain the adjusted weights.

For networks containing exclusively two-arm trials, the hat matrices from the two aggregate models are exactly equal. Therefore, in this scenario, our evidence flow networks are the same as those defined by König et al [8]. The differences arise in the presence of multi-arm trials. Our approach does not explicitly show the flow through multi-arm trials. Instead, the flow through each edge represents the pooled

contribution from all studies that make that comparison. This is only made possible by using the reduce weights method to define a network of two-arm trials. Since each edge is associated with only one value of evidence flow, our approach makes it easier to construct evidence flow networks for complicated networks, i.e. those with many nodes, many connections, and many different multi-arm trials. This also makes it possible to calculate the proportion contribution matrix for networks of multi-arm trials. With the evidence flow networks defined by König et al [8], this was not possible as the presence of multi-arm trials meant there were multiple values of flow associated with each edge.

4.9 Appendix B: Hat matrix for the fictional example

For the fictional example in Figure 4.2, the aggregate weight matrix is

$$\boldsymbol{W} = \text{diag}(1, 3, 4, 6, 5, 2, 7).$$

From Equation (4.3) in the main paper we recall that the edge incidence matrix is

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \tag{4.30}$$

where the columns represent treatments 1, 2, 3, 4, and 5, and the rows represent the edges (direct comparisons) 1-2, 1-3, 1-5, 2-3, 2-4, 3-5, and 4-5. The hat matrix is calculated using

$$\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{W} \boldsymbol{B})^+ \boldsymbol{B}^\top \boldsymbol{W}, \tag{4.31}$$

which is Equation (4.5) in the main paper. The resulting matrix, with values quoted to 2 decimal places, is

$$H = \begin{pmatrix} 0.21 & 0.40 & 0.39 & -0.46 & -0.33 & -0.07 & -0.33 \\ 0.13 & 0.53 & 0.33 & 0.28 & -0.14 & -0.19 & -0.14 \\ 0.10 & 0.25 & 0.65 & -0.09 & 0.19 & 0.16 & 0.19 \\ -0.08 & 0.14 & -0.06 & 0.74 & 0.18 & -0.12 & 0.18 \\ -0.07 & -0.09 & 0.15 & 0.22 & 0.72 & 0.13 & -0.28 \\ -0.03 & -0.28 & 0.32 & -0.37 & 0.33 & 0.35 & 0.33 \\ -0.05 & -0.06 & 0.11 & 0.16 & -0.20 & 0.09 & 0.80 \end{pmatrix}, \tag{4.32}$$

where the rows and columns represent the edges (direct comparisons) 1-2, 1-3, 1-5, 2-3, 2-4, 3-5, and 4-5. Figure 4.2 (b) in the main paper shows the evidence flow network for comparison 1-2, as indicated by the values in the first row of the matrix in Equation (4.32).

4.10 Appendix C: Further comments on choice of coefficients $f_{cd}^{(ab)}$ and interpretation of evidence flow

4.10.1 Convention for coefficients $f_{cd}^{(ab)}$

For each pair of nodes c and d, one of the coefficients $f_{cd}^{(ab)}$ and $f_{dc}^{(ab)}$ in Equation (4.6) of the main paper is positive, and the other is zero. The coefficients fulfill the relations labelled 1, 2 and 3 in Section 4.3.4 of the main paper. These properties in turn suggest that the positive coefficients $f_{cd}^{(ab)}$ have an interpretation as flows. Property 3 for example states that the sum of inflows equals the sum of outflows at nodes other than a and b.

Alternatively, one could have chosen a convention in which $f_{cd}^{(ab)} = -f_{dc}^{(ab)}$ for all pairs c and d. This is the choice made in König et al (2013) [8] and is perhaps more in-line with an expectation that the flow from c to d ought to be the negative of the flow from d to c. It is important to note though, that these options are alternative, but ultimately equivalent, parameterisations of the same problem. It is purely a matter of

choice and convenience which one to use.

Our choice follows the conventions in Papakonstantinou et al (2018) [9]. For each pair cd there is then only one non-zero flow variable. This minimises the number of relevant quantities in the ensuing equations. This in turn makes the definitions of the transition rates $U_{cd}^{(ab)}$ in Equation (4.20) straightforward. These have to be non-negative.

The broader idea is that for each pair c and d only one flow is non-zero, that from c to d, or that from d to c. Which one it is indicates the direction of the flow. The positive value of that flow variable describes the magnitude of the flow.

4.10.2 Interpretation of evidence flow

We adopted the term 'evidence flow' from existing literature [8, 9]. Although the term has been in use for a number of years, we find it hard to extract from the existing literature what exactly is the nature of these flows. For example, it is not easy to pinpoint what precisely the word 'evidence' means in mathematical terms. Neither is it immediately clear how evidence can be located at a node, and how it then 'flows' from one node to another. Nevertheless, it is apparent that the three properties of the coefficients $f_{cd}^{(ab)}$ in Section 4.3.4 of the main paper (previously stated by König et al [8]), describe properties that one would associate with a flow.

The random-walk picture developed in this paper can contribute to developing a better understanding of what exactly it is that is flowing. Namely, it is random walkers starting at a and ending at b that 'flow' along the network based on the rules defined by the transition matrix $\boldsymbol{T}$ defined in Equation (4.12) of the main manuscript. More precisely, when the coefficient $f_{cd}^{(ab)}$ is positive, it captures the net number of times a walker starting at a and ending at b passes through the edge cd. All walkers start at a and end at b. For such a walker, the net number of departures out of node a must be one (property 1 in Section 4.3.4), and the net number of arrivals into b is also one (property 2). No walkers can be created or destroyed at any of the other nodes, and neither can they remain indefinitely at any of these nodes. Therefore the total number of times the walker arrives at any node other than a and b is the same as the number of times it leaves that node. This is what property 3 in Section 4.3.4 describes.

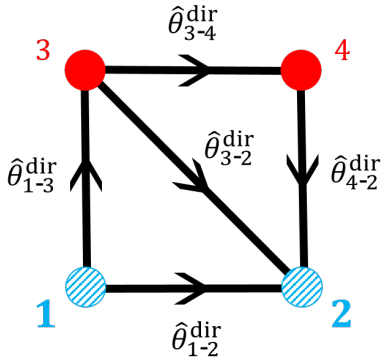


Figure 4.8: Meta-analytic graph of the example in Figure 4.5 (a). We focus on the comparison between treatments 1 and 2, as indicated by the blue striped colour of the nodes representing these treatments. Arrows show the sign conventions for the direction of evidence flow. Direct evidence for the relative treatment effects from the trial data are also indicated next to each comparison.

4.11 Appendix D: Heuristic argument for properties of the hat matrix and evidence flow

In this section we give a brief heuristic argument for the properties of the hat matrix in Section 4.3.4 of the main paper. These properties were stated in König et al (2013) [8], an algebraic proof for some of the properties was given in Papakonstantinou et al (2018) [9]. We present our argument using the example in Figure 4.5 of the main paper, but this can be generalised to more complex networks.

The network in Figure 4.5 (a) is the evidence flow network for the comparison between treatments 1 and 2. It contains four nodes. For illustration and to fix sign conventions for the flow of evidence, the network is shown again in Figure 4.8. Without loss of generality we assume that the direction of all edges are chosen such that $H^{(1\text{-}2)}_{cd} > 0$ for all edges cd shown in Figure 4.8. This means that $f^{(1\text{-}2)}_{cd} = H^{(1\text{-}2)}_{cd}$ for all cd.

The three properties in Section 4.3.4 translate into

1. $f^{(1\text{-}2)}_{1\text{-}2} + f^{(1\text{-}2)}_{1\text{-}3} = 1$;
2. $f^{(1\text{-}2)}_{1\text{-}2} + f^{(1\text{-}2)}_{3\text{-}2} + f^{(1\text{-}2)}_{4\text{-}2} = 1$;
3. $f^{(1\text{-}2)}_{1\text{-}3} = f^{(1\text{-}2)}_{3\text{-}2} + f^{(1\text{-}2)}_{3\text{-}4}$ and $f^{(1\text{-}2)}_{3\text{-}4} = f^{(1\text{-}2)}_{4\text{-}2}$.

We address these one-by-one. To do this we use Equation (4.7) from the main paper, $f^{(1\text{-}2)}_{cd} = H^{(1\text{-}2)}_{cd}$, and the above sign convention to note that

$$\hat{\theta}^{\text{net}}_{1\text{-}2} = f^{(1\text{-}2)}_{1\text{-}2}\hat{\theta}^{\text{dir}}_{1\text{-}2} + f^{(1\text{-}2)}_{1\text{-}3}\hat{\theta}^{\text{dir}}_{1\text{-}3} + f^{(1\text{-}2)}_{3\text{-}2}\hat{\theta}^{\text{dir}}_{3\text{-}2} + f^{(1\text{-}2)}_{4\text{-}2}\hat{\theta}^{\text{dir}}_{4\text{-}2} + f^{(1\text{-}2)}_{3\text{-}4}\hat{\theta}^{\text{dir}}_{3\text{-}4}. \tag{4.33}$$

4.11.1 $f^{(1\text{-}2)}_{1\text{-}2} + f^{(1\text{-}2)}_{1\text{-}3} = 1$

Imagine we have one set of direct estimates,

$$\hat{\boldsymbol{\theta}}^{\text{dir}} = (\hat{\theta}^{\text{dir}}_{1\text{-}2}, \hat{\theta}^{\text{dir}}_{1\text{-}3}, \hat{\theta}^{\text{dir}}_{3\text{-}2}, \hat{\theta}^{\text{dir}}_{3\text{-}4}, \hat{\theta}^{\text{dir}}_{4\text{-}2}), \tag{4.34}$$

resulting in a network estimate $\hat{\theta}^{\text{net}}_{1\text{-}2}$ via Equation (4.33).

Imagine now a different set of direct estimates

$$\hat{\boldsymbol{\theta}}'^{\text{dir}} = (\hat{\theta}'^{\text{dir}}_{1\text{-}2}, \hat{\theta}'^{\text{dir}}_{1\text{-}3}, \hat{\theta}'^{\text{dir}}_{3\text{-}2}, \hat{\theta}'^{\text{dir}}_{3\text{-}4}, \hat{\theta}'^{\text{dir}}_{4\text{-}2}), \tag{4.35}$$

such that

$$\begin{aligned}
\hat{\theta}'^{\text{dir}}_{1\text{-}2} &= \hat{\theta}^{\text{dir}}_{1\text{-}2} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{1\text{-}3} &= \hat{\theta}^{\text{dir}}_{1\text{-}3} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}2} &= \hat{\theta}^{\text{dir}}_{3\text{-}2}, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}4} &= \hat{\theta}^{\text{dir}}_{3\text{-}4}, \\
\hat{\theta}'^{\text{dir}}_{4\text{-}2} &= \hat{\theta}^{\text{dir}}_{4\text{-}2}.
\end{aligned} \tag{4.36}$$

We write $\hat{\theta}'^{\text{net}}_{1\text{-}2}$ for the network estimate from the dataset $\hat{\boldsymbol{\theta}}'^{\text{dir}}$.

Using the sign convention in which $\hat{\theta}^{\text{dir}}_{cd}$ denotes the effect of treatment d minus that of c, Equation (4.36) indicates that the direct effect of treatment 2 compared to treatment 1 in the dataset $\hat{\boldsymbol{\theta}}'^{\text{dir}}$ is Δ units greater than in dataset $\hat{\boldsymbol{\theta}}^{\text{dir}}$. Similarly, the relative effect of treatment 3 relative to treatment 1 is Δ units higher. Given that treatments 2 and 3 are the only ones treatment 1 is compared to directly in this network (see Figure 4.8) we would then expect

$$\hat{\theta}'^{\text{net}}_{1\text{-}2} = \hat{\theta}^{\text{net}}_{1\text{-}2} + \Delta. \tag{4.37}$$

Using Equation (4.33) and its analogue for the dashed treatment effects, we find

$$\hat{\theta}'^{\text{net}}_{1\text{-}2} - \hat{\theta}^{\text{net}}_{1\text{-}2} = \Delta\left(f^{(1\text{-}2)}_{1\text{-}2} + f^{(1\text{-}2)}_{1\text{-}3}\right), \tag{4.38}$$

and we therefore conclude

$$f^{(1\text{-}2)}_{1\text{-}2} + f^{(1\text{-}2)}_{1\text{-}3} = 1. \tag{4.39}$$

4.11.2 $f_{1\text{-}2}^{(1\text{-}2)} + f_{3\text{-}2}^{(1\text{-}2)} + f_{4\text{-}2}^{(1\text{-}2)} = 1$

We again imagine a second set of data, now with

$$\begin{aligned}
\hat{\theta}'^{\text{dir}}_{1\text{-}2} &= \hat{\theta}^{\text{dir}}_{1\text{-}2} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{1\text{-}3} &= \hat{\theta}^{\text{dir}}_{1\text{-}3}, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}2} &= \hat{\theta}^{\text{dir}}_{3\text{-}2} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}4} &= \hat{\theta}^{\text{dir}}_{3\text{-}4}, \\
\hat{\theta}'^{\text{dir}}_{4\text{-}2} &= \hat{\theta}^{\text{dir}}_{4\text{-}2} + \Delta.
\end{aligned} \tag{4.40}$$

This means that treatment 2 is now consistently doing better (or consistently doing worse) by Δ units in relation to all treatments it is compared to directly in the network. The overall effect of this must be that

$$\hat{\theta}'^{\text{net}}_{1\text{-}2} = \hat{\theta}^{\text{net}}_{1\text{-}2} + \Delta, \tag{4.41}$$

i.e. the effect of treatment 2 relative to that of treatment 1 is now Δ units greater. Using again Equation (4.33) for the data sets $\hat{\boldsymbol{\theta}}^{\text{dir}}$ and $\hat{\boldsymbol{\theta}}'^{\text{dir}}$ respectively, we now have

$$\hat{\theta}'^{\text{net}}_{1\text{-}2} - \hat{\theta}^{\text{net}}_{1\text{-}2} = \Delta \left(f_{1\text{-}2}^{(1\text{-}2)} + f_{3\text{-}2}^{(1\text{-}2)} + f_{4\text{-}2}^{(1\text{-}2)} \right), \tag{4.42}$$

from Equation (4.40). Therefore

$$f_{1\text{-}2}^{(1\text{-}2)} + f_{3\text{-}2}^{(1\text{-}2)} + f_{4\text{-}2}^{(1\text{-}2)} = 1. \tag{4.43}$$

4.11.3 $f_{1\text{-}3}^{(1\text{-}2)} = f_{3\text{-}2}^{(1\text{-}2)} + f_{3\text{-}4}^{(1\text{-}2)}$ and $f_{3\text{-}4}^{(1\text{-}2)} = f_{4\text{-}2}^{(1\text{-}2)}$

The first of these identities can be shown by looking at

$$\begin{aligned}
\hat{\theta}'^{\text{dir}}_{1\text{-}2} &= \hat{\theta}^{\text{dir}}_{1\text{-}2}, \\
\hat{\theta}'^{\text{dir}}_{1\text{-}3} &= \hat{\theta}^{\text{dir}}_{1\text{-}3} - \Delta, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}2} &= \hat{\theta}^{\text{dir}}_{3\text{-}2} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{3\text{-}4} &= \hat{\theta}^{\text{dir}}_{3\text{-}4} + \Delta, \\
\hat{\theta}'^{\text{dir}}_{4\text{-}2} &= \hat{\theta}^{\text{dir}}_{4\text{-}2},
\end{aligned} \tag{4.44}$$

and by realising that this means that treatment 3 now performs Δ units worse (or better) compared to all treatments it is directly compared to. This cannot affect the

network estimate of the treatment effect of 2 compared to 1, i.e. we expect $\hat{\theta}'^{\text{net}}_{1\text{-}2} = \hat{\theta}^{\text{net}}_{1\text{-}2}$. This leads to $f^{(1\text{-}2)}_{1\text{-}3} = f^{(1\text{-}2)}_{3\text{-}2} + f^{(1\text{-}2)}_{3\text{-}4}$.

The identity $f^{(1\text{-}2)}_{3\text{-}4} = f^{(1\text{-}2)}_{4\text{-}2}$ can be demonstrated in a similar way.

4.12 Appendix E: Electric current and evidence flow

In this section we demonstrate the relationship between electrical current and evidence flow. Consider an electrical network with N nodes and K edges. We define the vector of nodal or 'external' currents as $\boldsymbol{J} = (J_1, J_2, \ldots, J_N)^\top$. These represent currents flowing between a node of the network and an external sink or source. Our sign convention is such that a positive entry $J_a > 0$ indicates that a current goes into node a, whereas if $J_a < 0$, a current goes out of node a. We write $\boldsymbol{I} = (I_1, I_2, \ldots, I_K)^\top$ for the currents in the edges $k = ab, k = 1, 2, \ldots, K$. A positive value of I_{ab} indicates a flow of current from a to b, and we set $I_{ba} = -I_{ab}$.

We define $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_{ab}\}$ as the vector of voltages (potential differences) across the edges. That is, $\mathcal{V}_{ab} = v_a - v_b$ where v_a and v_b are the potentials at nodes a and b respectively. Ohm's law [51] can then be written as

$$\boldsymbol{I} = \boldsymbol{C}\boldsymbol{\mathcal{V}}, \tag{4.45}$$

where $\boldsymbol{C}$ is the $K \times K$ diagonal matrix of conductances (inverse resistances, $C_{ab} = (R_{ab})^{-1}$). Using this and Kirchhoff's laws, Rücker [7] demonstrated that $\boldsymbol{\mathcal{V}}$ can be written as

$$\boldsymbol{\mathcal{V}} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{C}\boldsymbol{B})^+ \boldsymbol{J}, \tag{4.46}$$

where $\boldsymbol{B}$ is the edge-incidence matrix of the network defined in Section 4.3 of the main paper. Substituting this into Ohm's Law (Equation (4.45)) yields the edge currents,

$$\boldsymbol{I} = \boldsymbol{C}\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{C}\boldsymbol{B}\right)^+ \boldsymbol{J}. \tag{4.47}$$

To make the analogy to evidence flow, we consider an electrical network with a battery attached across the nodes corresponding to the treatment comparison we are

interested in. For comparison ab the external current at node a is $J_a = +1$, at b we have $J_b = -1$. The current J_c at every other node $c \notin \{a, b\}$ is zero.

We can do this in turn for each of the K edges in the network. For convenience we label these $k = 1, \ldots, K$. We write $\boldsymbol{J}^{(k)}$ for the vector of nodal currents resulting in a situation where the battery is connected to the start and end points of edge k.

We then have K relations of the form in Equation (4.47),

$$\boldsymbol{I}^{(k)} = \boldsymbol{C}\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{C}\boldsymbol{B}\right)^+ \boldsymbol{J}^{(k)}. \tag{4.48}$$

We collect the internal currents $\boldsymbol{I}^{(k)}$ in a $K \times K$ matrix $\tilde{\boldsymbol{I}} = \begin{pmatrix} \boldsymbol{I}^{(1)} & \boldsymbol{I}^{(2)} & \ldots & \boldsymbol{I}^{(K)} \end{pmatrix}$. Similarly, we define the $N \times K$ matrix $\tilde{\boldsymbol{J}} = \begin{pmatrix} \boldsymbol{J}^{(1)} & \boldsymbol{J}^{(2)} & \ldots & \boldsymbol{J}^{(K)} \end{pmatrix}$. We then have

$$\tilde{\boldsymbol{I}} = \boldsymbol{C}\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{C}\boldsymbol{B}\right)^+ \tilde{\boldsymbol{J}}. \tag{4.49}$$

As an example, consider a simple network of three nodes 1,2,3 and where all possible edges (1-2, 1-3, 2-3) are present. Let $k = 1$ represent the edge 1-2, $k = 2$ represent 1-3, and $k = 3$ represent 2-3. The matrix of nodal currents is then

$$\tilde{\boldsymbol{J}} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{pmatrix}. \tag{4.50}$$

Each row of $\boldsymbol{J}$ represents a node, and each column represents a different placement of the battery. The first column corresponds to a battery attached across edge 1-2. Therefore, there is a $+1$ in the row corresponding to node 1, a -1 in the row corresponding to node 2 and a 0 for node 3. Similar reasoning is used to construct the other columns.

From this construction, it is clear that the matrix of nodal currents for this setup is equal to the transpose of the edge incidence matrix,

$$\tilde{\boldsymbol{J}} = \boldsymbol{B}^\top. \tag{4.51}$$

We can write the resulting matrix of edge currents in terms of its composite elements,

$$\tilde{\boldsymbol{I}} = \begin{pmatrix} I_{1\text{-}2}^{(1\text{-}2)} & I_{1\text{-}2}^{(1\text{-}3)} & I_{1\text{-}2}^{(2\text{-}3)} \\ I_{1\text{-}3}^{(1\text{-}2)} & I_{1\text{-}3}^{(1\text{-}3)} & I_{1\text{-}3}^{(2\text{-}3)} \\ I_{2\text{-}3}^{(1\text{-}2)} & I_{2\text{-}3}^{(1\text{-}3)} & I_{2\text{-}3}^{(2\text{-}3)} \end{pmatrix}, \tag{4.52}$$

where $I_{cd}^{(ab)}$ is the current through edge cd when a battery is attached across edge ab.

In the evidence flow analogy, we interpret the flow of current $I_{cd}^{(ab)}$ as the flow of evidence through edge cd for the network comparison ab. If the analogy holds (a proof follows below), we can write the elements of the hat matrix in terms of the edge currents. For the simple example above we have

$$\boldsymbol{H} = \begin{pmatrix} I_{1\text{-}2}^{(1\text{-}2)} & I_{1\text{-}3}^{(1\text{-}2)} & I_{2\text{-}3}^{(1\text{-}2)} \\ I_{1\text{-}2}^{(1\text{-}3)} & I_{1\text{-}3}^{(1\text{-}3)} & I_{2\text{-}3}^{(1\text{-}3)} \\ I_{1\text{-}2}^{(2\text{-}3)} & I_{1\text{-}3}^{(2\text{-}3)} & I_{2\text{-}3}^{(2\text{-}3)} \end{pmatrix}. \tag{4.53}$$

From Equations (4.52) and (4.53), it is clear that we need to prove that $\tilde{\boldsymbol{I}}^\top = \boldsymbol{H}$.

We now do this for a general setup. Taking the transpose of Equation (4.49), we find

$$\tilde{\boldsymbol{I}}^\top = \tilde{\boldsymbol{J}}^\top \left((\boldsymbol{B}^\top \boldsymbol{C} \boldsymbol{B})^+ \right)^\top \boldsymbol{B}^\top \boldsymbol{C}^\top. \tag{4.54}$$

From the definition of the pseudo-inverse it is possible to show that $(\boldsymbol{A}^+)^\top = (\boldsymbol{A}^\top)^+$ for a general matrix $\boldsymbol{A}$ (see Stoer and Bulirsch (2002) [52]). Using $\tilde{\boldsymbol{J}} = \boldsymbol{B}^\top$ and the fact that matrices $\boldsymbol{C}$ and $\boldsymbol{L} = \boldsymbol{B}^\top \boldsymbol{C} \boldsymbol{B}$ are symmetric ($\boldsymbol{C}^\top = \boldsymbol{C}$ and $\boldsymbol{L}^\top = \boldsymbol{L}$) we find

$$\tilde{\boldsymbol{I}}^\top = \boldsymbol{B} \left(\boldsymbol{B}^\top \boldsymbol{C} \boldsymbol{B} \right)^+ \boldsymbol{B}^\top \boldsymbol{C}. \tag{4.55}$$

We now recall that the hat matrix of the aggregate model is (see Equation (4.5) in the main paper)

$$\boldsymbol{H} = \boldsymbol{B} (\boldsymbol{B}^\top \boldsymbol{W} \boldsymbol{B})^+ \boldsymbol{B}^\top \boldsymbol{W}. \tag{4.56}$$

The weight associated with each edge in the aggregate network w_{ab} is given by the conductance (=inverse resistance) of that edge $C_{ab} = R_{ab}^{-1}$, see Section 4.4.1 in the main paper. The matrices $\boldsymbol{W}$ and $\boldsymbol{C}$ contain these weights and conductances on their respective diagonals ($\boldsymbol{W} = \text{diag}(w_{ab})$ and $\boldsymbol{C} = \text{diag}(C_{ab})$), and we therefore have

$$\boldsymbol{C} = \boldsymbol{W}. \tag{4.57}$$

Substituting this into Equation (4.55), we find

$$\tilde{\boldsymbol{I}}^\top = \boldsymbol{H}, \tag{4.58}$$

which is what we wanted to prove.

4.13 Appendix F: Random walks and electric networks

In this section, we demonstrate the relationship between electric current and random walks. This relationship is well known [37], and we include it here for completeness.

4.13.1 Dirichlet problem for electric circuits

We start from Ohm's law. Rather than using matrix notation as in Section 4.12, we formulate Ohm's law for the current I_{cd} in the edge cd,

$$I_{cd} = C_{cd}(v_c - v_d), \tag{4.59}$$

where v_c and v_d are the potentials at nodes c and d respectively. We have used the sign conventions of Doyle and Snell (2000) [37] to define the direction of current. As mentioned above we have $I_{cd} = -I_{dc}$.

In this section we focus on the scenario where a unit current flows into node a (from the exterior) and out of node b (to the exterior). No flows between the network and the exterior are possible at any other nodes. To create such a situation we imagine a battery connected to nodes a and b. The potential at b is set to zero, and that at a is $v_a = v_a^*$, with v_a^* such that the external current into a is equal to unity (the external current out of b is then also equal to unity). The asterisk indicates the choice of v_a resulting in a unit current into a. An illustration of this setup is shown in Figure 4.3 (b) in the main paper.

We use the superscript (ab) to indicate a battery attached across ab as described above, that is, we use $I_{cd}^{(ab)}$. Kirchhoff's law states that the total current at any node $c \neq a, b$ is zero,

$$\sum_d I_{cd}^{(ab)} = 0 \quad \forall c \neq a, b. \tag{4.60}$$

Substituting Equation (4.59) into Equation (4.60) and rearranging yields for $c \neq a, b$

$$v_c = \sum_d \frac{C_{cd}}{\sum_x C_{cx}} v_d = \sum_d v_d T_{cd}, \tag{4.61}$$

where we have used the definition of transition probabilities in Equation (4.8) in the main paper.

One can define a Laplacian matrix for this setup, $\boldsymbol{L}^{(ab)} = \mathbb{1} - \boldsymbol{T}^{(ab)}$, where $\mathbb{1}$ is the identity matrix [33]. A twice continuously differentiable function $\boldsymbol{f} : c \mapsto f_c$ is then called harmonic if it satisfies the Laplace equation [53], $\boldsymbol{L}^{(ab)}\boldsymbol{f} = \boldsymbol{0}$.

Equation (4.61) indicates that the function $c \mapsto v_c$ is harmonic at all points $c \neq a, b$. It also has boundary values at a and b: $v_a = v_a^*$ is chosen such that the current going into node a from the exterior is one, and we have $v_b = 0$. This constitutes a Dirichlet problem [34]. The uniqueness principle for Dirichlet problems then implies that v_c is uniquely determined for all c, given the boundary conditions at a and b. For further details see Doyle and Snell (2000) [37].

4.13.2 Dirichlet problem for random walks

We will now show that a quantity related to the expected net number of times a random walker visits a particular node c while travelling from a to b fulfills the same Laplace equation, and shares the same boundary conditions as the electric potentials in Section 4.13.1. The uniqueness of the solution of the Dirchlet problem then allows one to establish the analogy between electric networks and random walks. We now describe this in more detail.

We consider a walker starting at node a and reaching absorption when it arrives at node b. We write u_c for the expected number of times the walker visits node c before reaching b (with the convention that the final arrival at b does not constitute a visit to b, i.e. we have $u_b = 0$). The following relation then holds for $c \neq a, b$,

$$u_c = \sum_d u_d T_{dc}. \tag{4.62}$$

This equation can be understood as follows: In order to arrive at node c the walker must previously visit a neighbouring node d. The quantity u_d is the expected number of times this occurs. From such a node d the walker must then transition to c to contribute to u_c. This occurs with probability T_{dc}. Summing over all d results in Equation (4.62).

Equation (4.62) is of a similar form to Equation (4.61) in the electrical network. However T_{dc} appears on the right-hand side of Equation (4.62), whereas one has T_{cd} in Equation (4.61). We therefore write T_{dc} in terms of T_{cd}. Using Equation (4.8) from the

main paper, the definition $C_{ab} = R_{ab}^{-1}$, and the fact that $C_{cd} = C_{dc}$, we find

$$T_{dc} = \frac{T_{cd} \sum_x C_{cx}}{\sum_x C_{dx}}. \tag{4.63}$$

Substituting this into Equation (4.62) and re-arranging gives

$$\frac{u_c}{\sum_x C_{cx}} = \sum_d \frac{u_d}{\sum_x C_{dx}} T_{cd}. \tag{4.64}$$

Therefore, the object $c \mapsto u_c/(\sum_x C_{cx})$ is harmonic at all points $c \neq a, b$. Given that $u_b = 0$, we have the boundary condition $u_b/(\sum_x C_{bx}) = 0$. We note that Equation (4.64) and the boundary condition $u_b = 0$ can be derived for any quantity $\boldsymbol{u}$ that is proportional to the number of visits at the different nodes. The Laplace equation and the boundary condition therefore only fix u_c up to a factor. The uniqueness theorem for the Dirichlet problem also confirms that $u_c/(\sum_x C_{cx})$ is proportional to v_c from Section 4.13.1 for all c. The constant of proportionality is fixed by the boundary condition for u_a.

We now show that the choice $u_a = (\sum_x C_{ax}) v_a^*$ (with v_a^* as in Section 4.13.1) is required if we want u_c to be the expected number of times a walker starting at a visits node c before it reaches b. This choice implies

$$v_c = \frac{u_c}{\sum_x C_{cx}} \tag{4.65}$$

for all $c \neq a, b$ by virtue of the uniqueness theorem, and using Equations (4.61) and (4.64). In other words, $u_c/(\sum_x C_{cx})$ is then not only proportional to v_c, but identical to v_c for all c.

We now prove that this is the appropriate choice. All we need to check is that the normalisation of the u_c is consistent with the interpretation of u_c as the number of times the walker visits node c. To do this we keep in mind that the walker starts at a and finishes at b. Over the course of the walk returns to node a are possible. The net number of times the walker leaves node a however must be one, given that it starts at a and ends at b (this is the number of times the walker leaves a minus the number of times it arrives at a, not counting the initial placement of the walker at a). If u_c is the number of times a walker visits c during the walk, then the expected net number of departures from node c is given by $\sum_d (u_c T_{cd} - u_d T_{dc})$. Therefore we must have $\sum_c (u_a T_{ac} - u_c T_{ca}) = 1$. This condition is necessary for the correct normalisation of the u_c, and it is also sufficient to verify that the boundary condition $u_a = (\sum_x C_{ax}) v_a^*$ delivers this. This is what we will do next.

The boundary condition $u_a = (\sum_x C_{ax})v_a^*$ leads to Equation (4.65) as explained above. Substituting Equation (4.65) into Ohm's law (Equation (4.59)), we find

$$I_{cd}^{(ab)} = C_{cd}\left(\frac{u_c}{\sum_x C_{cx}} - \frac{u_d}{\sum_x C_{dx}}\right) \tag{4.66}$$

$$= u_c\frac{C_{cd}}{\sum_x C_{cx}} - u_d\frac{C_{dc}}{\sum_x C_{dx}}, \tag{4.67}$$

where, in the second step, we have used $C_{cd} = C_{dc}$. Finally, using Equation (4.8), we find

$$I_{cd}^{(ab)} = u_c T_{cd} - u_d T_{dc}. \tag{4.68}$$

The setup in Section 4.13.1 is such that the current into node a (from the exterior) is equal to one. This means that the total current from node a to all its neighbours in the network is also one, $\sum_c I_{ac}^{(ab)} = 1$. We conclude that $\sum_c(u_a T_{ac} - u_c T_{ca}) = 1$, confirming the correct normalisation of the u_c.

In Section 4.14 we show how to obtain these edge currents analytically.

4.14 Appendix G: Calculating the flow of evidence using the random walk approach

4.14.1 Details of the calculation

The interpretation of the flow of evidence as a random walk can be stated as follows: For the network comparison of treatments a and b, the hat matrix element $H_{cd}^{(ab)}$ that defines the flow of evidence through the direct comparison cd (via Equation (4.6) in the main paper) is equal to the expected *net* number of times a random walker, starting at a on the aggregate NMA network and walking until it reaches b, moves along the edge from c to d.

In Section 4.4.3 of the main paper we demonstrated how to construct a transition matrix for a random walker on the *aggregate network*. For a particular comparison ab, we can use the transition matrix $\boldsymbol{T}^{(ab)}$ to simulate a large ensemble of independent random walkers on the aggregate network starting their journey at a and stopping once they reach b. For each walker we count the number of times it moves across the different network edges in each direction. From this, we find the net number of times

the walker moves along a particular edge. By averaging these values over all of the simulated random walkers, we obtain an estimate of the evidence flow network for this comparison. The more walkers we simulate, the better our estimate of the evidence flow.

By using the analogy between random walks and electrical networks, we can also obtain an analytical result for the evidence flow. To do so we make use of the equations in Section 4.13. First, we apply a 1 volt battery between nodes a and b so that the voltage at a is $v_a = 1$ and at b is $v_b = 0$. With these boundary conditions we then solve the simultaneous equations described by Equation (4.61),

$$v_c = \sum_d v_d T_{cd}, \tag{4.69}$$

to obtain the nodal voltages, v_c, for all nodes $c \neq a, b$. Using Ohm's law, we find the edge currents for the case of a 1 volt battery,

$$I'^{(ab)}_{cd} = C_{cd}(v_c - v_d) = w_{cd}(v_c - v_d), \tag{4.70}$$

these are indicated by $I'^{(ab)}_{cd}$ to distinguish them from the normalised currents $I^{(ab)}_{cd}$ in Section 4.13. In Equation (4.70) we have used the fact that the conductance of edge cd is equal to the aggregate weight associated with that edge, $C_{cd} = w_{cd}$. To make the analogy to evidence flow we require that the total external current flowing into node a is 1. Therefore, to obtain the required currents we must normalise the currents $I'^{(ab)}_{cd}$ by dividing through by the total current flowing into a when $v_a = 1$, that is

$$I^{(ab)}_{cd} = \frac{I'^{(ab)}_{cd}}{\sum_x I'^{(ab)}_{ax}}. \tag{4.71}$$

As shown in Section 4.13, these currents are equal to the expected net number of times a random walker crosses each edge cd. Therefore, from Equation (4.71) we obtain an analytical expression for the evidence flow network in terms of random walkers as follows:

$$H^{(ab)}_{cd} = \overline{N^{(ab)}_{cd}} = \frac{w_{cd}(v_c - v_d)}{\sum_x w_{ax}(v_a - v_x)}, \tag{4.72}$$

and $f^{(ab)}_{cd}$ is obtained from $H^{(ab)}_{cd}$ via Equation (4.6). The potentials v_x are obtained from Equation (4.69).

4.14.2 Implementing the calculation

The above calculation can be written as a linear equation in matrix form. We provide this notation as it is useful for implementation. As above, we focus on the comparison ab in a network of N nodes such that our initial boundary conditions are $v_a = 1$ and $v_b = 0$. From Equation (4.69) we write, for $c \neq a, b$,

$$v_c = \sum_d v_d T_{cd} = T_{ca} + \sum_{d \neq a,b} v_d T_{cd}, \tag{4.73}$$

where we have inserted the known potentials, $v_a = 1$ and $v_b = 0$. Using the fact that $T_{cc} = 0$ (for $c \neq b$) we eliminate the term $d = c$ on the right-hand side, and obtain

$$v_c - \sum_{d \neq a,b,c} v_d T_{cd} = T_{ca} \tag{4.74}$$

for $c \neq a, b$. We collect the potentials v_c, $c \neq a, b$ in a vector $\boldsymbol{v}_{red}$ of length $N - 2$. This is the vector of unknown potentials we wish to calculate. Similarly, we write the transition probabilities T_{ca}, $c \neq a, b$ as an $(N-2)$-vector, $\boldsymbol{T}^{(ab)}_{\cdot a}$. Therefore, we re-write Equation (4.74) in matrix form as

$$(\mathbb{1} - \boldsymbol{T}^{(ab)}_{red})\boldsymbol{v}_{red} = \boldsymbol{T}^{(ab)}_{\cdot a} \tag{4.75}$$

where $\mathbb{1}$ is the $(N-2) \times (N-2)$ identity matrix, and $\boldsymbol{T}^{(ab)}_{red}$ is a reduced version of $\boldsymbol{T}^{(ab)}$ obtained from the full $N \times N$ transition matrix by removing the rows and columns corresponding to nodes a and b.

We then solve this equation for the vector of unknown potentials,

$$\boldsymbol{v}_{red} = (\mathbb{1} - \boldsymbol{T}^{(ab)}_{red})^{-1}\boldsymbol{T}^{(ab)}_{\cdot a}. \tag{4.76}$$

To obtain the full vector $\boldsymbol{v}$ of the potentials at all nodes, we use the fact that $v_a = 1$ and $v_b = 0$ and the entries of $\boldsymbol{v}_{red}$.

The set of potential differences $v_c - v_d$ in Equation (4.70) is then obtained by applying the edge-vertex incidence matrix to the vector of potentials, $\boldsymbol{Bv}$. Finally, multiplying by the weight matrix, $\boldsymbol{W}$, we obtain the vector of non-normalised edge currents (Equation (4.70) in matrix notation),

$$\boldsymbol{I}'^{(ab)} = \boldsymbol{WBv}. \tag{4.77}$$

The normalised currents are then found by dividing through by the total current flowing from node a into the network,

$$\boldsymbol{I}^{(ab)} = \frac{1}{\sum_x I'^{(ab)}_{ax}} \boldsymbol{I}'^{(ab)}. \tag{4.78}$$

4.15 Appendix H: Application to real data

4.15.1 Evidence flow from hat matrix

The edge-vertex incidence matrix for the aggregate network of the depression data set (see Figure 4.6 in the main paper) is

$$\boldsymbol{B} = \begin{array}{c} \\ \text{1-3} \\ \text{1-6} \\ \text{1-7} \\ \text{1-9} \\ \text{1-11} \\ \text{2-6} \\ \text{2-8} \\ \text{2-11} \\ \text{3-4} \\ \text{3-5} \\ \text{3-6} \\ \text{3-9} \\ \text{4-9} \\ \text{5-9} \\ \text{6-7} \\ \text{6-8} \\ \text{6-9} \\ \text{6-11} \\ \text{7-9} \\ \text{7-10} \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \left(\begin{array}{ccccccccccc} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \end{array}\right) \end{array} \tag{4.79}$$

where we have labelled the columns by the treatment and the rows by the direct treatment comparison that they represent. From the depression data (see Rücker and Schwarzer (2014) [39]) we obtain the adjusted weights using the adjustment method for multi-arm trials (see Refs. [7, 39, 42]). Using these weights, and Equations (4.1) and (4.2) in the main paper, we perform a pairwise meta-analysis across each edge. The resulting aggregate weight matrix $\boldsymbol{W}$ is given in Equation (4.81) on page 254. We have labelled the rows and columns by their respective direct treatment comparison. The values are rounded to 3 decimal places.

The hat matrix of the aggregate model is calculated using

$$\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{W}\boldsymbol{B})^+ \boldsymbol{B}^\top \boldsymbol{W}. \tag{4.80}$$

The resulting $\boldsymbol{H}$ matrix for the depression data set is given in Equation (4.82) on page 255. The numerical values for the matrix entries are shown to 3 decimal places. The rows and columns are labelled by the treatment comparison they represent. The first row of the hat matrix refers to the network comparison of treatments 1 and 3. By comparing this row to Figure 4.7 in the main text, it is clear that the evidence flow network defined by the hat matrix is equivalent to the evidence flow network obtained from the random-walk approach.

$$
\boldsymbol{W} = \begin{array}{c}
\begin{array}{cccccccccccccccccc}
\text{1-3} & \text{1-6} & \text{1-7} & \text{1-9} & \text{1-11} & \text{2-6} & \text{2-8} & \text{2-11} & \text{3-4} & \text{3-5} & \text{3-6} & \text{3-9} & \text{4-9} & \text{5-9} & \text{6-7} & \text{6-8} & \text{6-9} & \text{6-11} & \text{7-9} & \text{7-10}
\end{array} \\
\begin{pmatrix}
7.605 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4.432 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.785 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 9.410 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2.322 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 36.419 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 23.576 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 10.474 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13.559 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.118 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.187 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 87.697 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10.533 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 16.946 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.620 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 23.759 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 29.714 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12.154 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.713 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.894
\end{pmatrix}
\end{array}
\tag{4.81}
$$

$$
\boldsymbol{H} = \begin{array}{c}
\\
1\text{-}3 \\ 1\text{-}6 \\ 1\text{-}7 \\ 1\text{-}9 \\ 1\text{-}11 \\ 2\text{-}6 \\ 2\text{-}8 \\ 2\text{-}11 \\ 3\text{-}4 \\ 3\text{-}5 \\ 3\text{-}6 \\ 3\text{-}9 \\ 4\text{-}9 \\ 5\text{-}9 \\ 6\text{-}7 \\ 6\text{-}8 \\ 6\text{-}9 \\ 6\text{-}11 \\ 7\text{-}9 \\ 7\text{-}10
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccccccc}
1\text{-}3 & 1\text{-}6 & 1\text{-}7 & 1\text{-}9 & 1\text{-}11 & 2\text{-}6 & 2\text{-}8 & 2\text{-}11 & 3\text{-}4 & 3\text{-}5 & 3\text{-}6 & 3\text{-}9 & 4\text{-}9 & 5\text{-}9 & 6\text{-}7 & 6\text{-}8 & 6\text{-}9 & 6\text{-}11 & 7\text{-}9 & 7\text{-}10
\end{array} \\
\begin{pmatrix}
0.353 & 0.152 & 0.044 & 0.380 & 0.072 & 0.022 & 0.007 & -0.030 & -0.036 & -0.024 & -0.062 & -0.526 & -0.036 & -0.024 & -0.016 & -0.007 & 0.178 & -0.042 & 0.027 & 0.000 \\
0.261 & 0.236 & 0.052 & 0.339 & 0.111 & 0.035 & 0.011 & -0.046 & 0.010 & 0.007 & 0.098 & 0.146 & 0.010 & 0.007 & -0.040 & -0.011 & -0.514 & -0.065 & 0.012 & 0.000 \\
0.185 & 0.128 & 0.381 & 0.245 & 0.060 & 0.019 & 0.006 & -0.025 & 0.0098 & 0.007 & 0.024 & 0.145 & 0.010 & 0.007 & 0.299 & -0.006 & -0.086 & -0.035 & -0.321 & 0.000 \\
0.307 & 0.160 & 0.046 & 0.412 & 0.075 & 0.024 & 0.008 & -0.031 & 0.020 & 0.013 & -0.022 & 0.296 & 0.020 & 0.013 & -0.016 & -0.008 & 0.229 & -0.044 & 0.030 & 0.000 \\
0.235 & 0.213 & 0.046 & 0.305 & 0.201 & -0.250 & -0.081 & 0.331 & 0.009 & 0.006 & 0.089 & 0.132 & 0.009 & 0.006 & -0.036 & 0.081 & -0.463 & 0.468 & 0.011 & 0.000 \\
0.005 & 0.004 & 0.001 & 0.006 & -0.016 & 0.671 & 0.218 & 0.111 & 0.000 & 0.000 & 0.002 & 0.003 & 0.000 & 0.000 & -0.001 & -0.218 & -0.009 & -0.095 & 0.000 & 0.000 \\
0.002 & 0.002 & 0.000 & 0.003 & -0.008 & 0.337 & 0.607 & 0.056 & 0.000 & 0.000 & 0.001 & 0.001 & 0.000 & 0.000 & -0.000 & 0.393 & -0.005 & -0.048 & 0.000 & 0.000 \\
-0.022 & -0.020 & -0.004 & -0.028 & 0.073 & 0.386 & 0.125 & 0.489 & -0.001 & -0.001 & -0.008 & -0.012 & -0.001 & -0.001 & 0.003 & -0.125 & 0.042 & 0.438 & -0.001 & 0.000 \\
-0.020 & 0.003 & 0.001 & 0.014 & 0.002 & 0.000 & 0.000 & -0.001 & 0.587 & 0.016 & 0.017 & 0.359 & -0.413 & 0.016 & 0.000 & -0.000 & 0.022 & -0.001 & 0.001 & 0.000 \\
-0.035 & 0.006 & 0.002 & 0.024 & 0.003 & 0.001 & 0.000 & -0.001 & 0.043 & 0.260 & 0.031 & 0.631 & 0.043 & -0.740 & 0.000 & -0.000 & 0.039 & -0.002 & 0.002 & 0.000 \\
-0.091 & 0.084 & 0.008 & -0.041 & 0.040 & 0.012 & 0.004 & -0.016 & 0.045 & 0.030 & 0.161 & 0.673 & 0.045 & 0.030 & -0.023 & -0.004 & -0.692 & -0.023 & -0.015 & 0.000 \\
-0.046 & 0.007 & 0.003 & 0.032 & 0.003 & 0.001 & 0.000 & -0.001 & 0.056 & 0.037 & 0.039 & 0.822 & 0.056 & 0.037 & 0.000 & -0.000 & 0.051 & -0.002 & 0.003 & 0.000 \\
-0.026 & 0.004 & 0.002 & 0.018 & 0.002 & 0.001 & 0.000 & -0.001 & -0.532 & 0.021 & 0.022 & 0.463 & 0.468 & 0.021 & 0.000 & -0.000 & 0.029 & -0.001 & 0.002 & 0.000 \\
-0.012 & 0.002 & 0.001 & 0.007 & 0.001 & 0.000 & 0.000 & -0.000 & 0.013 & -0.223 & 0.009 & 0.191 & 0.013 & 0.777 & 0.000 & 0.000 & 0.012 & -0.000 & 0.001 & 0.000 \\
-0.076 & -0.108 & 0.329 & -0.094 & -0.051 & -0.016 & -0.005 & 0.021 & 0.000 & 0.000 & -0.075 & -0.001 & 0.000 & 0.000 & 0.338 & 0.005 & 0.428 & 0.030 & -0.333 & 0.000 \\
-0.002 & -0.002 & -0.000 & -0.003 & 0.008 & -0.334 & 0.389 & -0.055 & 0.000 & 0.000 & -0.001 & -0.001 & 0.000 & 0.000 & 0.000 & 0.611 & 0.005 & 0.047 & -0.000 & 0.000 \\
0.046 & -0.077 & -0.005 & 0.072 & -0.036 & -0.011 & -0.004 & 0.015 & 0.010 & 0.007 & -0.121 & 0.150 & 0.010 & 0.007 & 0.023 & 0.004 & 0.743 & 0.021 & 0.018 & 0.000 \\
-0.026 & -0.024 & -0.005 & -0.034 & 0.089 & -0.285 & -0.093 & 0.377 & -0.001 & -0.001 & -0.010 & -0.015 & -0.001 & -0.001 & 0.004 & 0.093 & 0.052 & 0.533 & -0.001 & 0.000 \\
0.121 & 0.031 & -0.334 & 0.166 & 0.015 & 0.005 & 0.002 & -0.006 & 0.010 & 0.007 & -0.046 & 0.151 & 0.010 & 0.007 & -0.315 & -0.002 & 0.315 & -0.009 & 0.351 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000
\end{pmatrix}
\end{array}
\tag{4.82}
$$

4.16 Appendix I: Implementation in netmeta

Here we outline how to use the updated `netmeta` package to obtain the results in this manuscript.

To obtain the aggregate hat matrix from Equation (4.5) for a `netmeta` object `net1` use:

```
HatAgg <- hatmatrix(net1, method = "Davies", type = "short")
```

then the hat matrix for the fixed effect model is `HatAgg$fixed` and for the random effect model it is `HatAgg$random`.

To obtain the proportion contribution matrix using the random-walk method use:

```
cont.rw <- netcontrib(net1, method = "randomwalk",
                                          hatmatrix.F1000 = FALSE)
```

The argument `hatmatrix.F1000 = FALSE` is the default but we include it here for transparency. As before, the fixed effect result is obtained from `cont.rw$fixed` and the random effect result from `cont.rw$random`.

To obtain the proportion contribution matrix using the Shortest algorithm (from Papakonstantinou et al (2018) [9]) use:

```
cont.sp <- netcontrib(net1, method = "shortestpath",
                                          hatmatrix.F1000 = FALSE)
```

Again, the fixed effect result is obtained from `cont.sp$fixed`, the random effect result from `cont.sp$random` and `hatmatrix.F1000 = FALSE` is the default argument.

If you wish to obtain results consistent with the original implementation of the algorithm in Papakonstantinou et al (2018) [9] set the argument `hatmatrix.F1000 = TRUE` as this uses the hat matrix which doesn't take into account multi-arm trials. This is not recommended in general but may be useful for reproducibility.

Bibliography

[1] 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*, Online, Last updated September 2016; Available from http://www.nicedsu.org.uk. Accessed March 2020, 2011.

[2] S. Dias, A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton, *Network meta-analysis for decision making* (Wiley, Oxford, UK, 2018).

[3] G. Salanti, "Indirect and mixed treatment comparison, network, or multiple treatments meta analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool", Res. Synth. Meth. **3**, 80–97 (2012).

[4] G. Lu and A. E. Ades, "Combination of direct and indirect evidence in mixed treatment comparisons", Stat. Med. **23**, 3105–3124 (2004).

[5] J. P. T. Higgins and A. Whitehead, "Borrowing strength from external trials in a meta-analysis", Stat. Med. **15**, 2733–2749 (1996).

[6] T. Lumley, "Network meta-analysis for indirect treatment comparisons", Stat. Med. **21**, 2313–2324 (2002).

[7] G. Rücker, "Network meta-analysis, electrical networks and graph theory", Res. Synth. Meth. **3**, 312–324 (2012).

[8] J. König, U. Krahn, and H. Binder, "Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons", Stat. Med. **32**, 5414–5429 (2013).

[9] T. Papakonstantinou, A. Nikolakopoulou, G. Rücker, A. Chaimani, G. Schwarzer, M. Egger, and G. Salanti, "Estimating the contribution of studies in network meta-analysis: paths, flows and streams", F1000Res. **7**, 610 (2018).

[10] A. Nikolakopoulou, J. P. T. Higgins, T. Papakonstantinou, A. Chaimani, C. Del Giovane, M. Egger, and G. Salanti, "CINeMA: an approach for assessing confidence in the results of a network meta-analysis", PLOS Med. **17**, e1003082 (2020).

[11] G. Lu, N. J. Welton, J. P. T. Higgins, I. White, and A. E. Ades, "Linear inference for mixed treatment comparison meta-analysis: a two-stage approach", Res. Synth. Meth. **2**, 43–60 (2011).

[12] S. Senn, F. Gavini, D. Magrez, and A. Scheen, "Issues in performing a network meta-analysis", Stat. Methods. Med. Res. **22**, 169–189 (2012).

[13] G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis, "Evaluation of networks of randomized trials", Stat. Methods. Med. Res. **17**, 279–301 (2008).

[14] G. Salanti, F. K. Kavvoura, and J. P. A. Ioannidis, "Exploring the geometry of treatment networks", Ann. Intern. Med. **148**, 544–553 (2008).

[15] A. L. Davies and T. Galla, "Degree irregularity and rank probability bias in network meta-analysis", Res. Synth. Meth. **12**, 316–332 (2021).

[16] F. S. Tonin, H. H. Borba, A. M. Mendes, A. Wiens, F. Fernandez-Llimos, and R. Pontarolo, "Description of network meta-analysis geometry: a metrics design study", PLOS ONE **14**, e0212650 (2019).

[17] A. A. Veroniki, S. E. Straus, G. Rücker, and A. C. Tricco, "Is providing uncertainty intervals in treatment ranking helpful in network meta-analysis?", J. Clin. Epidemiol. **100**, 122–129 (2018).

[18] T. Kibret, D. Richer, and J. Bayene, "Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study", Clin. Epidemiol. **6**, 451–460 (2014).

[19] V. Chiocchia, A. Nikolakopoulou, J. P. T. Higgins, M. J. Page, T. Papakonstantinou, A. Cipriani, T. A. Furukawa, G. C. M. Siontis, M. Egger, and G. Salanti, "ROB-MEN: a tool to assess risk of bias due to missing evidence in network meta-analysis", BMC Med. **19**, 304 (2021).

[20] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of networks: from biological networks to the internet and WWW* (Oxford University Press, Oxford, UK, 2003).

[21] M. Newman, *Networks*, 2nd ed. (Oxford University Press, Oxford, UK, 2018).

[22] E. Estrada, *The structure of complex networks: theory and applications* (Oxford University Press, Oxford, UK, 2011).

[23] G. Rücker and G. Schwarzer, "Ranking treatments in frequentist network meta-analysis works without resampling methods", BMC Med. Res. Methodol. **15**, 58 (2015).

[24] G. Rücker, M. Petropoulou, and G. Schwarzer, "Network meta-analysis of multicomponent interventions", Biometrical J. **62**, 808–821 (2020).

[25] G. Rücker, U. Krahn, J. König, O. Efthimiou, A. Davies, T. Papakonstantinou, and G. Schwarzer, *Netmeta: network meta-analysis using frequentist methods*, R package version 2.0-0. https://CRAN.R-project.org/package=netmeta, R Foundation for Statistical Computing (Vienna, Austria, 2021).

[26] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology", J. R. Soc. Interface **5**, 5813–834 (2008).

[27] A. Okubo and S. A. Levin, *Diffusion and ecological problems: modern perspectives*, 2nd ed. (Springer, New York, NY, USA, 2001).

[28] M. B. Isichenko, "Percolation, statistical topography, and transport in random media", Rev. Mod. Phys. **64**, 961–1043 (1992).

[29] W. J. Ewens, *Mathematical population genetics I. Theoretical introduction* (Springer, New York, NY, USA, 2010).

[30] R. N. Mantegna and H. E. Stanley, *An introduction to econophysics* (Cambridge University Press, Cambridge, UK, 1999).

[31] J. D. Noh and H. Rieger, "Random walks on complex networks", Phys. Rev. Lett. **92**, 118701 (2004).

[32] L. Lovász, *Random walks on graphs: a survey*, Online, YALE/DCS/TR-1029. Available from http://www.cs.yale.edu/publications/techreports/tr1029.pdf. Accessed March 2021, 1994.

[33] N. Masuda, M. A. Porter, and R. Lambiotte, "Random walks and diffusion on networks", Phys. Rep. **716-717**, 1–58 (2017).

[34] S. Kakutani, "Markov processes and the Dirichlet problem", Proc. Jap. Acad. **21**, 227–233 (1945).

[35] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Markov chains*, University Series in Higher Mathematics (Van Nostrand, New York, NY, UK, 1966).

[36] F. P. Kelly, *Reversibility and stochastic networks*, Probability and Statistics Series (Wiley, Chichester, UK, 1979).

[37] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, arXiv:math/0001057, 2000.

[38] K. Linde, L. Kriston, G. Rücker, and A. Schneider, *Treatment of depressive disorders in primary care — a multiple treatment systematic review of randomized controlled trials*, Online, Available from http://edok01.tib.uni-hannover.de/edoks/e01fb13/772211906.pdf. Accessed April 2021, 2013.

[39] G. Rücker and G. Schwarzer, "Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis", Stat. Med. **33**, 4353–4369 (2014).

[40] O. Efthimiou, T. P. Debray, G. van Valkenhoef, S. Trelle, K. Panayidou, K. G. Moons, J. B. Reitsma, A. Shang, and G. Salanti, "GetReal methods review group. GetReal in network meta-analysis: a review of the methodology", Res. Synth. Meth. **7**, 236–263 (2016).

[41] D. Jackson, I. R. White, and T. G. Simon, "Extending Dersimonian and Laird's methodology to perform multivariate random effects meta-analyses", Stat. Med. **29**, 1282–1297 (2010).

[42] I. Gutman and W. Xiao, "Generalized inverse of the Laplacian matrix and some applications", Bull. Acad. Serbe. Sci. Cl. Sci. Math. Nat. Sci. Nat. **129**, 15–23 (2004).

[43] U. Krahn, H. Binder, and J. König, "A graphical tool for locating inconsistency in network meta-analyses", BMC Med. Res. Methodol. **13**, 35 (2013).

[44] M. Urbano, "Kirchhoff's laws", in *Introductory electrical engineering with math explained in accessible language* (John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2019), pp. 197–213.

[45] T. Papakonstantinou, A. Nikolakopoulou, J. P. T. Higgins, M. Egger, and G. Salanti, "CINeMA: software for semiautomated assessment of the confidence in the results of network meta-analysis", Campbell Syst. Rev. **16**, e1080 (2020).

[46] C. A. Macfadyen, J. M. Acuin, and C. Gamble, "Topical antibiotics without steroids for chronically discharging ears with underlying eardrum perforations", Cochrane DB. Syst. Rev. **4**, CD004618 (2005).

[47] G. Csardi and T. Nepusz, "The igraph software package for complex network research", InterJournal Complex Systems **1695**, https://igraph.org (2006).

[48] technical-recipes.com, *A recursive algorithm to find all paths between two given nodes in C++ and C*, Online, Available from https://www.technical-recipes.com/2011/a-recursive-algorithm-to-find-all-paths-between-two-given-nodes/. Accessed December 2020, 2011.

[49] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", BMJ **372**, n71 (2021).

[50] G. Rücker, A. Nikolakopoulou, T. Papakonstantinou, G. Salanti, R. D. Riley, and G. S. Schwarzer, "The statistical importance of a study for a network meta-analysis estimate", BMC Med. Res. Methodol. **20**, 1–13 (2020).

[51] G. S. Ohm, *Die galvanische Kette, mathematisch bearbeitet* (T. H. Riemann, Berlin, Germany, 1827).

[52] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, 3rd ed. (Springer-Verlag, New York, NY, USA, 2002), p. 245.

[53] S. Axler, P. Bourdon, and W. Ramey, *Harmonic function theory* (Springer, New York, NY, USA, 2001).

Chapter 5

Introduction to Survival Analysis

The research project to be presented in Chapter 6 is on the topic of survival analysis. In this chapter we give a short introduction to some of the main concepts in survival analysis in order to bridge the gap between the projects. A number of relevant technical concepts, such as linear regression models and maximum likelihood, have already been introduced in the context of network meta-analysis. Most of the required mathematical details are provided in Chapter 6 and its accompanying appendices.

We do not aim to give a comprehensive overview of survival analysis in this chapter. Instead we present the main concepts that are relevant to the work in Chapter 6.

5.1 Time-to-event data

Survival analysis refers to a collection of statistical techniques for the analysis of data where the outcome of interest is the time until an event occurs [1–3]. This type of analysis appears in a number of areas such as engineering [4] (where one is interested in the lifetime of industrial components), and economics [5] (where events of interest include job acquisition, retirement or failure of a business). We focus on survival analysis in the context of medical research. Here, the subjects under study are patients with a particular condition, and the typical outcome is death or some other event connected to the condition such as a heart attack, stroke or relapse of cancer. The event times of each individual in the data set ($i = 1, \ldots, N$) are considered to be independent of each other. These observed event times are used to fit a survival model. A key aim of survival analysis is then to make predictions about the probability that

some patient or patients who are not part of data set used to fit the model survive to a particular time.

5.2 Censoring

A key feature of survival data is that not all individuals in the sample are observed to experience the event during the course of the study. Therefore, the event times of these individuals are unknown and the data is said to be 'censored'. The most common type of censoring is so-called 'right censoring'. This can occur in a number of ways, for example (i) studies do not proceed indefinitely and usually have a set end-date, meaning an individual may not have experienced the event by this date, (ii) the individual may be lost to follow-up (i.e. the researchers lose contact with the individual over the course of the study, perhaps due to withdrawal from the trial), or (iii) the individual experiences another event that makes further follow-up impossible (e.g. if the event of interest is a heart attack but the patient dies from some other unrelated cause before the end of the study) [1, 2]. With this type of censoring we have a lower-bound estimate for the event time. That is, we know that the true event time is later than the censoring time.

Other mechanisms for censored data include 'left' and 'interval' censoring. Left censoring is the least common in practice and refers to data in which the event occurs before the period of observation [6]. For example, if one is interested in the time it takes students in a class to learn a particular task, those who already know how to do the task before the beginning of the study are left censored. In the case of interval censored data, the available information is that the event occurred within a certain time window. For example, if the event is the relapse of a cancer, occurrence of the event can only be identified at discrete follow-up appointments (i.e. when a doctor makes a diagnosis). This means that once recurrence is observed, all we know is that the event time was some point between the previous appointment and the current one. The different types of censoring are illustrated in Figure 5.1.

Standard survival analysis methods assume that the censoring mechanism is *non-informative* [1, 3]. In other words, the fact that an individual is censored contains no information about the subsequent survival of that patient. *Informative* censoring may

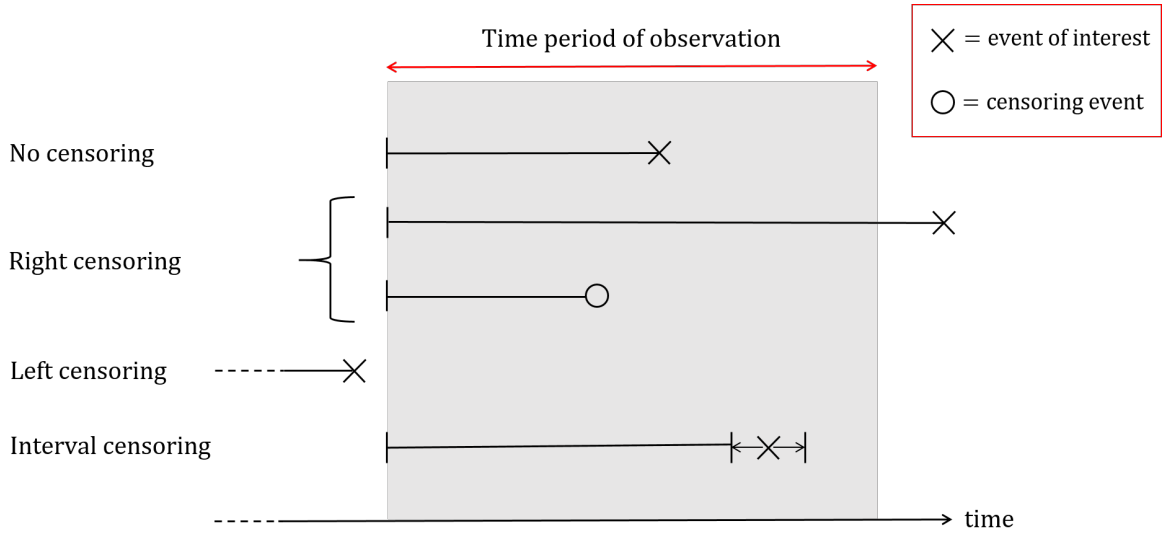


Figure 5.1: An illustration of the different censoring mechanisms. The individual experiences no censoring if the event of interest is observed within the time frame of the study. Right censoring occurs either when the event happens after the period of observation, or if the individual experiences a censoring event (such as withdrawal from the trial) during the study. An individual is left censored if the event happens before the period of observation. Interval censoring occurs when the event status is only observed at discrete times. In this scenario, the event time is not observed precisely but is known to occur within a certain time window.

occur, for example, if a patient withdraws from the study due to a worsening of the condition.

Censoring is a characteristic feature of time-to-event data that calls for specialised methods of analysis [1, 7]. Even when the censoring mechanism is said to be 'non-informative' the censoring time carries information about the survival of the patient. Therefore, in survival models individuals for whom the outcome is not observed are still involved in the analysis [6]. In our work we consider non-informative right censoring only.

5.3 Survival and hazard function

Of primary interest in survival analysis is the survival function, $S(t)$. For some specified time t relative to the time origin ($t = 0$), this is defined as the probability of survival to at least time t. Equivalently, it is the probability that the event occurs after time t,

$$S(t) = P(T > t), \tag{5.1}$$

where T is a random variable representing the event time. Fitting a survival function to the time-to-event data provides a crucial summary of the survival of the study cohort

which can then be used to make survival predictions about other individuals. The survival function has the following properties; (i) the probability of survival at the time origin is one, $S(0) = 1$, (ii) it is a non-increasing function, $S(t') \leq S(t)$ if $t' \geq t$, and (iii) the probability of survival approaches zero for large times, $\lim_{t \to \infty} S(t) = 0$ (i.e. if uncensored, everyone will eventually experience the event) [2, 3]. We note that (i) and (ii) are mathematical necessities, whereas (iii) is an assumption. For the standard approaches described here, it is also assumed that each patient can only experience the event once. An example of a survival function is shown in Figure 5.2 (b).

Another important function related to survival is the hazard rate, $h(t)$. This is defined as the probability per unit time that the event occurs at time t, given that no event has occurred prior to this time,

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \right]. \tag{5.2}$$

It can be thought of as the instantaneous event rate for an individual who has already survived to time t [1, 2].

While the hazard describes the instantaneous potential of an event occurring, the survival function is the cumulative probability that no event has happened by a given time. The two are intrinsically related via

$$S(t) = \mathrm{e}^{-\int_0^t h(t')\mathrm{d}t'}. \tag{5.3}$$

The hazard function is then a vehicle through which we can model the survival function [1]. We derive this expression in the following section.

5.3.1 The relation between survival function and hazard rate

To derive the relation in Equation (5.3), we define the probability that the event occurs by time t as $F(t) = P(T \leq t) = 1 - S(t)$. Defining $p(t)$ as the probability density of event times, we can write

$$F(t) = \int_0^t p(t')\mathrm{d}t'. \tag{5.4}$$

We then turn to the numerator in Equation (5.2). Using the expression for conditional probabilities, $P(A|B) = \frac{P(A,B)}{P(B)}$, we find

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{P(t < T \leq t + \Delta t)}{\Delta t \; S(t)} \right], \tag{5.5}$$

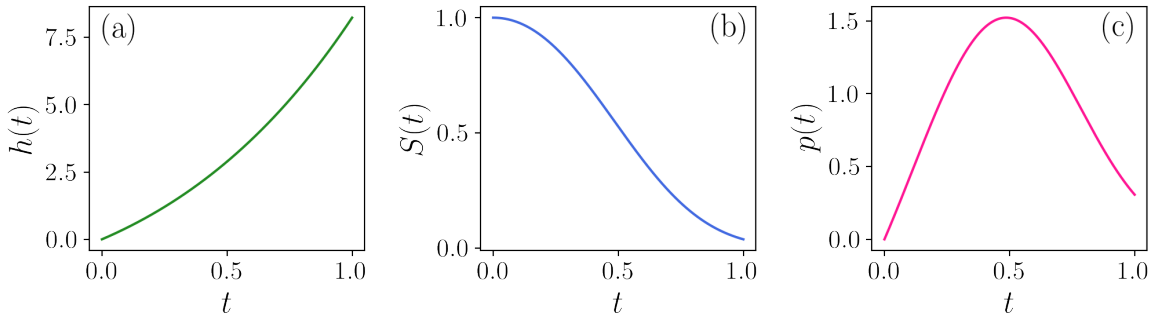


Figure 5.2: (a) An example of a (fictional) hazard function $h(t)$ with (b) its corresponding survival function $S(t)$, and (c) its probability density of event times, $p(t)$.

where we have used $P(t < T \leq t+\Delta t, T > t) = P(t < T \leq t+\Delta t)$ and $P(T > t) = S(t)$. Evaluating the numerator in Equation (5.5) gives

$$\begin{aligned} P(t < T \leq t + \Delta t) &= \int_t^{t+\Delta t} p(t')\mathrm{d}t' \\ &= \int_0^{t+\Delta t} p(t')\mathrm{d}t' - \int_0^t p(t')\mathrm{d}t' \\ &= F(t + \Delta t) - F(t), \end{aligned} \tag{5.6}$$

such that

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t} \right] \frac{1}{S(t)} = \frac{\mathrm{d}F(t)}{\mathrm{d}t} \frac{1}{S(t)}. \tag{5.7}$$

Given Equations (5.4) and (5.7), the distribution of event times is related to the survival and hazards functions via $p(t) = h(t)S(t)$. An illustration of these three functions is shown in Figure 5.2 for a particular (fictional) hazard function $h(t)$.

Using $F(t) = 1 - S(t)$ in Equation (5.7) leads to

$$h(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} \frac{1}{S(t)}, \tag{5.8}$$

which we can write as

$$h(t) = -\frac{\mathrm{d} \ln S(t)}{\mathrm{d}t}. \tag{5.9}$$

Rearranging this expression for $S(t)$ (and using the fact that $S(0) = 1$) leads directly to Equation (5.3).

5.4 Non-parametric survival model: Kaplan-Meier

A central aim in survival analysis is to obtain an estimate of the survival function from the (censored) time-to-event data. A simple non-parametric approach is the

so-called 'Kaplan-Meier' or 'product limit' estimator [8]. Essentially, one approximates the survival curve with a step function that changes value at every observed event time T_i. The estimator is written as

$$\hat{S}(t) = \prod_{i:T_i \leq t} (1 - \frac{r_i}{n_i}), \tag{5.10}$$

where r_i is the number of (non-censoring) events that occur at T_i, and n_i is the number of individuals in the sample known to be event free (i.e. have not yet been censored or experienced an event) by time T_i [3, 8].

There is clinical interest in investigating the effect of certain variables on survival probability. A categorical variable is one that takes only a finite number of distinct values. Individuals can then be categorised into groups according to this variable. For data of this type, one can construct different Kaplan-Meier curves for each group and compare their characteristics [1]. Often this is done for two treatment groups to investigate whether one treatment produces more preferable survival outcomes than another.

In this approach, comparisons are usually made via a statistical test that indicates whether or not there is a significant difference between the groups [3, 9]. However, this does not provide a quantitative summary of the extent of this difference. For example, we can conclude that treatment A has better survival outcomes than treatment B but we cannot say by how much [10]. Moreover, this approach only considers one characteristic of the cohort while in reality there are likely to be a number of patient characteristics that affect survival. This can be particularly problematic if the groups under investigation systematically differ in relation to some other medically relevant variable. For example, the patients in one treatment group may be older than those in the second group. This means that any observed differences in survival outcomes between the groups could be due to age rather than treatment. In this scenario age is a 'confounding variable' [1].

To control for confounding variables and to quantify the extent of their effect on survival, more sophisticated methods are required. The work in Chapter 6 builds on the celebrated Cox proportional hazards regression model. We introduce this model in the next section.

5.5 Semi-parametric survival model: Cox proportional hazards

The proportional hazards (PH) model introduced by Cox in 1972 [11] remains the most widely used method for modelling time-to-event data. It is a regression model that describes how the hazard rate, $h(t)$, depends on a set of p predictor variables z_μ, $\mu = 1, \ldots, p$. The survival function is then specified via Equation (5.3) and can be used to make predictions. The predictor variables are called *covariates* in survival analysis and refer to both continuous and categorical patient-specific characteristics measured at some time origin ($t = 0$). Typical covariates include age, gender, treatment group and measurements of so-called 'biomarkers'. A biomarker (or 'biological marker') refers to a measurable indicator of a biological system that provides insight into its physiological state. Examples include measurements of weight and blood pressure, as well as laboratory tests of samples such as blood, urine and biological tissue.

The Cox model is a linear regression on the *logarithm* of the hazard rate and is written as,

$$h(t) = h_0(t)\mathrm{e}^{\sum_{\mu=1}^{p}\beta_\mu z_\mu}, \tag{5.11}$$

where the regression coefficients β_μ are referred to as the *association* parameters. These quantify the relationship between each covariate and the hazard, e.g. if $\beta_\mu > 0$ then covariate z_μ is positively correlated with the hazard rate and thus has a negative association with survival. The larger β_μ, the stronger this association.

The function $h_0(t)$ in Equation (5.11) is the base hazard rate. This is the value the hazard takes if all covariates are zero, $z_\mu = 0$, $\forall\mu$. The log base hazard rate is the time varying intercept of the regression model on $\ln h(t)$. A key feature of Cox's PH model is that it makes no assumptions about the form of $h_0(t)$, and the function is estimated non-parametrically. This is particularly appealing as it means that the survival times are not assumed to follow a particular distribution thus allowing for extra flexibility [10, 12]. Because of this feature, the model is said to be semi-parametric.

As shown in Equation (5.11), the covariates are modelled such that they act multiplicatively on the hazard function. This is the 'proportional hazards' assumption; the hazard rate of the event in any particular covariate group is a constant multiple of

the hazard in any other [10]. Consider a particular covariate z_μ and assume all other covariates z_ν, $\nu \neq \mu$ are fixed. The ratio of hazards (HR) for two values $z_\mu = z'$ and $z_\mu = z''$ is then

$$\text{HR} = \frac{h_0(t)\mathrm{e}^{\beta_\mu z' + \sum_{\nu\neq\mu}\beta_\nu z_\nu}}{h_0(t)\mathrm{e}^{\beta_\mu z'' + \sum_{\nu\neq\mu}\beta_\nu z_\nu}} = \mathrm{e}^{\beta_\mu(z'-z'')}. \tag{5.12}$$

Hazard ratios are therefore characterised by the variables e^{β_μ}. For binary covariates (i.e. $z' = 1$ and $z'' = 0$), these variables are equal to the hazard ratios. The association parameters β_μ then define the *log hazard ratios*. If the proportional hazards assumption holds, plotting the function $h(t)$ against t for two covariate groups should lead to curves that do not cross at any point.

Estimating the parameters β_μ provides insight into the association between different covariates and survival. It also allows the survival model to be fitted so that we can make predictions about new individuals for whom we have covariate measurements. The standard method for obtaining these estimates is a maximum-likelihood approach.

5.5.1 Likelihood function for censored data

To work out the likelihood function for a censored data set one must consider what information is available about the individuals in the sample. For each individual, one either observes their true event time T_i^*, or some non-informative censoring time, C_i. The observed event time is then $T_i = \min(T_i^*, C_i)$. The event indicator $\delta_i = I(T_i^* \leq C_i)$ is defined such that it equals 1 when the event is observed ($T_i = T_i^*$) and 0 when it is censored ($T_i = C_i$).

In Chapter 6 we write the likelihood as $L(\theta|\mathcal{D}) = \mathcal{P}(\mathcal{D}|\theta)$ where $\mathcal{D}$ represents the data and θ is the set of model parameters we wish to infer. For the standard Cox PH model described here, $\mathcal{D} = \{T_i, \delta_i, z_\mu^i; i = 1, \ldots, N, \mu = 1, \ldots, p\}$ and $\theta = \{h_0(t), \beta_\mu; \mu = 1, \ldots, p\}$. We have written z_μ^i for the value of covariate μ associated with individual i.

The distribution of the true event times is governed by the survival function $S(t|\theta, z_\mu^i) = P(T_i^* > t|\theta, z_\mu^i)$ and corresponding density function $p(t|\theta, z_\mu^i)$ which depend on the parameters θ and covariates z_μ^i via the parameterisation of the hazard function in Equation (5.11). For non-informative censoring, the censoring times C_i are assumed to be stochastically independent from each other and the true event times. Their

distribution is governed by a survival function $P(C_i > t)$ and corresponding density function that *do not* depend on the model parameters θ [3].

Each individual in the data set contributes to the likelihood function. The nature of this contribution depends on whether we observe the event of interest or a censoring event for that individual. The likelihood describes the conditional probability (or probability density) of observing the data given a specific set of model parameters θ. Therefore, any information that does not depend on θ does not contribute to the likelihood. We now explain the contributions of observed (non-censored) and censored individuals in turn.

For individuals who are *not* subject to censoring, one knows (i) their precise event time, $T_i^* = T_i$, and (ii) that their censoring time is later than this time, $C_i > T_i$. Item (i) corresponds to the probability density $p(T_i|\theta, z_\mu^i)$, and item (ii) corresponds to the censoring survival function $P(C_i > T_i)$ (with no θ dependence). The contribution to the likelihood is then [3]

$$L_i^{\text{event}}(\theta|\mathcal{D}) = p(T_i|\theta, z_\mu^i) = S(T_i|\theta, z_\mu^i)h(T_i|\theta, z_\mu^i). \tag{5.13}$$

On the other hand, the available information for censored individuals is their precise censoring time $C_i = T_i$ and that their true event time exceeds this time, $T_i^* > T_i$. This information corresponds to the censoring probability density (which does not depend on θ), and the survival function $S(t|\theta, z_\mu^i) = P(T_i^* > t|\theta, z_\mu^i)$. The contribution of these individuals to the likelihood is then [3]

$$L_i^{\text{cens}}(\theta|\mathcal{D}) = S(T_i|\theta, z_\mu^i). \tag{5.14}$$

Both observed and censored individuals survive up to time T_i meaning both contributions to the likelihood contain the survival function $S(T_i|\theta, z_\mu^i)$. For observed samples, the additional contribution of the instantaneous hazard corresponds to the extra information that the event occurred at T_i.

The likelihood function of the whole data set is then [3]

$$\begin{aligned} L(\theta|\mathcal{D}) &= \prod_{i=1}^{N} L_i(\theta|\mathcal{D}) \\ &= \prod_{i=1}^{N} L_i^{\text{event}}(\theta|\mathcal{D})^{\delta_i} L_i^{\text{cens}}(\theta|\mathcal{D})^{(1-\delta_i)} \\ &= \prod_{i=1}^{N} h(T_i|\theta, z_\mu^i)^{\delta_i} S(T_i|\theta, z_\mu^i). \end{aligned} \tag{5.15}$$

Writing the survival function in terms of the hazard via Equation (5.3) leads to the log likelihood,

$$\ln L(\theta|\mathcal{D}) = \sum_{i=1}^{N} \delta_i \ln h(T_i|\theta, z_\mu^i) - \sum_{i=1}^{N} \int_0^{T_i} h(t|\theta, z_\mu^i) \mathrm{d}t. \tag{5.16}$$

To infer the parameters in the Cox model, we insert the semi-parameterised hazard function [Equation (5.11)] into Equation (5.16) and find

$$\ln L(\theta|\mathcal{D}) = \sum_{i=1}^{N} \delta_i \ln h_0(T_i) + \sum_{i=1}^{N} \delta_i \sum_{\mu=1}^{p} \beta_\mu z_\mu^i - \sum_{i=1}^{N} \int_0^{T_i} h_0(t) \mathrm{e}^{\sum_{\mu=1}^{p} \beta_\mu z_\mu^i} \mathrm{d}t. \tag{5.17}$$

The task now is to maximise the log likelihood with respect to the model parameters θ. We begin with the base hazard. Taking the functional derivative of Equation (5.17) with respect to $h_0(t)$ and setting it equal to zero yields an expression for the base hazard in terms of the association parameters β_μ,

$$h_0(t) = \frac{\sum_{i=1}^{N} \delta_i \delta(t - T_i)}{\sum_{i=1}^{N} I(t \in [0, T_i]) \mathrm{e}^{\sum_{\mu=1}^{p} \beta_\mu z_\mu^i}}. \tag{5.18}$$

This expression, known as the 'Breslow estimator'[1], was first derived by Breslow (1972) [13] in his discussion of Cox's original paper [14].

Inserting the Breslow estimator back into Equation (5.17) we obtain an expression for the log likelihood,

$$\begin{aligned}
\ln L(\theta|\mathcal{D}) &= \sum_i \delta_i \ln \left[\frac{\sum_k \delta_k \delta(T_i - T_k)}{\sum_j I(T_i \in [0, T_j]) \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^j}} \right] + \sum_{i=1}^{N} \delta_i \sum_{\mu=1}^{p} \beta_\mu z_\mu^i \\
&\qquad - \sum_{i=1}^{N} \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^i} \int_0^{T_i} \frac{\sum_k \delta_k \delta(t - T_k)}{\sum_j I(t \in [0, T_j]) \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^j}} \mathrm{d}t \\
&= -\sum_i \delta_i \ln \left[\sum_j I(T_i \in [0, T_j]) \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^j} \right] + \sum_{i=1}^{N} \delta_i \sum_{\mu=1}^{p} \beta_\mu z_\mu^i \\
&\qquad - \sum_{i=1}^{N} \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^i} \sum_k \delta_k \frac{I(T_k \in [0, T_i])}{\sum_j I(T_k \in [0, T_j]) \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^j}} + \text{const} \\
&= -\sum_i \delta_i \ln \left[\sum_j I(T_i \in [0, T_j]) \mathrm{e}^{\sum_\mu \beta_\mu z_\mu^j} \right] + \sum_{i=1}^{N} \delta_i \sum_{\mu=1}^{p} \beta_\mu z_\mu^i + \text{const},
\end{aligned} \tag{5.19}$$

where we have written 'const' for terms that are constant in β_μ. The log likelihood in Equation (5.19) is a function of the data and association parameters only. To fully specify the hazard and survival functions for the Cox PH model, one must now maximise this expression with respect to the parameters β_μ. This requires numerical methods. In Chapter 6 we make use of Powell minimisation.

[1] The term 'Breslow estimator' is also used to refer to the estimate of the cumulative base hazard $H_0(t) = \int_0^t h_0(t')\mathrm{d}t'$ with $h_0(t)$ given in Equation (5.18).

5.5.2 Powell minimisation

Powell's method [15] is a multidimensional minimisation algorithm. In our work we use it in the context of maximum likelihood, i.e. to minimise the negative log likelihood with respect to the vector of association parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.

Powell minimisation is based on a series of so-called 'line minimisations' that each use a simple one-dimensional minimisation scheme. Consider a function $f(.)$ to be minimised with respect to an n-dimensional parameter $\boldsymbol{x}$. The general idea is as follows: begin from some initial position vector $\boldsymbol{x}_0$ and choose some directional vector $\boldsymbol{u}$. Minimise $f(\boldsymbol{x})$ along the vector $\boldsymbol{u}$ using a one-dimensional minimisation algorithm (e.g. Brent's method [16]) and move the position vector to the position of this minimum. Then pick a new direction and minimise the function along this vector, and so on until the function is no longer decreasing.

The task is then to pick a set of directions along which to perform the line minimisations. A convenient choice is a set in which minimisation in one direction is not 'spoiled' by subsequent minimisation in another. Such non-interfering directions are called 'conjugate directions'. For a quadratic function with a set of n independent conjugate directions, one cycle of n line minimisations leads directly to the minimum of the function [16, 17]. For functions that are not quadratic, repeated cycles of the n line minimisations converge quadratically to the minimum [17].

Powell (1964) [15] proposed a method to produce n mutually conjugate directions, $\boldsymbol{u}_i$, $i = 1, \ldots, n$. We describe the basic procedure below. An illustration of the procedure for a 2-dimensional quadratic function is shown in Figure 5.3.

Powell minimisation (basic procedure):

1. Initialise $t = 0$. Initialise the set of directions as the basis vectors, $\boldsymbol{u}_i^{(0)} = \boldsymbol{e}_i$, $i = 1, \ldots, n$. Initialise the position vector $\boldsymbol{x}_0^{(0)}$.
2. For $i = 1, \ldots, n$: starting at $\boldsymbol{x}_{i-1}^{(t)}$, minimise $f(.)$ along the direction $\boldsymbol{u}_i^{(t)}$ and define $\boldsymbol{x}_i^{(t)}$ as the position of the minimum.
3. For $i = 1, \ldots, n-1$: update the direction vectors by setting $\boldsymbol{u}_i^{(t+1)} = \boldsymbol{u}_{i+1}^{(t)}$.
4. Update the final direction vector via $\boldsymbol{u}_n^{(t+1)} = \boldsymbol{x}_n^{(t)} - \boldsymbol{x}_0^{(t)}$.

5. Starting at $\boldsymbol{x}_n^{(t)}$, minimise $f(.)$ along the direction $\boldsymbol{u}_n^{(t+1)}$ and define $\boldsymbol{x}_0^{(t+1)}$ as the position of the minimum.

6. Set $t = t + 1$ and go to 2.

Powell showed that, for a quadratic function $f(.)$, k iterations of the above procedure produces a set of directions $\boldsymbol{u}_i$ whose final k members are mutually conjugate [15]. Therefore, n iterations of the basic procedure (i.e. $n(n+1)$ individual line minimisations) will exactly minimise a quadratic function [16, 17].

Non-quadratic functions require repeated cycles of the basic procedure. A problem with Powell's method is that the process of iteratively discarding $\boldsymbol{u}_1$ in favour of $\boldsymbol{x}_n - \boldsymbol{x}_0$ eventually results in a set of directions that are linearly dependent. When this happens, the algorithm only minimises $f(.)$ over a subspace of the full n-dimensions [17].

A number of modifications to Powell's method have been suggested to overcome this [16, 17]. A straightforward technique is to simply re-initialise the basis vectors $\boldsymbol{u}_i = \boldsymbol{e}_i$ after every n or $n + 1$ iterations of the basic procedure [17]. Another method involves discarding the direction of largest decrease [18]. In this approach, the direction $\boldsymbol{x}_n - \boldsymbol{x}_0$ is still assigned as the new direction but instead of discarding $\boldsymbol{u}_1$, one discards the direction in which $f(.)$ showed the greatest decrease during minimisation. This direction is likely to make the largest contribution to $\boldsymbol{x}_n - \boldsymbol{x}_0$ and therefore its removal should avoid a build up of linear dependence[2].

The modified procedure is repeated until the function no longer decreases (i.e. subsequent iterations produce a value of $f(\boldsymbol{x})$ which is the same as previous iterations to within some specified tolerance). At convergence one obtains the position of the minimum. In the context of Cox's PH model, this position corresponds to the maximum likelihood estimates of the regression parameters β_μ.

The maximum likelihood estimates of β_μ can be used to make statements about the relationship between different covariates and survival. Via Equation (5.18) one can then estimate the base hazard function and fully specify the hazard model via Equation (5.11). This leads to a fitted survival model [Equation (5.3)] from which

[2]There are some exceptions to this rule. For example, at certain iterations it may be better not to choose a new direction at all. See [17] for details.

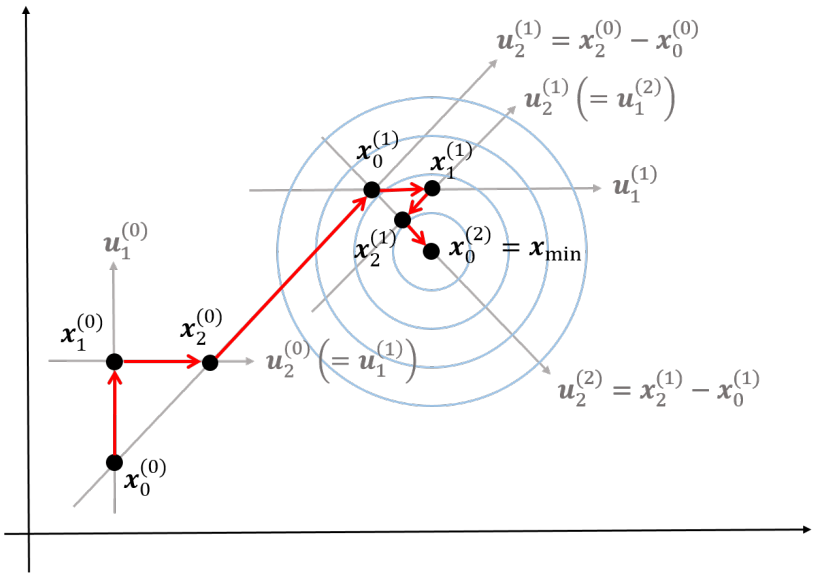


Figure 5.3: An illustration of the basic Powell minimisation procedure in 2 dimensions. The blue concentric circles indicate the quadratic function to be minimised. The thin grey arrows show the directions $\boldsymbol{u}_i$ at each iteration. Equalities in brackets indicate that the variable on the right is assigned the value of the variable on the left. The black dots show the position vector after each line minimisation (indicated by thick red arrows). The function is quadratic in 2-dimensions, therefore the algorithm converges to the exact minimum of the function after 2 iterations of the basic procedure (i.e. after $n(n+1) = 6$ line minimisations). The directions at the final iteration $(\boldsymbol{u}_1^{(2)}, \boldsymbol{u}_2^{(2)})$ are conjugate.

predictions can be made. For a new individual, measurements of their covariates z_μ^i are obtained and, using the fitted association parameters in the survival model, one can estimate the probability that they survive to some specified future time.

5.6 Extensions to standard survival analysis models

In this chapter we have introduced some of the main concepts of survival analysis with reference to the simplest type of time-to-event data. In reality, survival data may exhibit a number of characteristics that require more sophisticated modelling approaches. For example, individuals in the cohort may be at risk from more than one clinically relevant event. These events are considered 'competing' when the occurrence of one event means we are no longer able to observe another. This is related to the idea of informative censoring, i.e. the event of interest is censored by the occurrence of some competing event that itself carries information about the patient's health or condition. One may also be interested in modelling the risks from multiple events related to the condition of interest. Data of this type requires a 'competing risks' model [19].

Survival data could also include observations of multiple events for a single individual; these may be different (potentially correlated) events or repeated incidents of the same event. For example, a person may suffer from multiple strokes. Additional modelling considerations are needed for data with recurrent events [20].

Cox's model assumes that hazards are proportional, that is, the hazard rate in one covariate 'group' is a constant multiple of the hazard in another group. Researchers must check that this assumption is valid for their data set, for example by plotting the hazard curves for different groups [10, 12]. If the PH assumption does not hold, then a non-proportional hazards model is required [21, 22].

Another possible complication is 'dimensional mismatch' where one has access to large number of covariate measurements compared with the number of patients in the sample. For example, recent advancements in genome medicine mean that one often has access to genetic covariates which may include as many as 10^6 variables [7, 23]. To avoid 'overfitting' problems, regression models typically require the number of samples to be much larger than the number of fitted parameters [24, 25]. This means that standard methods are restricted in the number of covariates they can model and we lose out on a wealth of available information. Sophisticated analysis methods have been developed to overcome these issues [7, 23], though such methods are themselves not a panacea (see e.g. [26]).

In Chapter 6 we focus on data with time-varying covariate measurements. It is common that patients with a particular condition will attend follow-up medical appointments over the course of the study. Covariates that may fluctuate or progress over time can then be measured at these appointments. Temporal changes in these variables are expected to carry information about patient survival, for example, an increase in a certain covariate may indicate a worsening of the condition. It then makes sense to include the full history of these measurements in the survival model and to update predictions as new measurements become available. This process is known as 'dynamic prediction' and will be discussed in detail in the proceeding chapter.

Bibliography

[1] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part I: basic concepts and first analyses", Brit. J. Cancer **89**, 232–238 (2003).

[2] D. G. Kleinbaum and M. Klein, *Survival analysis. A self-learning text* (Springer-Verlang, New York, NY, USA, 2005).

[3] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data* (Wiley, NJ, USA, 2012).

[4] I. A. Ushakov, ed., *Handbook of reliability engineering* (John Wiley & Sons, New York, NY, USA, 1994).

[5] J. J. Heckman and B. Singer, "Econometric duration analysis", J. Econometrics **24**, 63–132 (1984).

[6] B. George, S. Seals, and I. Aban, "Survival analysis and regression models", J. Nucl. Cardiol. **21**, 686–694 (2014).

[7] S. Lee and H. Lim, "Review of statistical methods for survival analysis using genomic data", Genom. Inform. **17**, e41 (2019).

[8] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations", J. Am. Stat. Assoc. **53**, 457–481 (1958).

[9] N. E. Breslow, "Analysis of survival data under the proportional hazards model", Int. Stat. Rev. **43**, 45–58 (1975).

[10] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival analysis part II: multivariate data analysis – an introduction to concepts and methods", Brit. J. Cancer **89**, 431–436 (2003).

[11] D. R. Cox, "Regression models and life-tables", J. Roy. Stat. Soc. B Met. **34**, 187–202 (1972).

[12] M. Mills, "The Cox proportional-hazards regression model", in Introducing survival and event history analysis (2012), pp. 86–113.

[13] N. E. Breslow, "Discussion on Professor Cox's paper", J. Roy. Stat. Soc. B Met. **34**, 216–217 (1972).

[14] D. Y. Lin, "On the Breslow estimator", Lifetime Data Anal. **13**, 471–480 (2007).

[15] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives", Comput. J. **7**, 155–162 (1964).

[16] R. P. Brent, *Algorithms for minimization without derivatives* (Prentice-Hall, Englewood Cliffs, NJ, USA, 1973).

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C. The art of scientific computing*, 2nd ed. (Cambridge University Press, Cambridge, UK, 1992).

[18] F. S. Acton, *Numerical methods that work* (Mathematical Association of America, Washington, DC, USA, 1990).

[19] R. Varadhan, C. O. Weiss, J. B. Segal, A. W. Wu, D. Scharfstein, and C. Boyd, "Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications", Med. Care **48**, S96–S105 (2010).

[20] J. Cai and D. E. Schaubel, "Analysis of recurrent event data", in *Advances in survival analysis*, Vol. 23, edited by N. Balakrishnan and C. R. Rao, Handbook of Statistics (Elsevier, 2003), pp. 603–623.

[21] M. Schemper, "Cox analysis of survival data with non-proportional hazard functions", J. Roy. Stat. Soc. D Sta. **41**, 455–465 (1992).

[22] R. Giorgi, M. Abrahamowicz, C. Quantin, P. Bolard, J. Esteve, J. Gouvernet, and J. Faivre, "A relative survival regression model using B-spline functions to model non-proportional hazards", Stat. Med. **22**, 2767–2784 (2003).

[23] A. C. C. Coolen, J. E. Barrett, P. Paga, and C. J. Perez-Vicente, "Replica analysis of overfitting in regression models for time-to-event data", J. Phys. A: Math. Theor. **50** (2017).

[24] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction", Stat. Med. **3**, 143–152 (1984).

[25] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors", Stat. Med. **15**, 361–387 (1996).

[26] R. D. Riley, K. I. Snell, G. P. Martin, R. Whittle, L. Archer, M. Sperrin, and G. S. Collins, "Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small", J. Clin. Epidemiol. **132**, 88–96 (2021).

Chapter 6

Retarded kernels for longitudinal survival analysis and dynamic prediction

Preface

The contents of this chapter constitute a manuscript submitted to Statistical Methods in Medical Research. The preprint edition is available on ArXiv[1]. Originally, the sections that now appear as appendices were submitted as a separate supplementary file. The manuscript was authored by Annabel L Davies[2], Anthony C C Coolen[3,4] and Tobias Galla[2,5].

ALD designed the study, contributed to discussions guiding the work, carried out the mathematical calculations, performed the data analysis, wrote the first draft of the manuscript, produced all of the figures and edited the manuscript. ACCC and TG designed the study, contributed to discussions guiding the work and edited the manuscript.

[1]A. L. Davies, A. C. C. Coolen and T. Galla, "Retarded kernels for longitudinal survival analysis and dynamic prediction", *arXiv preprint.* arXiv:2110.11196 (2021).

[2]Theoretical Physics, School of Physics and Astronomy, The University of Manchester, Manchester, M13 9PL, United Kingdom.

[3]Department of Biophysics, Radboud University, The Netherlands

[4]Saddle Point Science Ltd, UK

[5]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

Abstract

Predicting patient survival probabilities based on observed covariates is an important assessment in clinical practice. These patient-specific covariates are often measured over multiple follow-up appointments. It is then of interest to predict survival based on the history of these longitudinal measurements, and to update predictions as more observations become available. The standard approaches to these so-called 'dynamic prediction' assessments are joint models and landmark analysis. Joint models involve high-dimensional parameterisations, and their computational complexity often prohibits including multiple longitudinal covariates. Landmark analysis is simpler, but discards a proportion of the available data at each 'landmark time'. In this work we propose a 'retarded kernel' approach to dynamic prediction that sits somewhere in between the two standard methods in terms of complexity. By conditioning hazard rates directly on the covariate measurements over the observation time frame, we define a model that takes into account the full history of covariate measurements but is more practical and parsimonious than joint modelling. Time-dependent association kernels describe the impact of covariate changes at earlier times on the patient's hazard rate at later times. Under the constraints that our model (i) reduces to the standard Cox model for time-independent covariates, and (ii) contains the instantaneous Cox model as a special case, we derive two natural kernel parameterisations. Upon application to three clinical data sets, we find that the predictive accuracy of the retarded kernel approach is comparable to that of the two existing standard methods.

6.1 Introduction

Survival analysis is a well-established field of medical statistics that involves modelling the probability of survival until some specified irreversible event such as death or the onset of disease. Of particular clinical interest is the prediction of patient-specific survival based on a set of observed biomarkers or 'covariates' [1]. Such predictions aid clinicians in making treatment and testing decisions, and provide personalised

information for patients about their health [2].

Cox's proportional hazards (PH) model [3] remains the most widely used model in survival analysis [4, 5]. In this context, survival is assumed to depend on a set of covariates, z_μ, $\mu = 1, \ldots, p$, measured at some baseline time. The hazard, $h(t)$, is the probability per unit time of the event happening at time t given that no event has occurred up to that time. In Cox's PH model this hazard is defined as

$$h(t) = h_0(t) \mathrm{e}^{\sum_{\mu=1}^{p} \beta_\mu z_\mu}, \tag{6.1}$$

where the β_μ (with $\mu = 1, \ldots, p$) are the so-called association parameters. The base hazard rate, $h_0(t)$, is the value of the hazard for covariate values $z_\mu = 0 \; \forall \mu$. The name 'proportional hazards' refers to the fact that, due to the exponential form of the hazard function, the effect of each covariate on the hazard is multiplicative. In this work we will call the model in Equation (6.1) the 'standard Cox model'.

Survival prediction in the standard Cox model is based on the survival function,

$$S(t) = \mathrm{e}^{-\int_0^t \mathrm{d}t' h(t')}, \tag{6.2}$$

that describes the probability that an individual with hazard function $h(\cdot)$ experiences the event after time t.

In reality, covariates are often measured repeatedly over time. This means that multiple observations of time-dependent covariates $\{z_\mu(t)\}$ are made for any particular patient. A simple extension to the standard Cox model involves modelling the hazard rate as dependent on the instantaneous value of the covariates [6–8], that is

$$h(t) = h_0(t) \mathrm{e}^{\sum_{\mu=1}^{p} \beta_\mu z_\mu(t)}. \tag{6.3}$$

We refer to this as the 'instantaneous Cox model'.

However, in practice, one does not have access to the full covariate trajectories $z_\mu(t)$. Instead observations are made at discrete follow-up times until some subject-specific final observation time. Since we do not have access to covariate measurements after this time, we cannot make predictions about future survival probabilities based on Equation (6.3). Of particular difficulty is the inclusion of so-called 'endogenous' covariates [9].

Due to these difficulties, survival predictions are commonly evaluated by treating the baseline covariate measurements as fixed values in a standard Cox model [2]. By

not including the follow-up observations, this standard practice discards a potentially considerable proportion of the available patient data.

Recently, there has been much interest in so-called 'dynamic prediction' [1, 10, 11]. These methods aim to make survival predictions based on the longitudinal history of biomarker data, and update these predictions as more data becomes available. Such analysis is clinically valuable as it allows patients and clinicians to review disease progression over time and update the prognosis at each follow-up visit [12]. Currently, there are two main approaches to dynamic prediction; joint modelling and landmarking.

Landmarking was an early approach to the problem [13], whereby a standard Cox model is fitted to patients in the original data set who are still at risk at the time point of interest, using their most recent covariate measurements.

More recently, joint modelling has become an established method [9, 14–16]. Here one models the time-dependent covariate trajectory using a parameterised longitudinal model, and this complete trajectory is then inserted into an instantaneous Cox-type survival model. A joint likelihood of the longitudinal and survival sub-models is constructed, and the model parameters are estimated via maximum likelihood or Bayesian inference.

Both methods have limitations. In particular, joint models are demanding both conceptually and computationally. Correctly modelling the longitudinal trajectories can be difficult when patient measurements exhibit varied non-linear behaviour [12] and misspecification of this trajectory has been found to lead to bias [17]. Furthermore, the number of model parameters increases rapidly with the inclusion of multiple longitudinal markers. This means that many software packages cannot handle more than one longitudinal covariate [18–20], and those that can quickly become computationally intensive [21, 22]. For these reasons, the landmarking model is often seen as the only practical option [12]. However, the relative simplicity of the landmarking approach comes with its own drawbacks. By using only the 'at risk' data set to make predictions at a certain time (discarding patients who had an event before the landmark time), landmarking makes use of only a subset of the available data. In standard landmarking approaches, the history of the covariate values are not taken into account directly, and a new model must be fitted every time one wishes to update the predictions.

In this work we present a new approach to dynamic prediction that conceptually and

in terms of computational complexity lies somewhere in between the joint modelling and landmarking methods. Rather than modelling the covariate trajectory at future times, as in the joint modelling approach, we model the probability of survival conditioned directly on the observed covariates measured from the baseline time up to a subject-specific final observation time. Unlike the landmark approach, a single model is fitted to all of the available data, using the full history of the covariate values. We do, however, maintain well-established and desirable features of the Cox model, so that our model contains the instantaneous Cox model as a special case, and reduces automatically to the standard Cox model for covariates that are observed to be fixed over time. Within these constraints, we define time-dependent parametric association kernels, $\beta_\mu(t, t', s)$, that describe the impact of changes of covariate μ at time t' on patient risk at some later time t. The kernel can also depend on the final observation time s for the patient. Building on ideas from weighted cumulative exposure models [23, 24], these kernels allow us to assign smaller effects to covariates that were measured further in the past. We refer to our method as the 'retarded kernel' approach.

The remainder of this article is set out as follows. In Section 6.2 we introduce the motivating data sets. In Section 6.3 we then provide details of the dynamic prediction models. We begin by describing the longitudinal and time-to-event data, and briefly outline the standard methods: joint modelling (Section 6.3.2) and landmarking (Section 6.3.3). In Section 6.3.4, we introduce the retarded kernel approach. We start by defining the hazard rate conditioned on the observed data, and then develop two natural parameterisations for the association kernels that meet our requirements. We outline the maximum likelihood method for parameter estimation for these models, and show how the retarded kernel approach can be used to make dynamic predictions. Via application to the real data sets, in Section 6.4 we compare the performance of the retarded kernel approach to the standard methods using an established measure of predictive accuracy. Finally, we discuss and summarise our results in Section 6.5.

6.2 Motivating data sets

In our work we will assess the predictive capabilities of the different models for dynamic prediction using three clinical data sets, that contain both longitudinal covariate

measurements and time-to-event data. All three data sets are publicly available in the `JMbayes` package [25], and were used in Rizopoulos (2012) [9] to illustrate the joint modelling method.

6.2.1 Primary biliary cirrhosis

The first motivating data set is from a study conducted by the Mayo Clinic from 1974 to 1984 on patients with primary biliary cirrhosis (PBC), a progressive chronic liver disease [26]. We will refer to this as the PBC data. The study involved $N = 312$ patients who were randomly assigned either a placebo (154 patients) or the D-penicillamine treatment (158 patients). Time-to-event data is available for the outcome of interest (death) or the censoring event (either the time at which the patient receives a liver transplant or the final follow-up time at which they were still alive). By the end of follow-up, 140 patients had died, 29 had received a transplant and 143 were still alive.

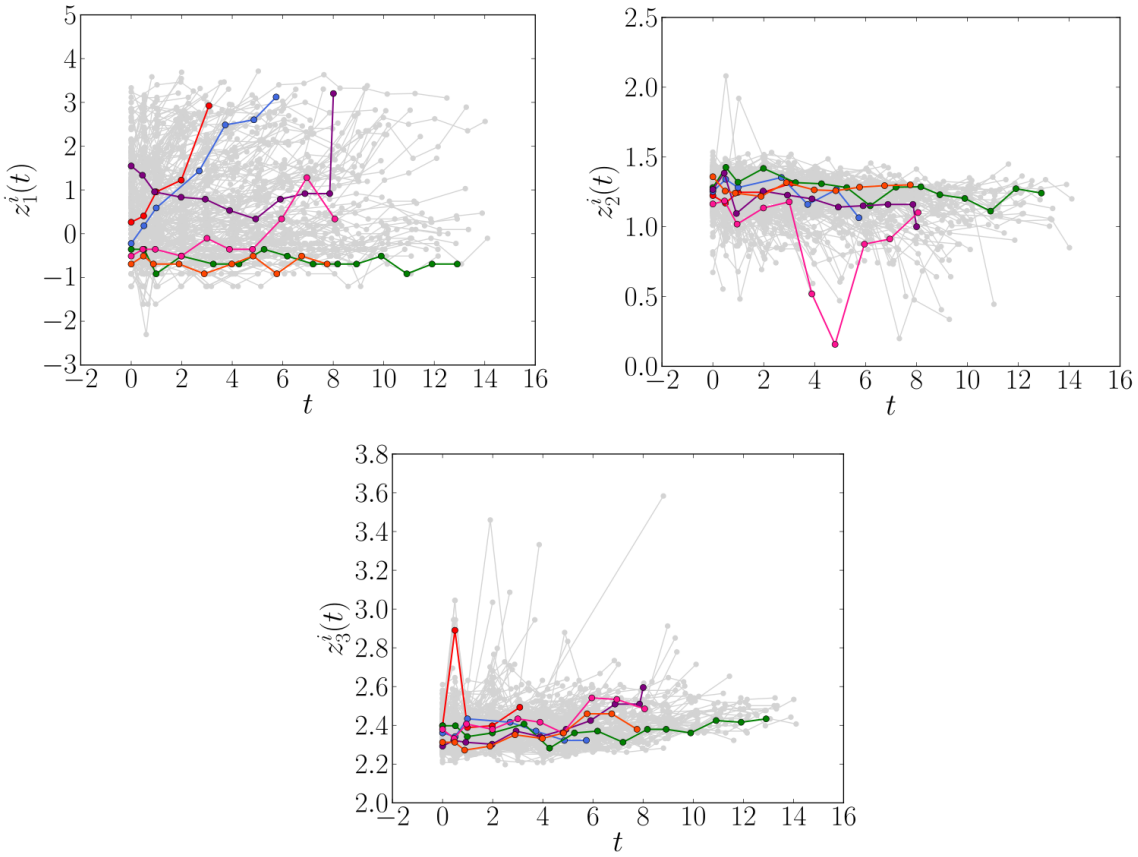


Figure 6.1: The longitudinal profiles of the time-dependent covariates log serum bilirubin ($z_1^i(t)$), log serum albumin ($z_2^i(t)$) and log prothrombin time ($z_3^i(t)$) for the $N = 312$ patients ($i = 1, \ldots, N$) in the PBC data set described in Section 6.2.1. For clarity, the trajectories of 6 individuals are highlighted. Time, t, on the x-axis is measured in years.

As well as baseline covariate measurements such as age at baseline and gender, multiple longitudinal biomarker measurements were collected for each patient over an average number of 6.2 visits from study entry to some subject-specific final observation time (prior to their event time). While the original aim of the study was to investigate the effect of the drug D-penicillamine, no effect was found and the data has since been used to study the progression of the disease based on longitudinal biomarkers [27]. With this in mind, we include age at baseline as our only fixed covariate, and focus on the longitudinal covariates log serum bilirubin, log serum albumin and log prothrombin time, which have previously been found to be indicators of patient survival [27]. Serum bilirubin and serum albumin indicate concentrations of these substances in the blood, measured in mg/dl and g/dl respectively. Prothrombin time measures the time (in seconds) it takes for blood to clot in a sample. Time series of these three longitudinal biomarkers are plotted in Figure 6.1.

6.2.2 AIDS

The second data set involves $N = 467$ HIV-infected patients who had failed to respond, or were intolerant to, zidovudine (previously called 'azidothymidine') therapy (AZT) [28]. The aim of the study was to compare two antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC). Patients were randomly assigned one of these drugs at baseline. Patients' CD4 cell counts were recorded at baseline and follow-up measurements were planned at 2, 6, 12 and 18 months. CD4 cells are white blood cells that fight infections. A decrease in the number of CD4 cells over time indicates a worsening of the immune system and higher susceptibility to infection. Therefore, the number of CD4 cells in a blood sample is an important marker of immune strength and hence a covariate of interest in HIV-infected patients. In line with previous analysis of this data [9, 29], we actually use the square root of the CD4 count as our longitudinal covariate. For brevity we will refer to this simply as the CD4 count. By the end of the study 118 patients had died, and the time to event (death) or censoring was recorded for all patients. Final observation times ($s_i \in [0, 2, 6, 12, 18]$ months) were always less than their corresponding event times, such that there is a time gap between when a subject was last observed and when they experienced an event. Following Guo and Carlin (2004) [29], we included, in addition to the longitudinal CD4 counts and the patients'

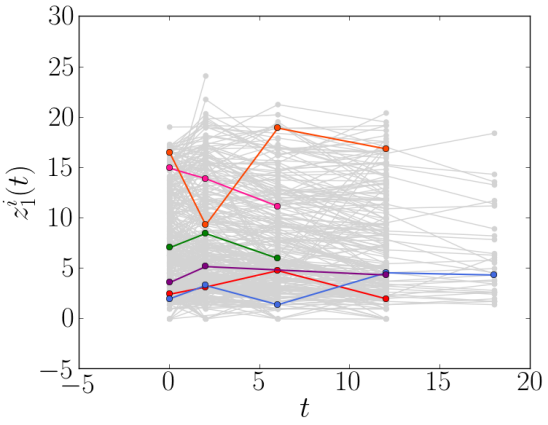


Figure 6.2: The longitudinal profiles of CD4 count ($z_1^i(t)$) in the $N = 467$ patients ($i = 1, \ldots, N$) in the AIDS data set in Section 6.2.2. For clarity, the trajectories of 6 individuals are highlighted. Time, t, is measured in months.

drug group, also three other binary fixed covariates in our analysis: gender, PrevOI (previous opportunistic infection – AIDS diagnosis – at study entry), and Stratum (whether the patient experienced AZT failure or AZT intolerance). We will refer to this data as the AIDS data set. The longitudinal profiles of the CD4 count for all patients are plotted in Figure 6.2.

6.2.3 Liver cirrhosis

The third data set is from a trial conducted between 1962 and 1974, involving $N = 488$ patients with liver cirrhosis, a general term including all forms of chronic diffuse liver disease [30]. We call this the Liver data set. At baseline, 251 patients were randomly assigned a placebo and 237 were assigned treatment with the drug prednisone. Follow-up appointments were scheduled at 3, 6 and 12 months and then yearly thereafter, though actual follow up times varied considerably. At these follow up appointments, multiple longitudinal biomarkers were measured. However, only the prothrombin index measurements are available from the `JMbayes` package. This is a measure of liver function based on a blood test of coagulation factors produced by the liver. For reproducibility, and following previous analyses of the Liver data set [9, 31], we include the prothrombin index as our only time-dependent biomarker. The drug group is included as a fixed baseline covariate. By the end of the study, 150 prednisone-treated, and 142 placebo-treated patients had died. Their time-to-event data was recorded. Of the 488 subjects, 120 were observed until their event time while all others were observed

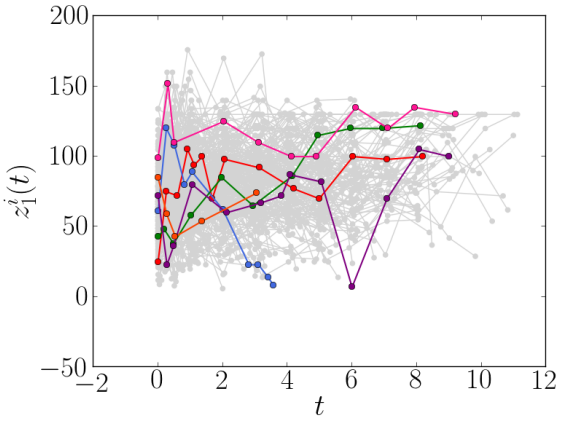


Figure 6.3: The longitudinal profiles of the prothrombin index ($z_1^i(t)$) as measured in the $N = 488$ patients ($i = 1, \ldots, N$) in the Liver data set described in Section 6.2.3. For clarity, the trajectories of 6 individuals are highlighted. Time, t, is measured in years.

until some subject-specific final observation time before their event time. Figure 6.3 shows the longitudinal prothrombin measurements for all patients.

6.3 Dynamic prediction models

6.3.1 Setup and notation

In this work, we consider longitudinal survival data of the form $\mathcal{D} = \{T_i, \delta_i, \mathcal{Z}^i; i = 1, \ldots, N\}$ where $T_i = \min(T_i^*, C_i)$ is the observed event time of individual i with T_i^* denoting the true event time and C_i denoting the censoring time. The event indicator $\delta_i = I(T_i^* \leq C_i)$ is equal to 1 when the true event time is observed and 0 when it is censored. Throughout this article we use the indicator function $I(A)$, defined as $I(A) = 1$ if A holds, and $I(A) = 0$ otherwise. $\mathcal{Z}^i = \{z_\mu^i(t_{i\ell});\ \mu = 1, \ldots, p,\ \ell = 1, \ldots, n_i,\ t_{i\ell} \in [0, s_i]\}$ denotes the set of time-dependent covariate observations of individual i. Individual i has n_i measurements of p longitudinal covariates from time $t = 0$ up to some subject-specific final observation time $s_i \leq T_i$. These measurements are taken at discrete (subject-specific) observation times, $t_{i\ell}$, $\ell = 1, \ldots, n_i$, where $t_{i1} = 0$ and $t_{in_i} = s_i$. We write $\mathcal{Z}^i_{[0,s_i]} = \{z_\mu^i(t);\ \mu = 1 \ldots p,\ t \in [0, s_i]\}$ for the 'true' (but non-accessible) continuous trajectories of the p covariates over the interval $t \in [0, s_i]$ for individual i. We develop our theory based on the assumption that we have access to these trajectories $\mathcal{Z}^i_{[0,s_i]}$. As we will see later, we estimate $\mathcal{Z}^i_{[0,s_i]}$ from the discrete observations $\mathcal{Z}^i$.

We are interested in predicting survival probabilities for some new subject with longitudinal measurements $\mathcal{Z} = \{z_\mu(t_\ell);\ \mu = 1, \ldots, p,\ \ell = 1, \ldots, n,\ t_\ell \in [0, s]\}$. The quantity we wish to estimate is the probability that this subject survives until some future time $u > s$, conditional on their survival to s, and on their covariate observations up to s. That is,

$$\pi(u|\mathcal{Z}_{[0,s]}, s) = \Pr(T^* \geq u|\ T^* > s, \mathcal{Z}_{[0,s]}, \mathcal{D}). \tag{6.4}$$

The quantity $\pi(u|\mathcal{Z}_{[0,s]}, s)$ is referred to as a 'dynamic predictor' due to the fact that it can be updated as more measurements become available at later times [2, 11].

6.3.2 Joint Models

In joint modelling one specifies two model components: a longitudinal model for the trajectory of the time-dependent covariates, and a survival model which relates to the covariate trajectory via shared parameters. In the `JMbayes` package, joint models are fitted using Bayesian inference by specifying a joint likelihood distribution for the two model components and a set of prior distributions on the model parameters. Details of this package and the joint modelling framework we follow are described in Rizopoulos (2016) [25] and Rizopoulos (2012) [9]. In this section we briefly outline the model.

6.3.2.1 Longitudinal modelling component

Mixed-effects models are typically specified for the longitudinal covariate trajectories, where it is assumed that the observed value $z_\mu(t)$ of the covariate at time t deviates from the true (unobserved) value $m_\mu(t)$ by an amount $\varepsilon_\mu(t)$. The error terms $\varepsilon_\mu(t)$ of all subjects are assumed to be statistically independent, and normally distributed with variance σ_μ^2:

$$\begin{aligned} &z_\mu^i(t) = m_\mu^i(t) + \varepsilon_\mu^i(t), \qquad m_\mu^i(t) = \boldsymbol{x}_\mu^{i\top}(t)\boldsymbol{\eta}_\mu + \boldsymbol{y}_\mu^{i\top}(t)\boldsymbol{b}_\mu^i \\ &\boldsymbol{b}^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \varepsilon_\mu^i(t) \sim \mathcal{N}(0, \sigma_\mu^2). \end{aligned} \tag{6.5}$$

Between-subject variability is modelled via estimation of subject-specific random effects $\boldsymbol{b}_\mu^i$, whereas effects that are shared between all subjects are modelled by the fixed effects $\boldsymbol{\eta}_\mu$. The vectors $\boldsymbol{x}_\mu^{i\top}(t)$ and $\boldsymbol{y}_\mu^{i\top}(t)$ denote the design vectors for these fixed and random effects respectively. For multivariate models, one can allow for association between the

different longitudinal markers via their corresponding random effects. In particular, we assume that the complete vector of random effects $\boldsymbol{b}^i = (\boldsymbol{b}_1^{i\top}, \ldots, \boldsymbol{b}_p^{i\top})^\top$ follows a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$ that describes the correlations between and variances of the random effects. For details we refer to References [2, 9, 10, 21, 32].

6.3.2.2 Survival modelling component

The hazard at time t is assumed to depend on the value of the longitudinal covariate at time t without measurement error, that is

$$h^{\mathrm{JM}}(t|\mathcal{M}_{[0,t]}) = h_0(t) \exp\left\{\sum_{\mu=1}^{p} \alpha_\mu m_\mu(t)\right\}, \tag{6.6}$$

where $\mathcal{M}_{[0,t]} = \{m_\mu(t');\ \mu = 1, \ldots, p,\ t' \in [0,t]\}$ denotes the history of the 'true' (unobserved) longitudinal covariates up to time t. Note that in Equation (6.6) the hazard rate depends only on the instantaneous values of the covariates, but this can be generalised as briefly highlighted below. Unlike in the Cox model, the baseline hazard, $h_0(t)$ cannot be expressed analytically in terms of the other model parameters during the maximum likelihood procedure, but must instead be specified. Often this is done using a flexible parametric model, for example using penalised spline functions [25]. Dependence of the hazard function on time-independent covariates $\{\zeta_\nu;\ \nu = 1, \ldots, q\}$ can be included through an additional term $\sum_{\nu=1}^{q} \gamma_\nu \zeta_\nu$ in the exponent in Equation (6.6), where γ_ν is the association parameter for fixed covariate ν. Alternative extensions allow the hazard to depend on the slope of the covariate trajectory, or on its cumulative effect, by replacing the term $\alpha_\mu m_\mu(t)$ with $\alpha_\mu^{(1)} m_\mu(t) + \alpha_\mu^{(2)} \frac{\mathrm{d}}{\mathrm{d}t} m_\mu(t)$ or with $\alpha_\mu \int_0^t \mathrm{d}t'\ m_\mu(t')$, respectively [2, 9].

It has also been proposed to introduce a weight function to capture cumulative effects, writing $\alpha_\mu \int_0^t \mathrm{d}t'\ w_\mu(t-t') m_\mu(t')$, with kernels $w_\mu(t-t')$ defined such that earlier covariate values have a smaller effect on the hazard than recent values [9, 33]. This idea is connected to the concept of 'weighted cumulative exposure' (WCE) [23, 24]. WCE models were developed in etiological research to describe the complex cumulative effect of time-dependent 'exposure', e.g. to a drug, on health outcomes [34]. In a survival context, these models rely on continuous knowledge of the exposure all the way up to the event time, and have hence been used almost exclusively for measuring the effects

of external exposures such as treatments or environmental factors [35]. Joint modelling allows the principles of WCE to be used for any longitudinal covariate. Through the prediction of future covariate trajectories, it is also possible for these ideas to be integrated into dynamic predictions [33]. As we will explain later, we build on the principles of WCE to develop our retarded kernel approach.

6.3.2.3 Dynamic prediction

Finally, in the joint modelling framework we use, the quantity $\pi(u|\mathcal{Z}_{[0,s]}, s)$ is estimated using a Bayesian approach, with posterior parameter distribution $p(\boldsymbol{\theta}_{\text{JM}}|\mathcal{D})$ and where $\boldsymbol{\theta}_{\text{JM}}$ is the vector of all the model parameters in the joint model. This leads to the estimator

$$\hat{\pi}^{\text{JM}}(u|\mathcal{Z}_{[0,s]}, s) = \int \Pr(T^* \geq u|\ T^* > s, \mathcal{Z}_{[0,s]}, \boldsymbol{\theta}_{\text{JM}}) p(\boldsymbol{\theta}_{\text{JM}}|\mathcal{D})\ \mathrm{d}\boldsymbol{\theta}_{\text{JM}}. \tag{6.7}$$

The parameter average in Equation (6.7) can generally not be evaluated analytically, and is computed via Monte Carlo methods. Again, we refer to References [2, 9, 10] for details.

6.3.2.4 Limitations

Joint models have undergone much development over recent years, with various extensions making the approach flexible in a range of different scenarios. However, the joint modelling approach requires the ability to correctly specify both the longitudinal and survival model. This can involve modelling assumptions which are not always easy to verify. Indeed, simulations have demonstrated that the joint modelling approach is biased under misspecification of the longitudinal model [17]. In addition, as more longitudinal outcomes are included, the dimensionality of the random effects increases, and fitting the joint model becomes computationally intensive. Depending on the longitudinal model specified, it can be difficult to include more than three or four longitudinal covariates [12, 21]. This is amplified when cumulative or weighted cumulative effects are used in the survival model (as numerical integration of the longitudinal model is required). As a result, there are cases where joint models are not a viable option and, instead, one must rely on approaches such as landmarking [12].

6.3.3 Landmarking

6.3.3.1 Description of the landmarking procedure

The landmarking approach to dynamic prediction is based on the standard Cox model [1, 11, 13]. Upon denoting with $\mathcal{R}(\upsilon) = \{i : \ T_i > \upsilon\}$ the set of individuals in the original data set who are still at risk at time υ, the landmarking model assumes that for a subject in the risk set $\mathcal{R}(\upsilon)$ the distribution of survival times, conditioned on the covariate measurements $\{z^i_\mu(\upsilon)\}$ at that time, follows a standard Cox model [12]. In general one does not have covariate measurements for all individuals at time υ. Instead, one uses for each individual the last observation $\{\tilde{z}_\mu(\upsilon),\ \mu = 1, \ldots, p\}$ of the covariates before time υ, and treats these as fixed covariates in a standard Cox model with υ as the baseline time. That is, for the so-called 'landmark time' υ one defines the hazard rate at times $t > \upsilon$ as

$$h^{\mathrm{LM}}(t|\mathcal{Z}, \upsilon) = h_0(t|\upsilon) \exp\left\{\sum_{\mu=1}^{p} \alpha_\mu(\upsilon)\tilde{z}_\mu(\upsilon)\right\}. \tag{6.8}$$

The baseline hazard function $h_0(t|\upsilon)$ is unspecified, and is estimated as in standard Cox models, via partial likelihood arguments, or via functional maximisation of the data log-likelihood. Subsequently the association parameters are estimated. This procedure is carried out for each choice of the landmark time υ, and leads to the Breslow estimator [36] $\hat{h}_0(t|\upsilon)$ and the association parameters $\hat{\alpha}_\mu(\upsilon)$. The main difference compared to standard Cox models is the dependence of association parameters and the base hazard rate on the landmark time υ.

To estimate the quantity $\pi(u|\mathcal{Z}_{[0,s]}, s)$, the landmark time υ in Equation (6.8) is set equal to s. Once this model is fitted, survival prediction to time $u > s$ is performed using the standard Cox survival probability,

$$\hat{\pi}^{\mathrm{LM}}(u|\mathcal{Z}_{[0,s]}, s) = \exp\left\{ - \mathrm{e}^{\sum_{\mu=1}^{p} \hat{\alpha}_\mu(s)\tilde{z}_\mu(s)} \int_s^u \hat{h}_0(t'|s)\mathrm{d}t' \right\}. \tag{6.9}$$

6.3.3.2 Limitations

Landmarking is computationally and conceptually much simpler than the joint modelling approach. For data sets with multiple longitudinal covariates, disparate non-linear covariate trajectories or categorical time-dependent covariates, landmarking is often the preferred approach [12]. However, it also has limitations. For example, the model

focuses only on the most recent value observed before time υ, and does not account for the earlier history of covariates. Furthermore, data from individuals who experience the event before time υ is not used for the parameter estimation at landmark time υ. Therefore, the landmark approach uses only a subset of the available data. In addition, a new Cox model has to be specified and fitted for each landmark time. Therefore, in order to update predictions after each time where subject j is observed, one must refit the model using a new risk set. The longer subject j is observed, the fewer individuals remain in the risk set and less data is available to do this.

6.3.4 Retarded kernel approach

We now introduce our retarded kernel approach to dynamic prediction. It aims to overcome some of the limitations of the standard joint modelling and landmarking methods. Unlike landmarking, the retarded kernel approach aims to incorporate the entire data set, including the full history of covariate values while, at the same time remaining conceptually and computationally simpler than joint models.

6.3.4.1 General setup

The starting point for the retarded kernel approach is an expression for the hazard rate that resembles that of weighted cumulative exposure models [23, 24],

$$h^{\mathrm{RK}}(t|\mathcal{Z}_{[0,s]}) = h_0(t)\exp\left\{\int_0^{\min(s,t)}\sum_{\mu=1}^{p}\beta_\mu(t,t',s)z_\mu(t')\mathrm{d}t'\right\}. \tag{6.10}$$

In this expression the $\{z_\mu(t')\}$ are time-dependent covariates, which we assume to be known from time 0 up to time s. To keep the notation compact we have left out time-independent covariates, as these can always be included trivially. This model differs from the joint model approach to WCE in how we deal with covariates that are only observed up to some final observation time s before the event time. When $t \leq s$ (i.e. when t is a point in time prior to the last observation of covariates) the hazard rate in Equation (6.10) only depends on covariates up to time t. For times $t \geq s$ covariates up to time s enter into the hazard rate.

The kernel $\beta_\mu(t,t',s)$ describes (potentially) retarded effects of covariates. More precisely, $\beta_\mu(t,t',s)$ quantifies the effect of the value of covariate μ at time t' on the hazard rate at a later time t, for a patient whose covariates are known up to time s.

The form of Equation (6.10) ensures causality, since only covariate values at times $t' \leq t$ contribute to the hazard at time t. We set $\beta(t, t', s) = 0$ for $t' > t$. In principle, the precise form of $\beta_\mu(t, t', s)$ could be chosen from a wide range of functions. We reduce this freedom via the following requirements which must hold for all μ:

(i) *Exponential decay of covariate impact.* We assume that the impact of each covariate μ at time t' on the hazard rate at a later time $t > t'$ decays exponentially with the time difference $t - t'$. How fast the effect of the covariate decays is governed by a covariate-specific impact time scale $\tau_\mu \geq 0$.

(ii) *Equivalence with standard Cox model for stationary covariates.* Our second requirement is that expression (6.10) reduces to the standard Cox model in Equation (6.1) in the case of a constant covariate, i.e. when $z_\mu(t) \equiv z_\mu$ for all t. This is achieved when there is a constant a_μ, which is independent of t and s, such that

$$\int_0^{\min(s,t)} \beta_\mu(t, t', s) \mathrm{d}t' = a_\mu. \tag{6.11}$$

(iii) *Equivalence with instantaneous Cox model for short impact time scales.* Finally, for $0 < t \leq s$ we require that expression (6.10) reduces to the instantaneous Cox model in Equation (6.3) in the limit $\tau_\mu \downarrow 0$, i.e. when the covariate impact on risk decays immediately. This is achieved, without violating (ii), if we have

$$\lim_{\tau_\mu \downarrow 0} \beta_\mu(t, t', s) = a_\mu \delta(t - t'). \tag{6.12}$$

From (i) it follows that our kernel $\beta_\mu(t, t', s)$ must have the following form:

$$\beta_\mu(t, t', s) = A_\mu(t, s) \tau_\mu^{-1} \mathrm{e}^{-(t-t')/\tau_\mu} + B_\mu(t, s), \tag{6.13}$$

where the quantities $A_\mu(t, s)$ and $B_\mu(t, s)$ can depend on τ_μ in general. Requirements (ii) and (iii) then translate into, respectively,

$$s, t \geq 0: \qquad A_\mu(t, s) \mathrm{e}^{-t/\tau_\mu} \left(\mathrm{e}^{\min(s,t)/\tau_\mu} - 1 \right) + \min(s, t) B_\mu(t, s) = a_\mu \tag{6.14}$$

$$0 < t \leq s: \qquad \lim_{\tau_\mu \downarrow 0} A_\mu(t, s) = a_\mu, \quad \lim_{\tau_\mu \downarrow 0} B_\mu(t, s) = 0. \tag{6.15}$$

6.3.4.2 Two models within this family

Finally, from the remaining family of models, i.e. those that satisfy Equations (6.14) and (6.15), we choose the two simplest members. These are defined by demanding that

either $B_\mu(t,s) = 0$ for *any* τ_μ (model A), or that $A_\mu(t,s) = a_\mu$ for *any* τ_μ (model B). Working out the details for these choices via Equations (6.14, 6.15) then leads to the following formulae:

$$\text{Model A:} \quad \beta_\mu^A(t,t',s) = \frac{a_\mu}{\tau_\mu} \frac{\mathrm{e}^{t'/\tau_\mu}}{\mathrm{e}^{\min(s,t)/\tau_\mu} - 1}, \tag{6.16}$$

$$\text{Model B:} \quad \beta_\mu^B(t,t',s) = \frac{a_\mu}{\tau_\mu} \mathrm{e}^{-(t-t')/\tau_\mu} + \frac{a_\mu}{\min(s,t)} \left\{ 1 - \mathrm{e}^{-t/\tau_\mu} \left(\mathrm{e}^{\min(s,t)/\tau_\mu} - 1 \right) \right\}. \tag{6.17}$$

Both models are built around the time-translation invariant factor $\exp\left[-(t-t')/\tau_\mu\right]$ and satisfy conditions (i), (ii), and (iii). So both reproduce the standard Cox model for time-independent covariates, as well as the instantaneous Cox model for longitudinal covariates with vanishing impact time scales, but they achieve this in distinct ways. We could have ensured a time-translation invariant kernel $\beta_\mu(t,t',s)$ by choosing in Equation (6.13) expressions for $A_\mu(t,s)$ and $B_\mu(t,s)$ that are independent of t. However, our models would then not reduce to the standard Cox model when covariates are constant. For $t > s$ we find that $\beta_\mu^{\mathrm{A}}(t,t',s)$ is independent of t. This describes an anomalous response: the system 'remembers' early changes in covariates without decay. This could describe e.g. irreversible damage to the organism. In contrast, $\beta_\mu^{\mathrm{B}}(t,t',s)$ retains a decaying dependence on t when $t > s$, with $\lim_{t\to\infty} \beta_\mu^{\mathrm{B}}(t,t',s) = a_\mu/s$. This could describe, for example, fluctuations in hormone levels that impact the hazard mostly in the short term, but also with persistent long-term effects.

Equations (6.16, 6.17) only hold for $s > 0$. In the data sets we study below there are some individuals whose longitudinal covariates are observed only once at the baseline time (i.e. their final observation time is $s = 0$). Given that Equations (6.16) and (6.17) cannot be used for such individuals, we must specify their association parameters $\beta_\mu(t)$ in some other way. Two possible options are a constant association, $\beta_\mu(t) = a_\mu$, or a decaying association, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$. Throughout the main paper we choose the former in the retarded kernel models. Results for the decaying association are presented in Appendix Section 6.10.

We note that we condition on knowledge of the covariates observed over a specific time interval $[0,s]$ in the model setup. As a consequence, the parameters in the retarded kernel models cannot necessarily be interpreted directly in terms of biophysical mechanisms. For example, τ_μ encapsulates both the possible decay in the physical effect of covariate μ, and the decay that occurs from conditioning on knowledge of the

covariate in the past. Parameter interpretations for the model therefore only make sense in a prediction context.

6.3.4.3 Maximum likelihood inference

As in the standard Cox model, we use maximum likelihood inference to determine the most plausible values of the model parameters based on the observed data. For simplicity, in this section we will mostly omit the superscript RK from the hazard function. We write θ for the full set of parameters, i.e. the model parameters $\{\tau_\mu, a_\mu\}$ described in Section 6.3.4.2 and the base hazard rate $h_0(t)$, and assume initially that for each sample i the covariates are known over the full time interval $[0, s_i]$. The optimal parameters are those for which the data likelihood $\mathcal{P}(\mathcal{D}|\theta)$ is maximised. For non-censored data this likelihood is given by

$$\mathcal{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}), \tag{6.18}$$

where $p(t|\theta, \mathcal{Z}^i_{[0,s_i]})$ is the probability density for individual i experiencing an event at time t given their covariate measurements. This probability density is expressed in terms of the parameterised hazard rate $h(t|\theta, \mathcal{Z}^i_{[0,s_i]})$ and the survival probability $S(t|\theta, \mathcal{Z}^i_{[0,s_i]}) = \exp[-\int_0^t \mathrm{d}t' h(t'|\theta, \mathcal{Z}^i_{[0,s_i]})]$ via

$$p(t|\theta, \mathcal{Z}^i_{[0,s_i]}) = h(t|\theta, \mathcal{Z}^i_{[0,s_i]}) S(t|\theta, \mathcal{Z}^i_{[0,s_i]}). \tag{6.19}$$

For right-censored data there are two contributions to the likelihood. Individuals for whom an event is observed at time $T_i = T_i^*$ contribute a density $p(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})$. Those that are censored at time $T_i = C_i$ contribute the survival probability $S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})$. Using the primary event indicator $\delta_i = I(T_i^* \leq C_i) \in \{0, 1\}$, the likelihood for censored data is then

$$\mathcal{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})^{\delta_i} S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}). \tag{6.20}$$

Upon defining $\Omega_{\mathrm{ML}}(\theta) = -\ln \mathcal{P}(\mathcal{D}|\theta)$, we can write the maximum likelihood parameter estimators as $\hat{\theta}_{\mathrm{ML}} = \mathrm{argmin}_\theta \Omega_{\mathrm{ML}}(\theta)$.

A full derivation of the maximum likelihood equations for models of the form in Equation (6.10) is provided in Appendix Section 6.6.1. Here we present only the results.

The maximum likelihood estimator of the base hazard rate is the direct analogue of the Breslow estimator [36]:

$$\hat{h}_0(t) = \frac{\sum_{i=1}^{N} \delta_i \delta(t - T_i)}{\sum_{i=1}^{N} I(t \in [0, T_i]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_i,t)} \beta_\mu(t,t',s_i) z_\mu^i(t') \mathrm{d}t'}}, \tag{6.21}$$

recalling from Section 6.3.1 that $I(A) = 1$ if A holds, and $I(A) = 0$ otherwise. The remaining parameters $\{a_\mu, \tau_\mu\}$ in Equations (6.16) and (6.17) are found by minimisation of

$$\begin{aligned}\Omega_{\mathrm{ML}}[\{a_\mu, \tau_\mu\}] \;=\; & \sum_{i=1}^{N} \delta_i \ln \left(\sum_{j=1}^{N} I(T_i \in [0, T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,T_i)} \beta_\mu(T_i,t',s_j) z_\mu^j(t') \mathrm{d}t'} \right) \\ & - \sum_{i=1}^{N} \delta_i \sum_\mu \int_0^{s_i} \beta_\mu(T_i, t', s_i) z_\mu^i(t') \mathrm{d}t', \end{aligned} \tag{6.22}$$

where we have disregarded terms that are independent of $\{a_\mu, \tau_\mu\}$. As in all Cox-type models, the final minimisation of Equation (6.22) with respect to the remaining parameters (here, the associations and time-scales) must be performed numerically, for example using Powell's method [37].

6.3.4.4 Dynamic Prediction

Using the maximum likelihood estimates $\hat{\theta}_{\mathrm{ML}}$ for the model parameters, we can use the retarded kernel models to estimate the quantity $\pi(u|\mathcal{Z}_{[0,s]}, s)$ in Equation (6.4), representing the probability that a subject has not experienced an event by time $u > s$, conditional on their survival to s and on their covariate values $\mathcal{Z}_{[0,s]}$ up to that time. That is,

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}_{[0,s]}, s) = \exp\left\{ - \int_s^u \hat{h}^{\mathrm{RK}}(t'|\mathcal{Z}_{[0,s]}) \mathrm{d}t' \right\}, \tag{6.23}$$

with $\hat{h}^{\mathrm{RK}}(t|\mathcal{Z}_{[0,s]})$ as defined by Equation (6.10), with kernels of the form in Equations (6.16, 6.17) and with the ML estimators for the parameters in those kernels. Using the ML estimator in Equation (6.21) of the base hazard rate we can perform the integration in Equation (6.23) to find

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}_{[0,s]}, s) = \exp\left\{ - \sum_{j=1}^{N} \frac{\delta_j \; I(T_j \in [s, u]) \mathrm{e}^{\sum_{\mu=1}^{p} \int_0^s \hat{\beta}_\mu(T_j,t',s) z_\mu(t') \mathrm{d}t'}}{\sum_{k=1}^{N} I(T_j \in [0, T_k]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_k,T_j)} \hat{\beta}_\mu(T_j,t',s_k) z_\mu^k(t') \mathrm{d}t'}} \right\}, \tag{6.24}$$

where $\hat{\beta}_\mu(t, t', s)$ indicates the association kernel obtained from the ML estimators of the parameters $\{a_\mu, \tau_\mu\}$. In the numerator we have used the fact that the prefactor $I(T_j \in [s, u])$ ensures that $\min(s, T_j) = s$.

6.3.4.5 Covariate interpolation

So far, we have defined the retarded kernel models conditional on covariate trajectories $\mathcal{Z}_{[0,s]}$ over the entire interval $[0, s]$. In reality, we do not have full knowledge of these trajectories. Instead for each subject i we have a finite number of discrete measurements that coincide with follow up appointments, $\mathcal{Z}^i = \{z^i_\mu(t_{i\ell});\ \mu = 1, \ldots, p, \ell = 1, \ldots, n_i, t_{i\ell} \in [0, s_i]\}$. In order to perform the integrals in Equations (6.22) and (6.24) we must interpolate between these discrete observed values.

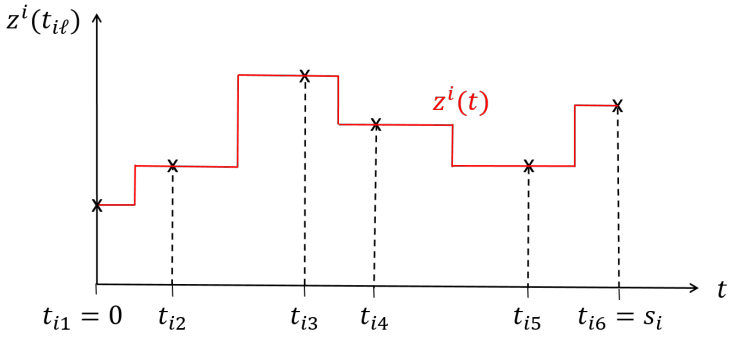


Figure 6.4: An illustration of the interpolation method for covariates. For each subject i, there are a discrete number of covariate observations. The observation times $t_{i\ell}$ are labelled on the horizontal axis. The covariate measurement at each observation time is indicated by a cross. The solid line shows the interpolated covariate trajectory based on these discrete observations. The value of a covariate at time $t \neq t_{i\ell}$ is taken to be equal to the observed value of the covariate at the observation time closest to t. This yields a step function that changes value half way between each pair of consecutive observations.

We choose a simple 'nearest neighbour' approach, that is we set $z^i_\mu(t) = z^i_\mu(t_{i\ell})$ where $t_{i\ell}$ is the observation time closest to t. The approximate covariate trajectory is then a step function that changes value half way between each pair of consecutive observation times. Figure 6.4 illustrates this idea. Using this method, the integrals in Equations (6.22) and (6.24) can be evaluated analytically (see Appendix Section 6.6.3). This reduces the computational effort required to perform the minimisation and the dynamic prediction. Other, smoother interpolation procedures such as Gaussian convolutions [38, 39] are also possible and may improve estimations (at some computational cost). While interpolation makes assumptions about the values of the covariate within the

observation interval $[0, s_i]$, we do not make assumptions about the covariates after the final observation time s_i.

6.4 Application to clinical data

6.4.1 Methods

6.4.1.1 Training and test data

For each of the data sets we assess the predictive accuracy of the different dynamic prediction models by splitting the data in half into a training data set and a test data set. Each model is fitted to the training data and the resulting model is used to make survival predictions about individuals in the test data set. Predictive accuracy is assessed by comparing these predictions to the true event times in the test data (see Section 6.4.1.2). The procedure was repeated for 20 random splits of the data and the corresponding prediction error was averaged over these repetitions.

6.4.1.2 Measuring predictive accuracy

Following Rizopoulos et al (2017) [2], we quantify the predictive accuracy of the different models using the expected error of predicting future events. Dynamic prediction is concerned with predicting the survival of individuals to a given time u, based on their survival to some earlier time $t < u$, and covariate measurements for the individual up to this time. The expected prediction error for a given 'prediction time' u and 'base time' t is then defined as [40]

$$\mathrm{PE}(u|t) = \mathbb{E}\left[\mathcal{L}\{N_i(u) - \pi(u|\mathcal{Z}^i_{[0,t]}, t)\}\right] \tag{6.25}$$

where $N_i(u) = I(T_i^* > u)$ is the true event status of subject i at time u, and $\pi(u|\mathcal{Z}^i_{[0,t]}, t)$ is the model's predicted survival probability for subject i based on information about this subject (covariate measurements and survival status) up to the base time t. The notation $\mathbb{E}$ stands for an average over the distribution of covariates and event times. $\mathcal{L}(.)$ denotes a loss function which defines how we measure the difference between survival status and predicted survival probability. Commonly choices are $\mathcal{L}(x) = |x|$ and the squared loss $\mathcal{L}(x) = x^2$ [2, 31, 40]. We choose the latter. The definition of

prediction error is such that $\text{PE}(u|t) = 0$ if the survival status of all individuals is predicted with full accuracy (i.e. $\pi(u|\mathcal{Z}^i_{[0,t]}, t) = 1$ for all subjects who are alive at time u and $\pi(u|\mathcal{Z}^i_{[0,t]}, t) = 0$ for subjects who are dead by time u). If the reverse is true ($\pi(u|\mathcal{Z}^i_{[0,t]}, t) = 1$ for subjects who are dead at time u and $\pi(u|\mathcal{Z}^i_{[0,t]}, t) = 0$ for subjects who are alive) then $\text{PE}(u|t) = 1$. We obtain $\text{PE}(u|t) = 0.25$ if every individual has predicted survival probability $\pi(u|\mathcal{Z}^i_{[0,t]}, t) = 0.5$.

Again following Rizopoulos et al (2017) [2], in this paper we use the overall prediction error $\text{PE}(u|t)$ proposed by Henderson et al (2002) [31], that in addition takes into account censoring,

$$\begin{aligned}\widehat{\text{PE}}(u|t) = \frac{1}{n(t)} \sum_{i;T_i \geq t} & I(T_i \geq u)\mathcal{L}\{1 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\} + \delta_i I(T_i < u)\mathcal{L}\{0 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\} \\ & + (1 - \delta_i)I(T_i < u)\Big[\hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, T_i)\mathcal{L}\{1 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\} \\ & + \{1 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, T_i)\}\mathcal{L}\{0 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\}\Big]. \quad (6.26)\end{aligned}$$

The sum extends over the $n(t)$ subjects in the test data set who are still at risk at time t. The first term of Equation (6.26) corresponds to individuals in the test data who are still alive after time u. These have survival status $N_i(u) = 1$, and therefore contribute a loss function based on the difference between their estimated survival probability and 1, i.e. $\mathcal{L}\{1 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\}$. The second term refers to individuals who have experienced an event by time u (i.e. $T_i = T_i^* < u$). Their survival status is 0 and therefore they contribute a loss function $\mathcal{L}\{0 - \hat{\pi}(u|\mathcal{Z}^i_{[0,t]}, t)\}$. The final term represents individuals who were censored before time u (i.e. $T_i = C_i < u$) so we do not know their survival status at time u. Here the estimated probability of survival based on information up to time t is compared with the probability of survival given that we know subject i survived up until their censoring time $T_i \geq t$.

To compare the predictive accuracy of joint modelling, landmarking and the retarded kernel approach we insert into Equation (6.26) the respective estimators $\hat{\pi}^{\text{JM}}(u|\mathcal{Z}^i_{[0,t]}, t)$, $\hat{\pi}^{\text{LM}}(u|\mathcal{Z}^i_{[0,t]}, t)$ and $\hat{\pi}^{\text{RK}}(u|\mathcal{Z}^i_{[0,t]}, t)$. This requires that we calculate the probability of a subject's survival to time u, based on survival and covariate observations until a general base time $t < u$ that need not be the individual's final observation time s_i. For the joint model and landmarking estimators we replace the final observation time with t in Equations (6.7) and (6.9). For the retarded kernel estimator $\hat{\pi}^{\text{RK}}(u|\mathcal{Z}^i_{[0,t]}, t)$ in Equation (6.24) we replace $I(T_j \in [s_i, u])$ with $I(T_j \in [t, u])$ since we know subject i is

alive until t. However, we only have covariate observations up to the latest observation time $t_{i\ell}$ that is $\leq t$. In line with our chosen interpolation procedure we only integrate the covariate trajectory up to this time. Specifically, for any general base time t we have

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,t]},t)=\exp\left\{-\sum_{j=1}^{N}\frac{\delta_j\ I(T_j\in[t,u])\mathrm{e}^{\sum_\mu\int_0^{\max\{t_{i\ell}:t_{i\ell}\leq t\}}\hat{\beta}_\mu(T_j,t',\max\{t_{i\ell}:t_{i\ell}\leq t\})z^i_\mu(t')\mathrm{d}t'}}{\sum_{k=1}^{N}I(T_j\in[0,T_k])\mathrm{e}^{\sum_\mu\int_0^{\min(s_k,T_j)}\hat{\beta}_\mu(T_j,t',s_k)z^k_\mu(t')\mathrm{d}t'}}\right\}, \tag{6.27}$$

where index i labels the individual (in the test data) for whom we are making predictions, while the sums over j and k refer to individuals in the training data set used for inference. The integral limit $\max\{t_{i\ell} : t_{i\ell} \leq t\}$ labels the last observation time of individual i before (or at) the base time t.

The term $\hat{\pi}(u|\mathcal{Z}^i_{[0,t]},T_i)$ in Equation (6.26) represents the probability of survival to u given subject i survived to their censoring time $T_i = C_i$. To calculate this using the retarded kernel model we replace $I(T_j\in[t,u])$ with $I(T_j\in[T_i,u])$ in Equation (6.27). For joint modelling $\hat{\pi}^{\mathrm{JM}}(u|\mathcal{Z}^i_{[0,t]},T_i)$ is obtained by replacing $\mathcal{Z}_{[0,s]}$ with $\mathcal{Z}^i_{[0,t]}$ and by replacing the condition $T^* > s$ with $T^*_i > T_i$ in Equation (6.7). Since this term is only calculated for censored individuals ($T_i = C_i$), the condition $T^*_i > T_i$ means 'the true event time of individual i is greater than their censoring time'. Finally, for landmarking we use $\hat{\pi}^{\mathrm{LM}}(u|\mathcal{Z}^i_{[0,t]},T_i) = \hat{\pi}^{\mathrm{LM}}(u|\mathcal{Z}^i_{[0,t]},t)/\hat{\pi}^{\mathrm{LM}}(T_i|\mathcal{Z}^i_{[0,t]},t)$ which is equivalent to replacing s with t in Equation (6.9) except in the integral limits where we replace $\int_s^u$ with $\int_{T_i}^u$.

To perform the prediction error calculation for the retarded kernel models we use our own `C++` code following Equations (6.26) and (6.27). For the joint model and landmarking model we use a version of the function `prederrJM` in the `JMbayes` package subject to minor modifications (see Appendix Section 6.9 for details).

6.4.1.3 Fixed base time

First we compare the predictive accuracy of the three methods by specifying a fixed base time t and varying the prediction time u. Based on Figures 6.1 and 6.3, for the PBC and Liver data sets we choose a fixed base time of $t = 3$ years. This value is chosen so that a large number of individuals are still alive after this time (and we can hence make predictions about them), but also so that these individuals have had their

covariates measured multiple times before this time. We then vary the prediction time u from the base time $t = 3$ years in steps of 0.2 years up to 8 years for the PBC data, and up to 10 years for the Liver data. For the AIDS data set we choose $t = 6$ months as the base time, so that most individuals have been observed three times. We then vary the prediction time u from this base time up to 18 months in steps of 0.2 months.

6.4.1.4 Fixed prediction window

In our second test, we vary the base time t, while keeping the prediction window $w = u - t$ fixed (i.e. the time difference between prediction and base time). Since we are varying the base time t, we must then fit a new landmark model for each choice of t (where the landmark time $\upsilon = t$). On the other hand, for the retarded kernel approach and the joint model we need only fit the model once, and can make the error assessments at each iteration using this single fitted model.

Based on previous analysis of the PBC and Liver data [9, 31] we choose three prediction windows: $w_1 = 1$ year, $w_2 = 2$ years, and $w_3 = 3$ years. Given the event time distributions, we do not make predictions for either data set beyond $u = 10$ years. Therefore, for w_1 we vary the base time from $0 - 9$ years, for w_2 we vary it from $0 - 8$ years and for w_3 this is $0 - 7$ years. In all cases we increase the base time in increments of 0.2 years.

Based on the event time distribution of the AIDS data, we choose prediction windows $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. Here covariates are observed at 0, 2, 6, 12, and 18 months only. As a result, predictions will only be updated at these time steps, and we can only make a small number of distinct measurements of predictive accuracy. Due to the event times in the AIDS data set, we do not make predictions past 18 months. Therefore, for window w_1 we use base times $t = 0, 2, 6, 12$ months and for windows w_2 and w_3 we use $t = 0, 2, 6$ months only.

6.4.2 PBC data set

We fit each model to the PBC training data set using $p = 3$ time-dependent covariates and a single fixed covariate: $z_1^i(t)$ denotes log serum bilirubin, $z_2^i(t)$ denotes log serum albumin, $z_3^i(t)$ is log prothrombin time, and the fixed covariate ζ_1^i is the subject's age at baseline.

The PBC data set contains event-time information for two events, death and liver transplant. The most appropriate way of analysing this data is to use a competing risks model. However, for simplicity we here treat the transplant event as a censoring event. Another simple way to analyse this data is to treat the two events as a single composite event. We provide the results of the latter analysis in Appendix Section 6.8. The two analyses are found to give similar results.

6.4.2.1 Models

For the joint model we first fit a simple multivariate linear mixed model to each of the three time-dependent covariates,

$$z^i_\mu(t) = m^i_\mu(t) + \varepsilon^i_\mu(t) = \eta_{\mu,0} + b^i_{\mu,0} + (\eta_{\mu,1} + b^i_{\mu,1})t + \varepsilon^i_\mu(t), \tag{6.28}$$

where the random effects $\boldsymbol{b}^i$ are assumed to follow a joint multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$.

Figure 6.1 suggests that the covariate trajectories in the PBC data may be non-linear for some individuals. Hence, for extra flexibility we also fit a second joint model that includes natural cubic splines in both the fixed and random effects parts of the model. Following Rizopoulos (2016) [25], the log serum bilirubin ($\mu = 1$) is modelled using natural cubic splines with 2 degrees of freedom,

$$\begin{aligned} z^i_1(t) &= m^i_1(t) + \varepsilon^i_1(t) \\ &= \eta_{1,0} + b^i_{1,0} + (\eta_{1,1} + b^i_{1,1})B^n_1(t,\lambda) + (\eta_{1,2} + b^i_{1,2})B^n_2(t,\lambda) + \varepsilon^i_1(t) \end{aligned} \tag{6.29}$$

where $\{B^n_k(t,\lambda); k = 1,2\}$ denotes the B-spline basis matrix for a natural cubic spline of time [9, 41]. We write analogous equations for both the log albumin and the log prothrombin covariates. Again, the random effects of all three longitudinal covariates are assumed to follow a joint multivariate normal distribution.

For both the linear and spline longitudinal models, the hazard function of the survival sub-model in the joint modelling framework is

$$h^{\text{JM}}(t|\mathcal{M}^i_{[0,t]}) = h_0(t)\exp\left\{\gamma_1\zeta^i_1 + \alpha_1 m^i_1(t) + \alpha_2 m^i_2(t) + \alpha_3 m^i_3(t)\right\}, \tag{6.30}$$

where we recall from Section 6.3.2 that $\mathcal{M}^i_{[0,t]} = \{m^i_\mu(t');\ \mu = 1,\ldots,p,\ t' \in [0,t]\}$ denotes the history of the 'true' (unobserved) longitudinal covariates up to time t for

subject i. For the landmark model the hazard is instead specified for a given landmark time, υ,

$$h^{\text{LM}}(t|\mathcal{Z}^i, \upsilon) = h_0(t|\upsilon) \exp\left\{\gamma_1\zeta_1^i + \alpha_1(\upsilon)\tilde{z}_1^i(\upsilon) + \alpha_2(\upsilon)\tilde{z}_2^i(\upsilon) + \alpha_3(\upsilon)\tilde{z}_3^i(\upsilon)\right\}, \quad (6.31)$$

where $\tilde{z}_\mu^i(\upsilon)$ is again the last observed value of covariate μ for patient i before time υ.

For the retarded kernel approach, we specify the hazard function as

$$h^{\text{RK}}(t|\mathcal{Z}^i_{[0,s_i]}) = h_0(t) \exp\left\{\gamma_1\zeta_1^i + \int_0^{\min(s_i,t)} \left(\beta_1(t,t',s_i)z_1^i(t') + \beta_2(t,t',s_i)z_2^i(t') + \beta_3(t,t',s_i)z_3^i(t')\right) \mathrm{d}t'\right\}. \quad (6.32)$$

The parameterisations of the time-dependent association parameters $\beta_\mu(t,t',s)$ are given in Equations (6.16) and (6.17) for models A and B, respectively.

6.4.2.2 Results

Figure 6.5 shows plots of the overall prediction error $\widehat{\text{PE}}(u|t)$ against the prediction time u for a fixed base time of $t = 3$ years averaged over the 20 random splits of the data. Results for the linear joint model, spline joint model, landmarking model and models A and B of the retarded kernel approach are plotted on the same graph. All five models have similarly accurate predictions up to $u = 5$ years. For later prediction times, the standard approaches have a lower average prediction error than the retarded kernel models. The largest disparity in prediction error is observed at $u = 8$ years between the spline joint model ($\widehat{\text{PE}}(u|t) = 0.126$) and the retarded kernel models which both have $\widehat{\text{PE}}(u|t) = 0.146$.

Plots of the average prediction error $\widehat{\text{PE}}(u|t)$ against the base time t are shown in Figure 6.6, for fixed prediction windows $w_1 = 1$ year, $w_2 = 2$ years, and $w_3 = 3$ years. Again results for the five different models are plotted in the same graphs. For the shortest prediction window w_1, all models are similarly accurate for base times up to $t = 7.5$ years, after which the landmarking model performs slightly worse than the others. For the larger prediction windows, models A and B of the retarded kernel approach show slightly larger prediction errors than the other models over the range $t = 0 - 5$ years. At larger base times the joint models and retarded kernel models again exhibit similar prediction errors while landmarking has the largest error. The largest

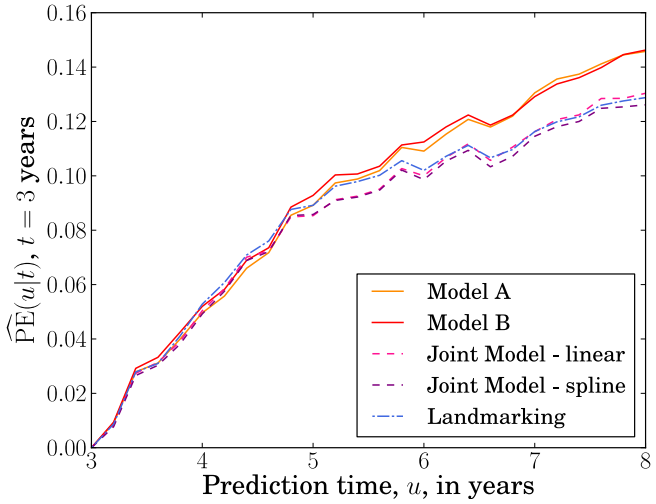


Figure 6.5: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of prediction time u (in years) for the PBC data with fixed base time $t = 3$ years. Prediction error is calculated for u values from 3 to 8 years, with 0.2 year increments. A squared loss function was used in Equation (6.26). The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models (one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines).

difference in performance occurs at $t = 9$ years for the prediction window w_1, between the spline joint model ($\widehat{\mathrm{PE}}(u|t) = 0.062$) and the landmarking model ($\widehat{\mathrm{PE}}(u|t) = 0.107$).

The results of the above tests suggest that, for the PBC data, the retarded kernel approach performs as well as existing methods for prediction windows < 2 years but less well for larger windows. However, as the base time increases, the retarded kernel models behave similarly to the joint models while the landmarking model exhibits the highest prediction error. Care should be taken when interpreting these results, as we have not used a competing risks model in our analysis. This data set does, however, serve as an illustration that with only a modest drop in accuracy the retarded kernel model can serve as a simpler alternative to joint modelling when considering multiple longitudinal covariates. Unlike the landmarking approach, the retarded kernel model takes into account the full history of covariate observations which, along with the fact that landmarking discards more data as the base time increases, may explain why the landmarking model performs worst for later base times.

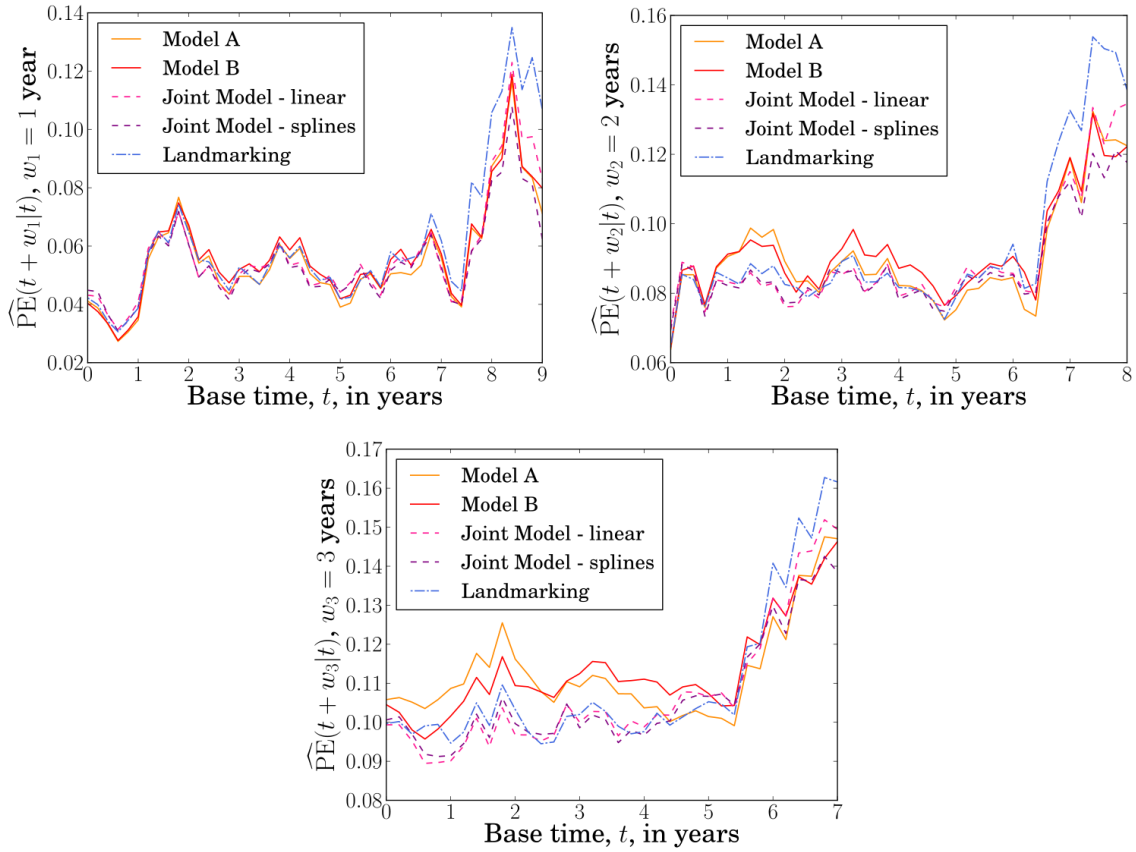


Figure 6.6: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in years) for the PBC data, with prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The prediction error is calculated for t ranging from 0 to 9,8 or 7 years for w_1, w_2 and w_3 respectively, with 0.2 year increments. A squared loss function was used in Equation (6.26). The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. Results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models; one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines.

6.4.3 AIDS data set

In the AIDS data set we focus on a single longitudinal covariate, the CD4 count $z_1^i(t)$. We also include four fixed binary covariates: drug group ($\zeta_1^i = 1$ for ddI and $\zeta_1^i = 0$ for ddC), gender ($\zeta_2^i = 1$ for male and $\zeta_2^i = 0$ for female), PrevOI ($\zeta_3^i = 1$ for AIDS diagnosis at study entry and $\zeta_3^i = 0$ for no AIDS diagnosis) and Stratum ($\zeta_4^i = 1$ for AZT failure and $\zeta_4^i = 0$ for AZT intolerance). See Section 6.2.2 for a description of these variables.

6.4.3.1 Models

The joint modelling framework allows us to model the dependence of CD4 count on the patients drug group. Following Rizopoulos (2012) [9] we fit the linear mixed model,

$$\begin{aligned} z_1^i(t) &= m_1^i(t) + \varepsilon_1^i(t) \\ &= \eta_{1,0} + b_{1,0}^i + (\eta_{1,1} + b_{1,1}^i)t + \eta_{1,2}\zeta_1^i t + \varepsilon_1^i(t), \end{aligned} \tag{6.33}$$

where the term $\eta_{1,2}\zeta_1^i t$ denotes the effect of the interaction of treatment (drug group) with time. As usual, the random effects $\boldsymbol{b}^i$ are assumed to follow a normal distribution. To complete the joint model, the hazard function is then chosen as

$$h^{\mathrm{JM}}(t|\mathcal{M}_{[0,t]}^i) = h_0(t)\exp\left\{\gamma_1\zeta_1^i + \gamma_2\zeta_2^i + \gamma_3\zeta_3^i + \gamma_4\zeta_4^i + \alpha_1 m_1^i(t)\right\}. \tag{6.34}$$

For the landmark model with landmark time υ one has

$$h^{\mathrm{LM}}(t|\mathcal{Z}^i, \upsilon) = h_0(t|\upsilon)\exp\left\{\gamma_1\zeta_1^i + \gamma_2\zeta_2^i + \gamma_3\zeta_3^i + \gamma_4\zeta_4^i + \alpha_1(\upsilon)\tilde{z}_1^i(\upsilon)\right\}, \tag{6.35}$$

and for the retarded kernel approach we specify the survival model as follows

$$h^{\mathrm{RK}}(t|\mathcal{Z}_{[0,s_i]}^i) = h_0(t)\exp\left\{\gamma_1\zeta_1^i + \gamma_2\zeta_2^i + \gamma_3\zeta_3^i + \gamma_4\zeta_4^i + \int_0^{\min(s_i,t)}\beta_1(t,t',s_i)z_1^i(t')\mathrm{d}t'\right\}. \tag{6.36}$$

As before, the parameterisations of $\beta_\mu(t,t',s)$ in models A and B are given in Equations (6.16) and (6.17) respectively.

6.4.3.2 Results

The plots of $\widehat{\mathrm{PE}}(u|t)$ against prediction time u with base time $t = 6$ months are shown in Figure 6.7 for the four models. As before, the data for $\widehat{\mathrm{PE}}(u|t)$ is an average over

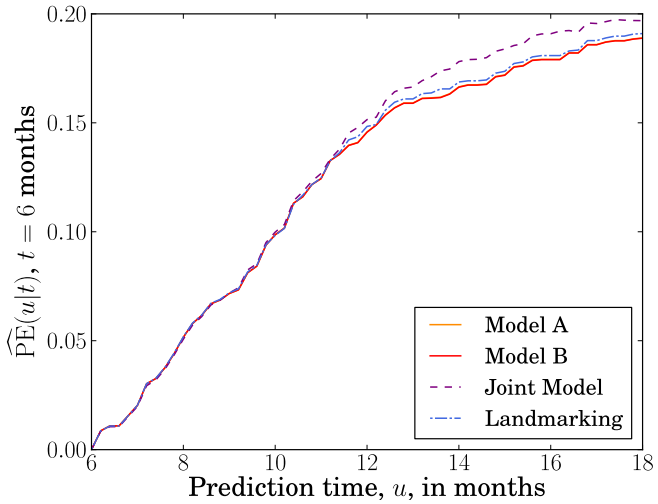


Figure 6.7: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in months) for the AIDS data with fixed base time $t = 6$ months. This error is calculated for u ranging from 6 to 18 months, at 0.2 month intervals. In Equation (6.26) a squared loss function was used. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. The results from model A cannot be seen because they overlap with the results from model B.

20 random splits of the data into training and test sets. All models show comparable accuracy up to $u = 11$ months. After this time, the joint model shows slightly worse prediction than the other three (whose accuracies remain almost equal). The largest difference in predictive error occurs at $u = 16.2$ months, between models A and B of the retarded kernel approach on the one hand and the joint model on the other. At this value of u, both versions of the retarded kernel model lead to $\widehat{\mathrm{PE}}(u|t) = 0.179$ while the joint model has $\widehat{\mathrm{PE}}(u|t) = 0.192$.

Figure 6.8 shows plots of $\widehat{\mathrm{PE}}(u|t)$ against base time for the AIDS data set with three prediction windows, $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. For the shortest prediction window w_1, all four models have similar prediction error at $t = 0$ and 2 months. The joint model has the largest error at $t = 6$ months (where models A and B are lowest), but has the same error as the retarded kernel models at $t = 12$ months (where landmarking has the highest error). For windows w_2 and w_3 the joint model demonstrates the worst prediction at all base times. The other three models exhibit similar errors at $t = 0$ for both these windows as well as at $t = 2$ for window w_2. In all other scenarios, models A and B of the retarded kernel approach have the lowest prediction error. The largest difference in prediction error is for w_2 at $t = 6$

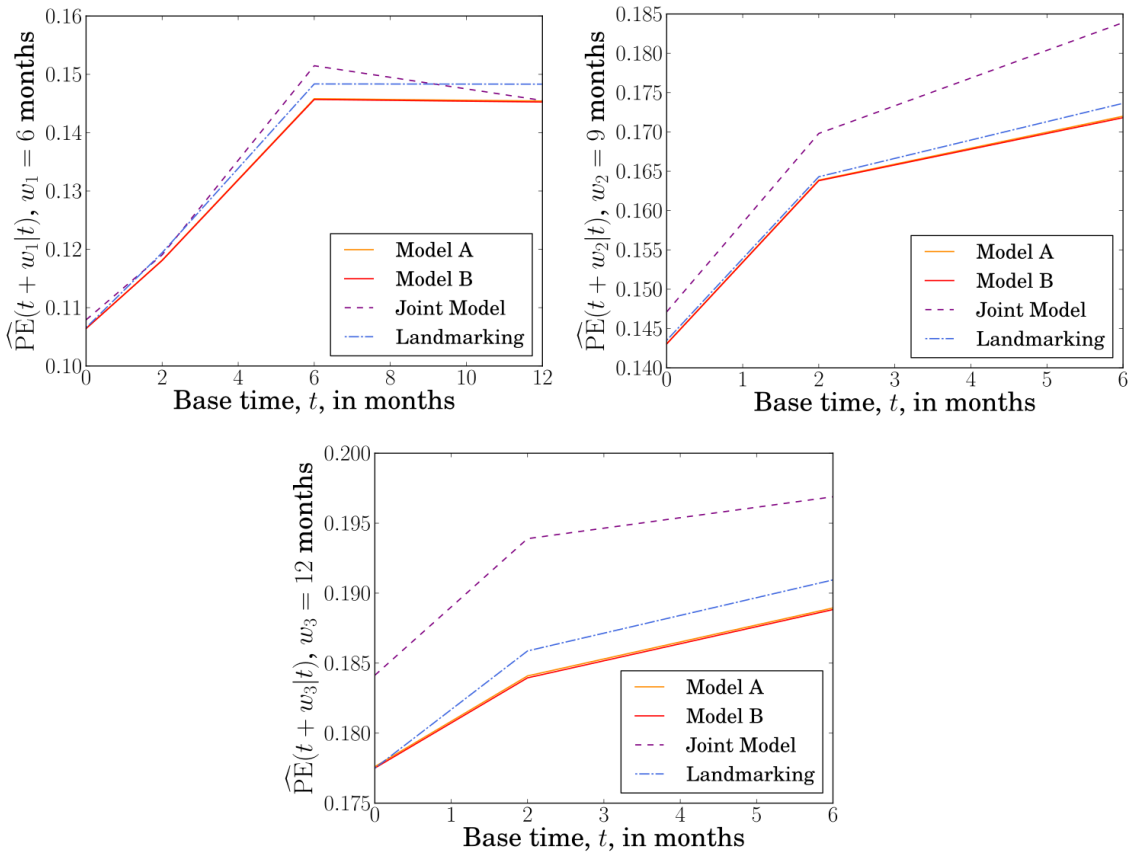


Figure 6.8: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in months) for the AIDS data with three fixed prediction windows: $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. The prediction times are $u = t + w$. Observations are made at times $0, 2, 6, 12, 18$ months for all individuals in this data set. Prediction errors are hence only updated at these time points. For prediction window w_1, prediction error is measured for $t = 0, 2, 6$ and 12 months. For windows w_2 and w_3, the error is measured at $t = 0, 2$ and 6 months only. In Equation (6.26) we used a squared loss function. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. The results from model A cannot be seen clearly because they overlap with the results from model B.

months where the joint model has $\widehat{\text{PE}}(u|t) = 0.184$, landmarking has $\widehat{\text{PE}}(u|t) = 0.174$ and the retarded kernel models both have $\widehat{\text{PE}}(u|t) = 0.172$.

The above results suggest that, for the AIDS data set, the joint model has the worst predictive accuracy overall, while the two retarded kernel models perform best.

6.4.4 Liver data set

For the liver data set we model prothrombin index as our one longitudinal covariate $z_1^i(t)$, and drug group as our single fixed covariate ζ_1^i. The fixed covariate is defined such that $\zeta_1^i = 1$ for individuals in the treatment (prednisone) group, and $\zeta_1^i = 0$ for those in the placebo group.

6.4.4.1 Models

Following Rizopoulos (2012) [9], we define a flexible longitudinal model for the subject-specific prothrombin trajectories, using natural cubic splines with different average profiles for each drug group. Rizopoulos (2012) [9] also suggests to include a separate indicator variable of the baseline measurement, to capture sudden changes in the prothrombin index in the early part of follow up. The longitudinal model then takes the form

$$\begin{aligned} z_1^i(t) &= m_1^i(t) + \varepsilon_1^i(t) \\ &= \eta_{1,0} + b_{1,0}^i + (\eta_{1,1} + b_{1,1}^i)B_1^n(t,\lambda) + (\eta_{1,2} + b_{1,2}^i)B_2^n(t,\lambda) \\ &\quad +(\eta_{1,3} + b_{1,3}^i)B_3^n(t,\lambda) + \eta_{1,4}\zeta_1^i B_1^n(t,\lambda) + \eta_{1,5}\zeta_1^i B_2^n(t,\lambda) + \eta_{1,6}\zeta_1^i B_3^n(t,\lambda) \\ &\quad +\eta_{1,7}\zeta_1^i + \eta_{1,8} I(t = t_{i,1}) + \eta_{1,9}\zeta_1^i I(t = t_{i,1}) + \varepsilon_1^i(t) \end{aligned} \tag{6.37}$$

where $I(t = t_{i,1})$ is the indicator variable for the baseline time and, as before, $\{B_k^n(t,\lambda); k = 1,2,3\}$ is the B-spline basis matrix for a natural cubic spline of time. This time, two internal knots are placed at 33% and 66.7% percentiles of the follow up times. The random effects are assumed to have a diagonal covariance matrix.

The hazard functions for the joint model and the landmark model (with landmark time υ) are then

$$h^{\text{JM}}(t|\mathcal{M}_{[0,t]}^i) = h_0(t)\exp\left\{\gamma_1\zeta_1^i + \alpha_1 m_1^i(t)\right\}, \tag{6.38}$$

$$h^{\text{LM}}(t|\mathcal{Z}^i, \upsilon) = h_0(t|\upsilon)\exp\left\{\gamma_1\zeta_1^i + \alpha_1(\upsilon)\tilde{z}_1^i(\upsilon)\right\}. \tag{6.39}$$

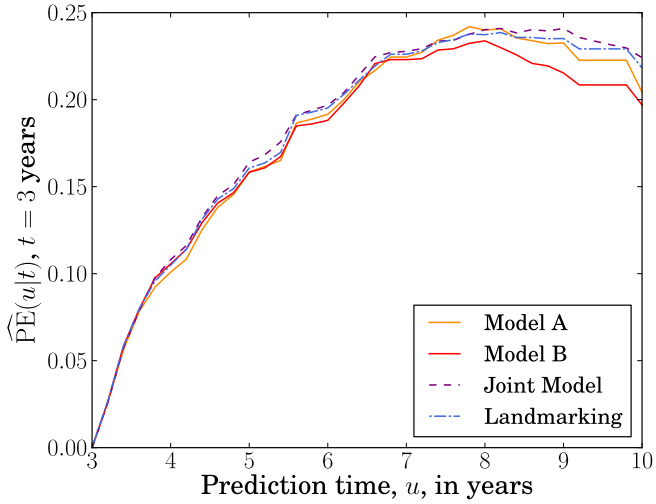


Figure 6.9: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in years) for the Liver data with fixed base time $t = 3$ years. This error is calculated for u ranging from 3 to 10 years, with 0.2 year increments. In Equation (6.26) we used a squared loss function. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model.

For the retarded kernel models we have

$$h^{\mathrm{RK}}(t|\mathcal{Z}^i_{[0,s_i]}) = h_0(t)\exp\left\{\gamma_1\zeta_1^i + \int_0^{\min(s_i,t)}\beta_1(t,t',s_i)z_1^i(t')\mathrm{d}t'\right\}. \tag{6.40}$$

6.4.4.2 Results

Figure 6.9 shows prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of u for a fixed base time $t = 3$ years for all four models. Again, each value of $\widehat{\mathrm{PE}}(u|t)$ is an average over the 20 random splits of the data into training and test data sets. The four models show similar prediction error up to $u = 7$ years. After this point, retarded kernel model B has slightly lower prediction error than the other models. For example, at $t = 9.2$ years, the joint model has $\widehat{\mathrm{PE}}(u|t) = 0.236$, the landmark model has $\widehat{\mathrm{PE}}(u|t) = 0.229$, model A has $\widehat{\mathrm{PE}}(u|t) = 0.223$ and model B has $\widehat{\mathrm{PE}}(u|t) = 0.208$.

Plots of average $\widehat{\mathrm{PE}}(u|t)$ against base time t are shown in Figure 6.10 for fixed prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. For all three windows the four models exhibit very similar accuracy levels, with no model showing consistently superior predictions.

For the liver data set the above results suggest that the retarded kernel models have a predictive accuracy that is comparable to those of standard methods.

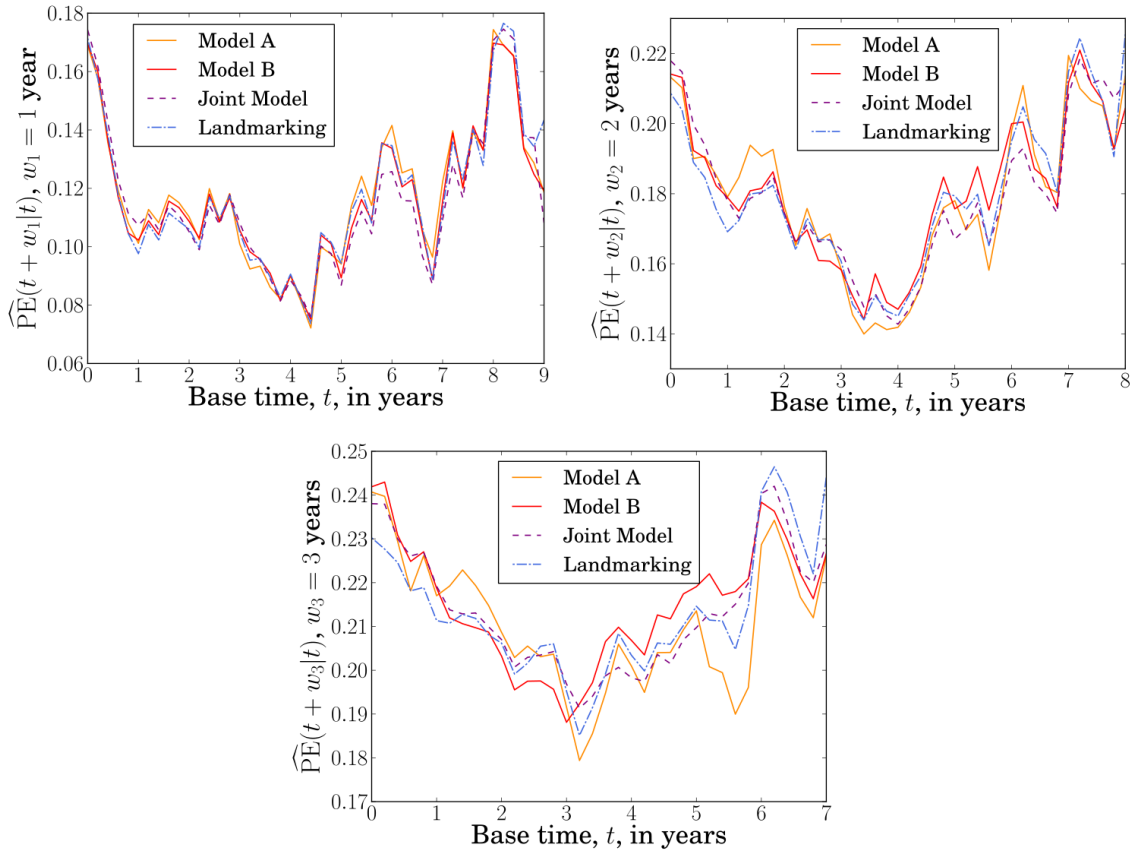


Figure 6.10: Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted against base time t (in years) for the Liver data with three fixed prediction windows, $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The error is calculated for t ranging from 0 to 9,8 or 7 years, for w_1, w_2 and w_3 respectively, with 0.2 year intervals. In Equation (6.26) a squared loss function was used. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model.

6.5 Summary and Discussion

In this work we propose a retarded kernel approach to dynamic prediction in survival analysis. In terms of complexity, our method comes somewhere between the two standard approaches, joint modelling and landmarking. It is more parsimonious than joint modelling, as it does not model the longitudinal covariate trajectory, and it makes no assumptions about the base hazard rate. This makes the method more practical for data with multiple time-dependent covariates. The retarded kernel approach conditions only on the observed covariates and, unlike joint modelling, makes no assumptions about covariate values in the future. This makes it more suitable for covariates that cannot easily be predicted, such as categorical ones. Compared to landmarking, the retarded kernel approach makes use of more of the available data. In landmarking, a new model is fitted at each landmark time, discarding individuals in the data set who have experienced the event before this landmark time. Additionally, standard landmarking only uses covariate measurements that are most recent before the landmark time. In contrast, the retarded kernel approach fits a single model that incorporates information from all individuals in the data set, using the full history of their covariate measurements. We note however, that in extensions of the landmarking approach one could, in principle, fit a landmarking model using multiple covariate measurements, for example, by replacing the most recent measurement with the mean value of measurements up to that time.

The retarded kernel approach relies on parameterisation of the association kernels $\beta_\mu(t, t', s)$. In this work we focused on two specific parameterisations, motivated by practical considerations. We required that our models reduce to the standard Cox model for static covariates, and that they contain the instantaneous Cox model as a special case, so that they are natural extensions of familiar models. However, alternative parameterisations or extensions (if demanded by the data at hand) can be incorporated without much effort. For example, one could include a 'hard' time delay between covariate variations and their effect on hazard, or use parameterisations that favour time-translation invariance over consistency with standard Cox models. Furthermore, one need not be restricted to the exponential model proposed here but could instead make use of a more flexible parametric model for the decay.

In tests on medical data, we found that the retarded kernel approach performs similarly to the two standard approaches in terms of predictive accuracy. Depending on the data set, base time and prediction window, each method (joint modelling, landmarking, or retarded kernels) had at some point the highest or the lowest prediction error; none appeared to be consistently superior or inferior across the scenarios we tested. These initial comparisons indicate that the retarded kernel method is a reasonable approach to dynamic prediction, worthy of further research and development.

An important factor to consider when assessing model performance is overfitting. This issue occurs as the ratio of the number of model parameters to the number of data points used to fit the model increases. The model then reflects the fitted data set too closely which leads to poor performance in terms of parameter estimation and prediction. In this work, we compared models of varying complexity, for example joint models involve highly parameterised longitudinal trajectories while the retarded kernel approach uses multiple parameters in the survival model. Observed differences in predictive accuracy between the various models could therefore be due to overfitting in some models. It is possible that for very large data sets (i.e. where the number of data points is much larger than the number of parameters in any of the models) that all methods perform equally well. In future work it would be interesting to test this, for example using simulations with very large values of N.

Future development of the retarded kernel approach could also involve attempts to correct for overfitting within the model. For example, regularisation methods are commonly used for this purpose in survival analysis with high-dimensional covariates [42, 43]. Here, one adds a penalty term to the maximum likelihood equation to suppress the number or magnitude of the model parameters. Since our method is an extension of the standard Cox model and is based on maximum likelihood estimation, a similar approach may also be possible for the retarded kernel model. Furthermore, one of us [44, 45] has previously employed methods from statistical mechanics in order to correct for overfitting in the standard Cox model. Extending these methods for the retarded kernel approach would be another interesting avenue for future research.

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) reporting guidelines [46] highlight the importance of transparency in the development of prediction models. In order for others to

independently validate a model, the equations used for prediction should be relatively easy to explicitly write down. The equation for predictive probability in the retarded kernel model (6.24) involves integration of the covariates and their associations over time. The result of this integration depends on the parameterisation of the association kernels and on the interpolation procedure used for the covariates. For step function interpolation and the kernels of models A and B, this integration can be performed analytically (the derivation and final result can be found in Appendix 6.6.3). Therefore, while the resulting final equations for the retarded kernel model are fairly involved, they can be obtained explicitly. As before, this places the retarded kernel approach between the joint model and landmarking methods. Predicted survival probabilities in the joint model involve a numerical procedure and therefore cannot be explicitly written down. On the other hand, the predictive probabilities for a single landmark time in the landmarking model are simple and transparent. However, if prediction is to be made at various landmark times then multiple equations must be specified which increases complexity and reduces the transparency of this approach.

There is scope to further develop the retarded kernel approach. For example, we used a naïve interpolation procedure (step functions) but could try smoother interpolation methods such as Gaussian convolutions [38, 39]. We could also take a Bayesian inference approach, using non-informative prior distributions or incorporating existing knowledge into informative ones. While joint modelling takes into account measurement errors, we have not attempted to do this for the retarded kernel model. Cox models were indeed found to be biased in the presence of such errors [17, 47, 48]. Hence, future work could involve building measurement error effects into the retarded kernel approach. One could also build on methods from WCE models [34] and use a wider class of association kernels, for example those estimated via spline functions.

In this work we compared against standard landmarking models, though extensions to these models exist [2, 11, 49]. Similarly, we only considered joint models with instantaneous dependence on the 'true' covariate trajectory $m_\mu^i(t)$ in the hazard function. This study serves as a 'proof of concept' and as a starting point for future investigations; we leave systematic comparison of alternative model variants to future work. Such comparative research should also make use of more sophisticated evaluation techniques. In our work we used a simple data-splitting technique in order to reduce computational

effort (especially in the joint modelling framework). This serves as a preliminary illustration of model performance but methods such as cross-validation or bootstrapping would be more appropriate in practice [50, 51]. Furthermore, it would also be instructive to look at other measures of model performance. In particular, we used a measure of prediction error that averages over the data set and, as a result, may mask potentially useful information. For example, a model may predict high risk accurately but perform less well for low risk individuals. In practice, this may not be a problem if the aim of the prognosis is to identify high risk patients (e.g. for further assessment or treatment). To investigate these effects one can use calibration plots which show the agreement between observed and predicted risks over the whole spectrum of predicted risks [50, 52]. Calibration plots for a range of different prediction and base times would then provide a detailed analysis of model performance and give insight into the relative performance of the different approaches in various scenarios.

Future work could also benefit from recent developments in simulation methods for dynamic predictions with time-varying covariates [12]. Generating data according to the retarded kernel model with dependence on the period over which covariates are observed is non-trivial, but could possibly be achieved by extending the permutational algorithm developed by Sylvestre and Abrahamowicz (2008) [53]. Such simulations could provide valuable tests for internal consistency.

In summary, we have developed a 'retarded kernel' approach to dynamic prediction that overcomes some limitations of existing methods. By conditioning the hazard rate on observed covariates over a given time frame, it offers a simpler alternative to joint models without disregarding portions of longitudinal covariate data, as is the case with landmarking methods. Using three different clinical data sets we have demonstrated that retarded kernels can have a predictive accuracy comparable to that of established methods. We therefore believe that the retarded kernel method is a promising addition to the toolbox of dynamic prediction methods.

6.6 Appendix A: Mathematical details

6.6.1 Maximum Likelihood Inference

As in standard Cox survival analysis, we use maximum likelihood inference to determine the most plausible values of the model parameters for our models, based on the observed data. We write θ for the full set of parameters, this includes the base hazard rate $h_0(t)$. That is, $\theta = \{h_0(t), a_\mu, \tau_\mu; \mu = 1, \ldots, p\}$ for both models A and B. The optimal parameters are those for which the data likelihood $\mathcal{P}(\mathcal{D}|\theta)$ is maximised. We use the primary event indicator $\delta_i = I(T_i^* \leq C_i) \in \{0, 1\}$, where the indicator function $I(A)$ is defined as $I(A) = 1$ if A holds, and $I(A) = 0$ otherwise. The data likelihood for censored data is then

$$\mathcal{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})^{\delta_i} S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}), \qquad S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) = \mathrm{e}^{-\int_0^{T_i} h(t|\theta, \mathcal{Z}^i_{[0,s_i]})\mathrm{d}t}, \tag{6.41}$$

where, in this section, we refer only to the retarded kernel model and therefore omit the superscript 'RK' from the hazard for clarity. Maximising $\mathcal{P}(\mathcal{D}|\theta)$ is equivalent to minimising the negative log likelihood, i.e. $\hat{\theta}_{\mathrm{ML}} = \mathrm{argmin}_\theta\ \Omega_{\mathrm{ML}}(\theta)$, with

$$\begin{aligned}\Omega_{\mathrm{ML}}(\theta) &= -\ln \prod_{i=1}^{N} h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})^{\delta_i} S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) \\ &= -\sum_{i=1}^{N} \delta_i \ln h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) + \sum_{i=1}^{N} \int_0^{T_i} h(t|\theta, \mathcal{Z}^i_{[0,s_i]})\mathrm{d}t. \end{aligned} \tag{6.42}$$

The hazard rate for the retarded kernel approach is

$$h(t|\theta, \mathcal{Z}^i_{[0,s_i]}) = h_0(t) \exp\left\{ \sum_{\mu=1}^{p} \int_0^{\min(s_i,t)} \beta_\mu(t, t', s_i) z^i_\mu(t')\mathrm{d}t' \right\}. \tag{6.43}$$

Substituting this into $\Omega_{\mathrm{ML}}(\theta)$ yields

$$\begin{aligned}\Omega_{\mathrm{ML}}(\theta) = -\sum_{i=1}^{N} \delta_i \ln h_0(T_i) - \sum_{i=1}^{N} \delta_i \int_0^{s_i} \sum_\mu \beta_\mu(T_i, t', s_i) z^i_\mu(t')\mathrm{d}t' \\ + \sum_{i=1}^{N} \int_0^{T_i} h_0(t') \mathrm{e}^{\int_0^{\min(s_i,t')} \mathrm{d}t'' \sum_\mu \beta_\mu(t',t'',s_i) z^i_\mu(t'')} \mathrm{d}t' \ , \end{aligned} \tag{6.44}$$

where we have used the fact that $s_i \leq T_i$. For simplicity we have specified the hazard in Equation (6.44) without fixed (or baseline) covariates. To include these explicitly (if desired), one can simply add the term $\sum_\nu \gamma_\nu \zeta^i_\nu$ to the exponent of the hazard function.

Extremisation of Equation (6.44) functionally over $h_0(t)$ gives the maximum likelihood estimator of the base hazard rate, given $\beta_\mu(t,t',s)$,

$$\hat{h}_0(t) = \frac{\sum_{i=1}^N \delta_i \delta(t-T_i)}{\sum_{i=1}^N I(t\in[0,T_i])\mathrm{e}^{\int_0^{\min(s_i,t)}\sum_\mu \beta_\mu(t,t',s_i)z_\mu^i(t')\mathrm{d}t'}}, \tag{6.45}$$

as quoted in Equation (6.21) in the main paper. Equation (6.45) is the analogue of the standard Breslow estimator [36]. Inserting this expression back into Equation (6.44) leaves the following function to be extremised over the remaining model parameters $\{a_\mu,\tau_\mu\}$ in the kernels $\beta_\mu(t,t',s)$ (where we denote all terms that do not contain $\{a_\mu,\tau_\mu\}$ simply as 'constant'):

$$\begin{aligned}
\Omega_{\mathrm{ML}}[\{a_\mu,\tau_\mu\}] &= -\sum_{i=1}^N \delta_i \ln\left(\frac{\sum_{k=1}^N \delta_k\delta(T_i-T_k)}{\sum_{j=1}^N I(T_i\in[0,T_j])\mathrm{e}^{\sum_\mu\int_0^{\min(s_j,T_i)}\beta_\mu(T_i,t',s_j)z_\mu^j(t')\mathrm{d}t'}}\right)\\
&\quad -\sum_{i=1}^N\delta_i\sum_\mu\int_0^{s_i}\beta_\mu(T_i,t',s_i)z_\mu^i(t')\mathrm{d}t'\\
&\quad +\sum_{i=1}^N\left\{\int_0^{T_i}\left(\frac{\sum_{k=1}^N\delta_k\delta(t'-T_k)}{\sum_{j=1}^N I(t'\in[0,T_j])\mathrm{e}^{\sum_\mu\int_0^{\min(s_j,t')}\beta_\mu(t',t'',s_j)z_\mu^j(t'')\mathrm{d}t''}}\right)\right.\\
&\quad \left.\times\mathrm{e}^{\sum_\mu\int_0^{\min(s_i,t')}\beta_\mu(t',t'',s_i)z_\mu^i(t'')\mathrm{d}t''}\mathrm{d}t'\right\}\\
&= \sum_{i=1}^N\delta_i\left\{\ln\left(\sum_{j=1}^N I(T_i\in[0,T_j])\mathrm{e}^{\sum_\mu\int_0^{\min(s_j,T_i)}\beta_\mu(T_i,t',s_j)z_\mu^j(t')\mathrm{d}t'}\right)\right.\\
&\quad \left.-\sum_\mu\int_0^{s_i}\beta_\mu(T_i,t',s_i)z_\mu^i(t')\mathrm{d}t'\right\}\\
&\quad +\sum_{k=1}^N\delta_k\left(\frac{\sum_{i=1}^N I(T_k\in[0,T_i])\mathrm{e}^{\sum_\mu\int_0^{\min(s_i,T_k)}\beta_\mu(T_k,t'',s_i)z_\mu^i(t'')\mathrm{d}t''}}{\sum_{j=1}^N I(T_k\in[0,T_j])\mathrm{e}^{\sum_\mu\int_0^{\min(s_j,T_k)}\beta_\mu(T_k,t'',s_j)z_\mu^j(t'')\mathrm{d}t''}}\right)+\text{ constant}\\
&= \sum_{i=1}^N\delta_i\left\{\ln\left(\sum_{j=1}^N I(T_i\in[0,T_j])\mathrm{e}^{\sum_\mu\int_0^{\min(s_j,T_i)}\beta_\mu(T_i,t',s_j)z_\mu^j(t')\mathrm{d}t'}\right)\right.\\
&\quad \left.-\sum_\mu\int_0^{s_i}\beta_\mu(T_i,t',s_i)z_\mu^i(t')\mathrm{d}t'\right\}+\text{ constant.}
\end{aligned} \tag{6.46}$$

This is the formula quoted in Equation (6.22) in the main paper. Minimisation of Equation (6.46) with respect to the remaining model parameters $\{a_\mu,\tau_\mu;\mu=1\ldots p\}$ must be performed numerically. Finally, if we define the N^2 integrals

$$\mathcal{I}_{ij}[\{a_\mu,\tau_\mu\}] = \int_0^{\min(s_j,T_i)}\sum_{\mu=1}^p\beta_\mu(T_i,t',s_j)z_\mu^j(t')\mathrm{d}t', \tag{6.47}$$

then we can re-write expression (6.46) as

$$\Omega_{\mathrm{ML}}[\{a_\mu,\tau_\mu\}] = \sum_{i=1}^{N} \delta_i \ln \left(\sum_{j=1}^{N} I(T_i \in [0,T_j])\ \mathrm{e}^{\mathcal{I}_{ij}[\{a_\mu,\tau_\mu\}]-\mathcal{I}_{ii}[\{a_\mu,\tau_\mu\}]} \right) + \text{constant}. \quad (6.48)$$

6.6.2 Survival probability

We recall from Equation (6.23) in the main paper that the estimated probability that subject i has not experienced an event by time $u > s_i$ conditional on their survival to s_i and on their covariate values $\mathcal{Z}^i$ up to that time is given by

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) = \exp\left\{ -\int_{s_i}^{u} \hat{h}(t'|\mathcal{Z}^i_{[0,s_i]})\mathrm{d}t' \right\}. \quad (6.49)$$

Substituting into this equation the base hazard estimator in Equation (6.45), in combination with Equation (6.43), yields

$$\begin{aligned}
\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) &= \exp\left\{ -\int_{s_i}^{u} \frac{\sum_{j=1}^{N} \delta_j \delta(t'-T_j) \mathrm{e}^{\int_0^{\min(s_i,t')} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(t',t'',s_i) z^i_\mu(t'')}}{\sum_{k=1}^{N} I(t' \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,t')} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(t',t'',s_k) z^k_\mu(t'')}} \mathrm{d}t' \right\} \\
&= \exp\left\{ -\sum_{j=1}^{N} \frac{\delta_j\ I(T_j \in [s_i,u]) \mathrm{e}^{\int_0^{\min(s_i,T_j)} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_i) z^i_\mu(t'')}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,T_j)} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_k) z^k_\mu(t'')}} \right\} \\
&= \exp\left\{ -\sum_{j=1}^{N} \frac{\delta_j\ I(T_j \in [s_i,u]) \mathrm{e}^{\int_0^{s_i} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_i) z^i_\mu(t'')}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,T_j)} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_k) z^k_\mu(t'')}} \right\},
\end{aligned} \quad (6.50)$$

where in the last line we replaced $\min(s_i, T_j) = s_i$, which holds by virtue of the factor $I(T_j \in [s_i, u])$. We have also used the notation $\hat{\beta}_\mu(t, t', s)$ to indicate the association kernel obtained from the ML estimators of the parameters $\{a_\mu, \tau_\mu\}$. Using the integral defined in Equation (6.47) we can re-write Equation (6.50) as

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) = \exp\left\{ -\sum_{j=1}^{N} \delta_j\ I(T_j \in [s_i,u]) \frac{\mathrm{e}^{\mathcal{I}_{ji}[\{\hat{a}_\mu \hat{\tau}_\mu\}]}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\mathcal{I}_{jk}[\{\hat{a}_\mu \hat{\tau}_\mu\}]}} \right\}, \quad (6.51)$$

where we recall that i labels the individual for whom we are making predictions, while the sums over j and k refer to individuals in the data set used for inference.

6.6.3 Step function interpolation

In the main paper we use staircase functions as a straightforward method to interpolate between discrete measurements of the covariates. We take a 'nearest neighbour'

approach, that is we set $z^i_\mu(t) = z^i_\mu(t_{i\ell})$ where $t_{i\ell}$ is the observation time closest to t. The approximated continuous time covariate trajectory then changes value half way between each pair of consecutive observation times. That is,

$$z^i_\mu(t) = \sum_{\ell=1}^{n_i} I(t \in [U_{i\ell}, U_{i\ell+1}]) z^i_\mu(t_{i\ell}) \tag{6.52}$$

where $U_{i\ell}$ denote the 'switch' times with $U_{i1} = 0$ (the first observation time), $U_{in_i+1} = s_i$ (the final observation time), and all other $U_{i\ell}$ occur half way between consecutive observation times, i.e. $U_{i\ell} = \frac{1}{2}(t_{i\ell-1} + t_{i\ell})$ $\ell = 2, \ldots, n_i$. Using Equation (6.52) along with the parameterisations of the association kernels, we can evaluate the integral in Equation (6.47) analytically. We do this in the following sections for retarded kernel models A and B.

6.6.3.1 Model A

We recall from Equation (6.16) in the main paper that the association kernel for model A is defined as

$$\beta_\mu(t, t', s) = \frac{a_\mu}{\tau_\mu} \frac{\exp\left(t'/\tau_\mu\right)}{\exp\left(\min(s,t)/\tau_\mu\right) - 1}. \tag{6.53}$$

Therefore, using the step function defined by $\theta(z > 0) = 1$ and $\theta(z < 0) = 0$, we have

$$\begin{aligned}
\mathcal{I}^{(\mathrm{A})}_{ij}[\{a_\mu, \tau_\mu\}] &= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} \frac{a_\mu z^j_\mu(t_{j\ell})}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \int_0^{\min(s_j,T_i)} \frac{1}{\tau_\mu} \mathrm{e}^{t'/\tau_\mu} I(t' \in [U_{j\ell}, U_{j\ell+1}]) \mathrm{d}t' \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_i} \frac{a_\mu z^j_\mu(t_{j\ell})}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \theta(\min(s_j, T_i, U_{j\ell+1}) - U_{j\ell}) \left[\mathrm{e}^{t'/\tau_\mu}\right]_{U_{j\ell}}^{\min(s_j,T_i,U_{j\ell+1})} \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} a_\mu z^j_\mu(t_{j\ell})\ \theta(\min(s_j, T_i, U_{j\ell+1}) - U_{j\ell}) \frac{\mathrm{e}^{\min(s_j,T_i,U_{j\ell+1})/\tau_\mu} - \mathrm{e}^{U_{j\ell}/\tau_\mu}}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} a_\mu z^j_\mu(t_{j\ell})\ \theta(T_i - U_{j\ell}) \frac{\mathrm{e}^{\min(T_i,U_{j\ell+1})/\tau_\mu} - \mathrm{e}^{U_{j\ell}/\tau_\mu}}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1}
\end{aligned} \tag{6.54}$$

where in the last line we used the fact that $U_{j\ell} < U_{j\ell+1} \leq s_j$.

6.6.3.2 Model B

We recall from Equation (6.17) in the main paper that the association kernel for model B is defined as

$$\beta_\mu(t, t', s) \quad = \quad \frac{a_\mu}{\tau_\mu} \mathrm{e}^{-(t-t')/\tau_\mu} + \frac{a_\mu}{\min(s,t)} \left[1 - \mathrm{e}^{[\min(s,t)-t]/\tau_\mu} + \mathrm{e}^{-t/\tau_\mu}\right]. \tag{6.55}$$

Substituting this into Equation (6.47) gives

$$\begin{aligned}
\mathcal{I}_{ij}^{(\mathrm{B})}[\{a_\mu,\tau_\mu\}] &= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\mathrm{e}^{-T_i/\tau_\mu}\int_0^{\min(s_j,T_i)} I(t'\in[U_{j\ell},U_{j\ell+1}])\frac{1}{\tau_\mu}\mathrm{e}^{t'/\tau_\mu}\mathrm{d}t' \\
&\quad+\sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\left(\int_0^{\min(s_j,T_i)} I(t'\in[U_{j\ell},U_{j\ell+1}])\mathrm{d}t'\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}\right. \\
&\quad\left.+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right) \\
&= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\theta(\min(s_j,T_i,U_{j,\ell+1})-U_{j\ell})\left\{\mathrm{e}^{-T_i/\tau_\mu}\left(\mathrm{e}^{\min(s_j,T_i,U_{j\ell+1})/\tau_\mu}-\mathrm{e}^{U_{j\ell}/\tau_\mu}\right)\right. \\
&\quad\left.+\left(\min(s_j,T_i,U_{j\ell+1})-U_{j\ell}\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right)\right\} \\
&= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\theta(T_i-U_{j\ell})\left\{\mathrm{e}^{-T_i/\tau_\mu}\left(\mathrm{e}^{\min(T_i,U_{j\ell+1})/\tau_\mu}-\mathrm{e}^{U_{j\ell}/\tau_\mu}\right)\right. \\
&\quad\left.+\left(\min(T_i,U_{j\ell+1})-U_{j\ell}\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right)\right\}, \qquad (6.56)
\end{aligned}$$

where in the last line, we have again used the property $U_{j\ell} < U_{j\ell+1} \leq s_j$.

6.7 Appendix B: R code for joint models

All joint models were fitted using the `R` package `JMbayes` [25]. All the data analysed in the main paper are available in this package:

1. PBC data
   (a) `pbc2` contains the PBC data set with time varying measurements of covariates
   (b) `pbc2.id` contains the PBC data set with only baseline covariate measurements per individual
2. AIDS data
   (a) `aids` contains the AIDS data set with time varying measurements of covariates
   (b) `aids.id` contains the AIDS data set with only baseline covariate measurements per individual
3. Liver data

(a) `prothro` contains the Liver data set with time varying measurements of covariates

(b) `prothros` contains the Liver data set with only baseline covariate measurements per individual.

The models fitted below are specified for the full data sets listed above. For the results presented in the main paper, the data sets were split randomly 20 times into training and test data sets and the models were actually fitted to the training data at each iteration.

6.7.1 PBC data

For the results in the main paper we treat the transplant event as a censoring event. To describe this we define a variable `status2` using

```
pbc2.id$status2 <- as.numeric(pbc2.id$status == "dead")
pbc2$status2 <- as.numeric(pbc2$status == "dead")
```

where `status2` = 1 if the individual's event is death and = 0 otherwise.

For the composite event (results shown in Section 6.8) we replace `status2` with `status3`,

```
pbc2.id$status3 <- as.numeric(pbc2.id$status != "alive")
pbc2$status3 <- as.numeric(pbc2$status != "alive")
```

defined as 1 if the individual experiences an event (death or a liver transplant) and 0 otherwise (still alive by end of study).

6.7.1.1 Linear longitudinal model

Extract of `R` code used to fit the PBC data set using the simple linear model described in Section 6.4.2.1 in the main paper. Based on code in Rizopoulos (2012) [9] and Rizopoulos (2018) [54]:

```
long.pbc.linear<-mvglmer(list(log(serBilir)~year+(year|id),
                              log(albumin)~year+(year|id),
                              log(prothrombin)~year+(year|id)),
```

```
            data=pbc2, families=list(gaussian, gaussian, gaussian))
surv.pbc<-coxph(Surv(years, status2)~age, data=pbc2.id, model=TRUE)
JM.pbc.linear<-mvJointModelBayes(long.pbc.linear, surv.pbc,
                                                    timeVar = "year")
```

6.7.1.2 Spline model

Extract of R code to fit the PBC data set using the natural cubic spline model described in Section 6.4.2.1 of the main paper. Based on code in Rizopoulos (2016) [25] and Rizopoulos (2018) [54]:

```
long.pbc.spline<-mvglmer(list(log(serBilir)~ns(year,2,B=c(0,14.4))
                                          +(ns(year,2,B=c(0,14.4))|id),
                              log(albumin)~ns(year,2,B=c(0,14.4))
                                          +(ns(year,2,B=c(0,14.4))|id),
                              log(prothrombin)~ns(year,2,B=c(0,14.4))
                                         +(ns(year,2,B=c(0,14.4))|id)),
               data=pbc2, families=list(gaussian, gaussian, gaussian))
surv.pbc<-coxph(Surv(years, status2)~age, data=pbc2.id, model=TRUE)
JM.pbc.spline<-mvJointModelBayes(long.pbc.spline, surv.spline,
                                                      timeVar = "year")
```

6.7.2 AIDS data

Extract of R code to fit the AIDS data set using the model described in Section 6.4.3.1 of the main paper. Based on code in Section 4.2 of Rizopoulos (2012) [9]:

```
long.aids<-lme(CD4~obstime+obstime:drug, random=~obstime|patient,
                                                          data=aids)
surv.aids<-coxph(Surv(Time, death)~drug+prevOI+AZT+gender,
                                                  data=aids.id, x=TRUE)
JM.aids<-jointModelBayes(long.aids, surv.aids, timeVar="obstime")
```

6.7.3 Liver data

Extract of `R` code to fit the Liver data set using the model described in Section 6.4.4.1 of the main paper. Replicated from code in Section 5.1.2 of Rizopoulos (2012) [9]:

```
prothro$t0<-as.numeric(prothro$time==0)
long.proth<-lme(pro~treat*(ns(time, 3) + t0),
            random=list(id=pdDiag(form=~ns(time,3))), data = prothro)
surv.proth<-coxph(Surv(Time, death)~treat, data=prothros, x=TRUE)
JM.proth<-jointModelBayes(long.proth, surv.proth, timeVar="time")
```

6.8 Appendix C: PBC data with composite event

In the main paper we present the results for the PBC data set for models that treat death as the event of interest and transplant events as censoring events. Here we show the results for models that treat the two events (death or transplant) as a single composite event. Figure 6.11 shows the result for a fixed base time $t = 3$ years and varying prediction time u. Figure 6.12 shows the results for three fixed prediction windows and varying base time t. With comparison to Figures 6.5 and 6.6 in the main paper, we see that the relative accuracy between the models in the two analyses are similar (though overall prediction error for all models is slightly higher for the composite event analysis).

6.9 Appendix D: Edits made to prederrJM

The definition of prediction error $\widehat{\text{PE}}(u|t)$ is given in Equation (6.26) of the main paper. This is identical to the equation for prediction error quoted on pg. 34 in Rizopoulos (2016) [25]. For retarded kernel models A and B prediction error is calculated using a `C++` code that exactly follows this equation.

The `JMbayes` package provides the function `prederrJM` to calculate prediction error for joint models (as described in Rizopoulos (2016) [25]). The function can also be used for standard Cox models and, therefore, landmarking models. However, the source code for `prederrJM` varies very slightly from Equation (6.26). Specifically,

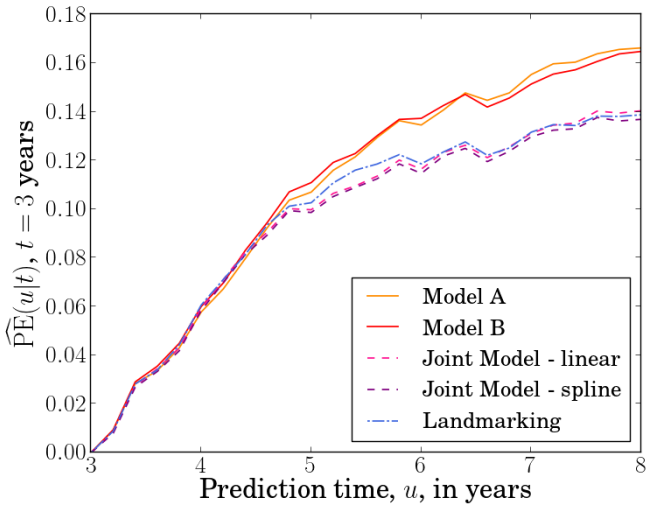


Figure 6.11: Fixed base time results for the PBC data with models fitted treating the two events (death and transplant) as a single composite event. The plot shows overall prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of prediction time u (in years) with fixed base time $t = 3$ years. Prediction error is calculated for u values from 3 to 8 years, with 0.2 year increments. A squared loss function was used in Equation (6.26) in the main paper. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models (one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines). Other than the definition of the composite event, the models fitted are the same as those described in the main paper.

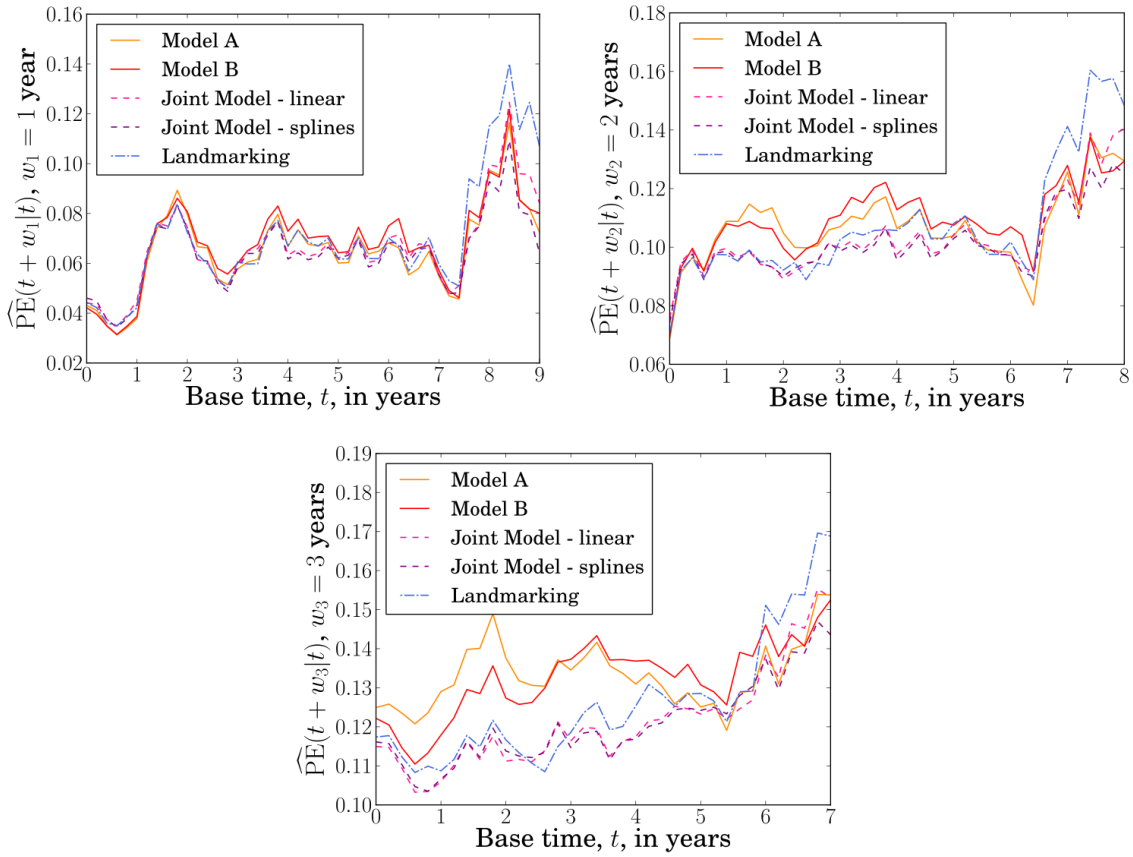


Figure 6.12: Fixed prediction window results for the PBC data with models fitted treating the two events (death and transplant) as a single composite event. Plots show overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in years), with prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The prediction error is calculated for t ranging from 0 to 9,8 or 7 years for w_1, w_2 and w_3 respectively, with 0.2 year increments. A squared loss function was used in Equation (6.26) in the main paper. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. Results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models; one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines. Other than the definition of the composite event, the models fitted are the same as those described in the main paper.

1. `prederrJM` uses $\sum_{i;T_i>t}$ instead of the $\sum_{i;T_i\geq t}$ in Equation (6.26),

2. for the first term (indiviudals who are still alive), `prederrJM` specifies the condition $I(T_i > u)$ instead of $I(T_i \geq u)$,

3. and for the second term (individuals who have experienced the event), `prederrJM` specifies $\delta_i I(T_i \leq u)$ instead of $\delta_i I(T_i < u)$.

These inconsistencies only have an effect when u or t are exactly equal to one (or more) of the event times T_i in the test data. In the PBC and Liver data sets, event times T_i are quoted to a large number of decimal places meaning we never encounter $u = T_i$ or $t = T_i$ (since we vary t and u in steps of 0.2). However, for the AIDS data set, event times are stored to a lower number of decimal places and we do encounter $u = T_i$ or $t = T_i$ for some values of t and u. For the joint model and landmarking results presented in the main paper we use an edited version of `prederrJM` where inequalities exactly match Equation (6.26) (and hence the equation for prediction error in Rizopoulos (2016) [25]). This code can be found at the GitHub repository https://github.com/AnnieDavies/Supplement_Davies_Coolen_Galla_2021. For the PBC and Liver data sets the results in the main paper are the same as those using the `prederrJM` code without these changed inequalities. Figures 6.13 and 6.14 show the results for the AIDS data without these changes. Comparing these to Figures 6.7 and 6.8 in the main paper, it is clear the effect of these changes is very minor.

The handling of exceptions in `prederrJM` is such that the function generates an output `NA` if no-one experiences a (non-censoring) event in the window $[t, u]$. Because we are splitting the data sets randomly into training and test sets at different iterations, we occasionally encounter this scenario for certain windows. For the PBC data this occurred for the fixed base time ($t = 3$ years) analysis at prediction time $u = 3.2$ years for iterations 8, 15 and 16, and for the window $w_1 = 1$ year analysis at base time $t = 7.2$ for iteration 20, $t = 7.4$ for iterations 9 and 20, and $t = 7.6, 7.8$ for iteration 16. For the AIDS data (with code edited to match Equation (6.26)) this occurred only in the fixed base time ($t = 6$ months) analysis at $u = 6.2$ months for iterations 17 and 18 and at $u = 6.4, 6.6$ months for iteration 18. For the Liver data this only occurred for window $w_1 = 1$ year at $t = 8.6, 8.8$ for iteration 16 and $t = 9$ for iteration 5. If there are no non-censoring events in a given window

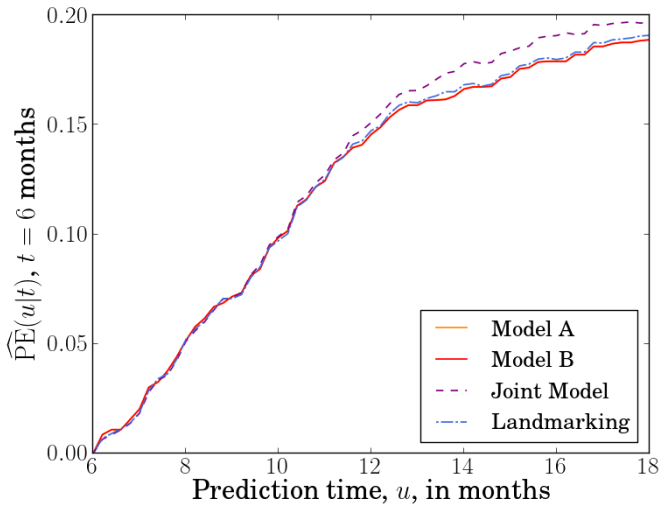


Figure 6.13: Fixed base time results for the AIDS data set using the prederrJM code (for the joint model and landmarking model) without changes made to the inequalities. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in months) for the AIDS data with fixed base time $t = 6$ months. This error is calculated for u ranging from 6 to 18 months, at 0.2 month intervals. In Equation (6.26) in the main paper a squared loss function was used. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. The results from model A (orange line) cannot be seen because they overlap with the results from model B (red line).

$[t, u]$, the second term in Equation (6.26) is equal to zero. Therefore, we edited the `prederrJM` source code to handle this scenario (see the Github repository https://github.com/AnnieDavies/Supplement_Davies_Coolen_Galla_2021). The results in the main paper are for this edited code. Compared to the original `prederrJM` code, these edits have a negligible effect on results.

For the version of `prederrJM` for Cox models, we also obtain an output of `NA` if there is no-one censored in the interval $[t, u]$. In the joint model version of `prederrJM` this is handled by including the argument `na.rm=TRUE` when we perform the sum $\sum_{i;T_i \geq t}$. We therefore added this argument to the function for Cox models.

All changes made to the `prederrJM` source code were very minor and had an almost negligible effect on all results. Changes were made to the code only to ensure that all models were evaluated with exactly the same prediction error equation consistent with the equation quoted in literature [2, 25, 31].

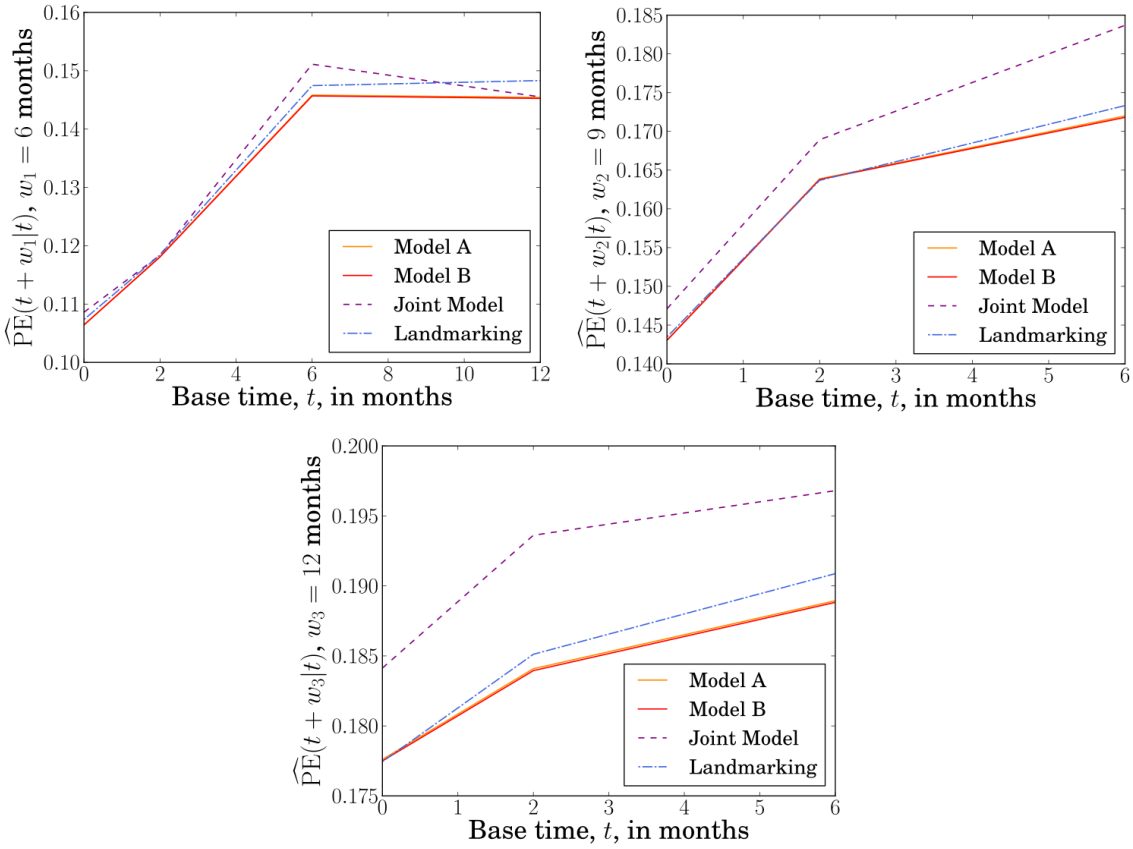


Figure 6.14: Fixed prediction window results for the AIDS data set using the prederrJM code (for the joint model and landmarking model) without changes made to the inequalities. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in months) for the AIDS data with three fixed prediction windows: $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. The prediction times are $u = t + w$. Observations are made at times $0, 2, 6, 12, 18$ months for all individuals in this data set. Prediction errors are hence only updated at these time points. For prediction window w_1, prediction error is measured for $t = 0, 2, 6$ and 12 months. For windows w_2 and w_3, the error is measured at $t = 0, 2$ and 6 months only. In Equation (6.26) in the main paper we used a squared loss function. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. The results from model A (orange line) cannot be seen clearly because they overlap with the results from model B (red line).

6.10 Appendix E: Results with decaying association parameter at s=0

The association kernels $\beta_\mu(t, t', s)$ for models A and B as specified in Equations (6.16) and (6.17) of the main paper do not hold for $s = 0$. In the data sets we analyse, some individuals are observed only once meaning their final observation time is $s = 0$. For the results presented in the main paper we treat the association parameter of these individuals as fixed, $\beta_\mu(t) = a_\mu$. Another option is to define a decaying parameter, $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$. The results for this latter choice are shown in Figures 6.15 and 6.16 for the PBC data (treating transplants as a censoring event), in Figures 6.17 and 6.18 for the AIDS data, and in Figures 6.19 and 6.20 for the Liver data. For the PBC and Liver data, the results with $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$ and $\beta_\mu(t) = a_\mu$ are similar when the base time $t \gtrsim 1$ year. When we restrict the individuals in the test data to having observations over a smaller time frame, the prediction error for these models is much larger. This effect is increased for the larger prediction windows. This can be understood because for smaller t many individuals in the test data will have been observed only once and the parameter $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$ means the effect of this observation is decayed at later times. Similarly in the AIDS data set, the results are similar to the results in the main paper except when the individuals are restricted to only one observation (at $t = 0$). Perhaps another reasonable choice of association parameter for $s = 0$ is a hybrid of the fixed and decaying association, e.g. $\beta_\mu(t) = a_\mu(1 + e^{-t/\tau_\mu})$.

Data Availability Statement

`C++` and `R` codes used to perform the data analysis in this manuscript are available at the GitHub repository https://github.com/AnnieDavies/Supplement_Davies_Coolen_Galla_2021. The data sets analysed are available publicly via the `JMbayes R` package [25].

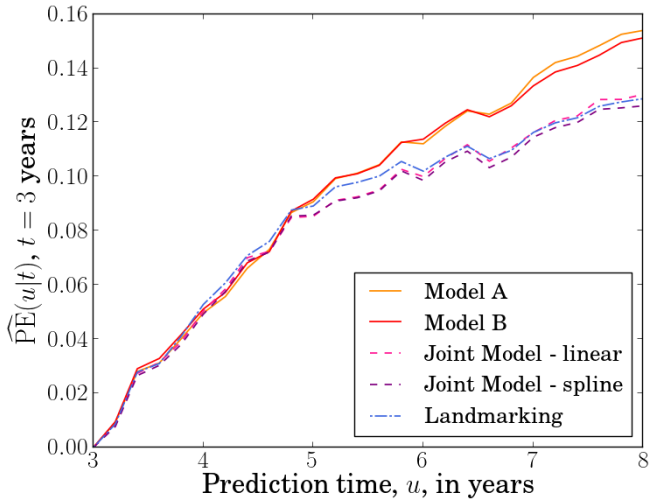


Figure 6.15: Fixed base time results for the PBC data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. The plot shows overall prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of prediction time u (in years) with fixed base time $t = 3$ years. Prediction error is calculated for u values from 3 to 8 years, with 0.2 year increments. A squared loss function was used in Equation (6.26) in the main paper. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models (one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines). Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper (i.e. we treat transplant events as a censoring event).

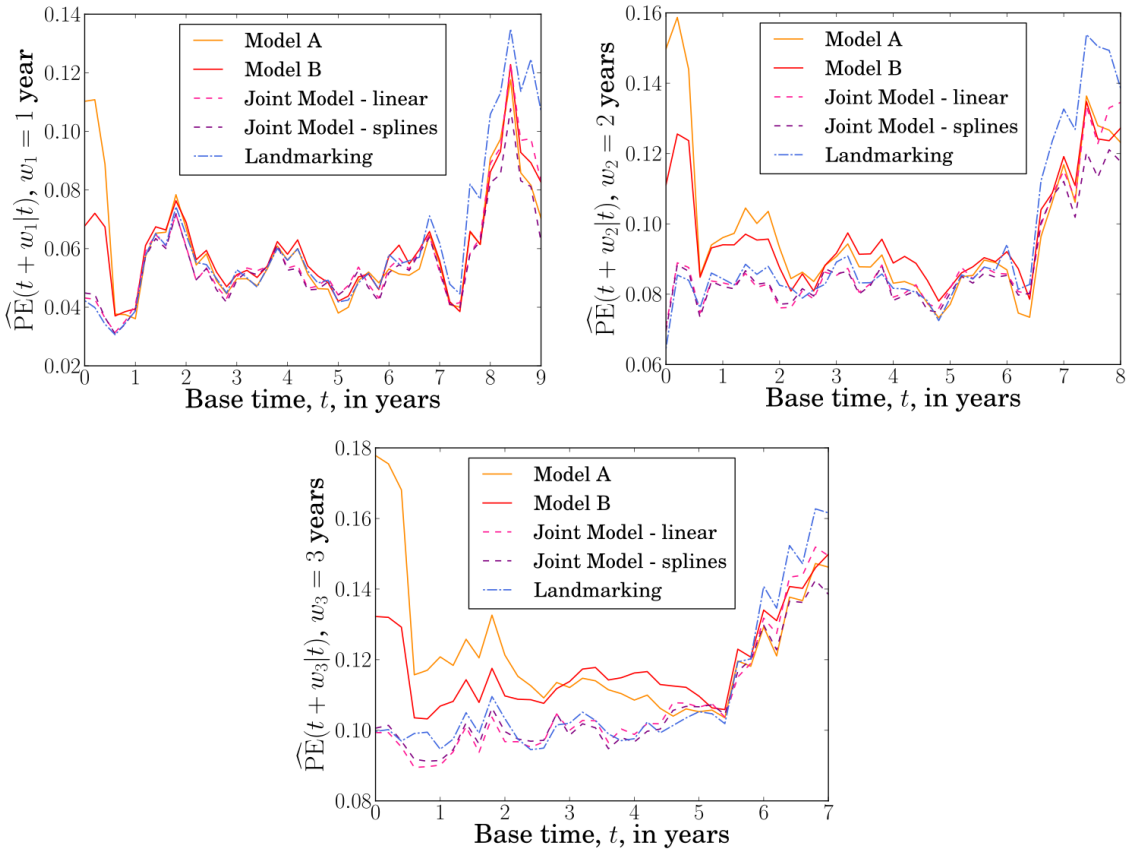


Figure 6.16: Fixed prediction window results for the PBC data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Plots show overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in years), with prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The prediction error is calculated for t ranging from 0 to 9,8 or 7 years for w_1, w_2 and w_3 respectively, with 0.2 year increments. A squared loss function was used in Equation (6.26) in the main paper. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. Results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models; one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper (i.e. we treat transplant events as a censoring event).

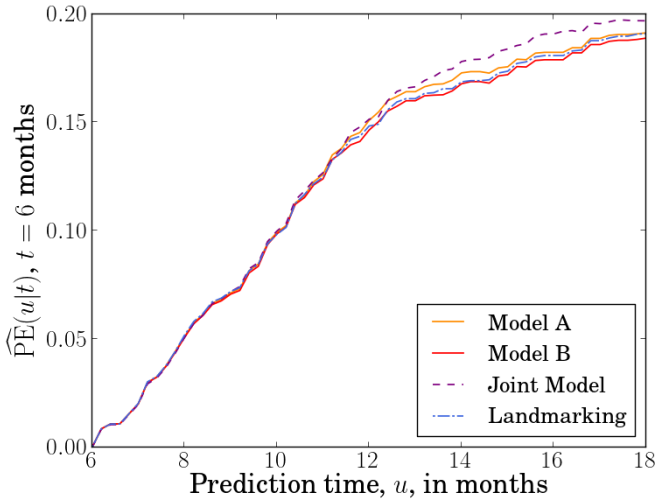


Figure 6.17: Fixed base time results for the AIDS data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in months) with fixed base time $t = 6$ months. This error is calculated for u ranging from 6 to 18 months, at 0.2 month intervals. In Equation (6.26) in the main paper a squared loss function was used. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper. The results from model A (orange line) cannot be seen because they overlap with the results from model B (red line).

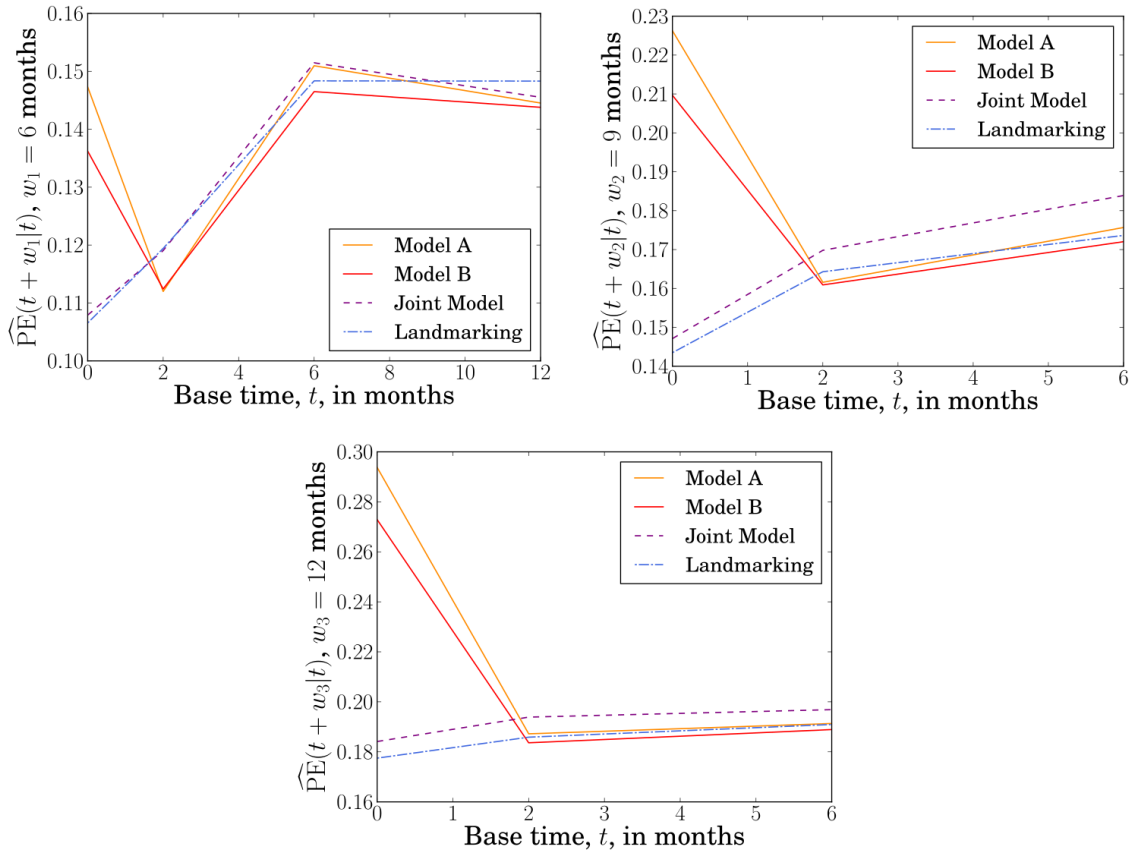


Figure 6.18: Fixed prediction window results for the AIDS data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in months) with three fixed prediction windows: $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. The prediction times are $u = t + w$. Observations are made at times $0, 2, 6, 12, 18$ months for all individuals in this data set. Prediction errors are hence only updated at these time points. For prediction window w_1, prediction error is measured for $t = 0, 2, 6$ and 12 months. For windows w_2 and w_3, the error is measured at $t = 0, 2$ and 6 months only. In Equation (6.26) in the main paper we used a squared loss function. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper. The results from model A (orange line) cannot be seen clearly because they overlap with the results from model B (red line).

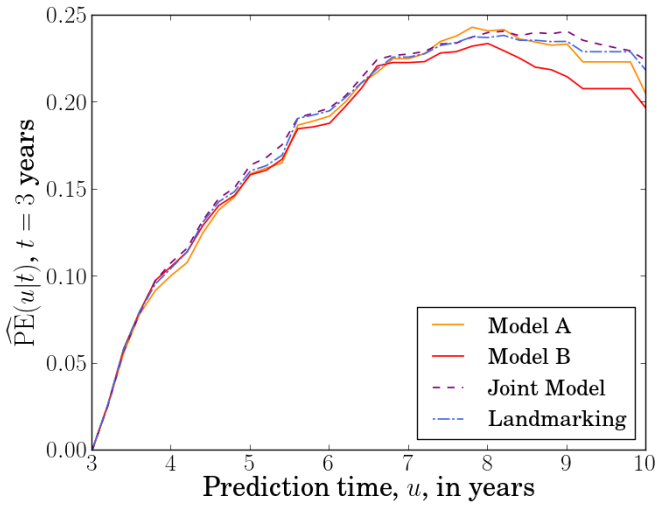


Figure 6.19: Fixed base time results for the Liver data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in years) with fixed base time $t = 3$ years. This error is calculated for u ranging from 3 to 10 years, with 0.2 year increments. In Equation (6.26) in the main paper we used a squared loss function. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper.

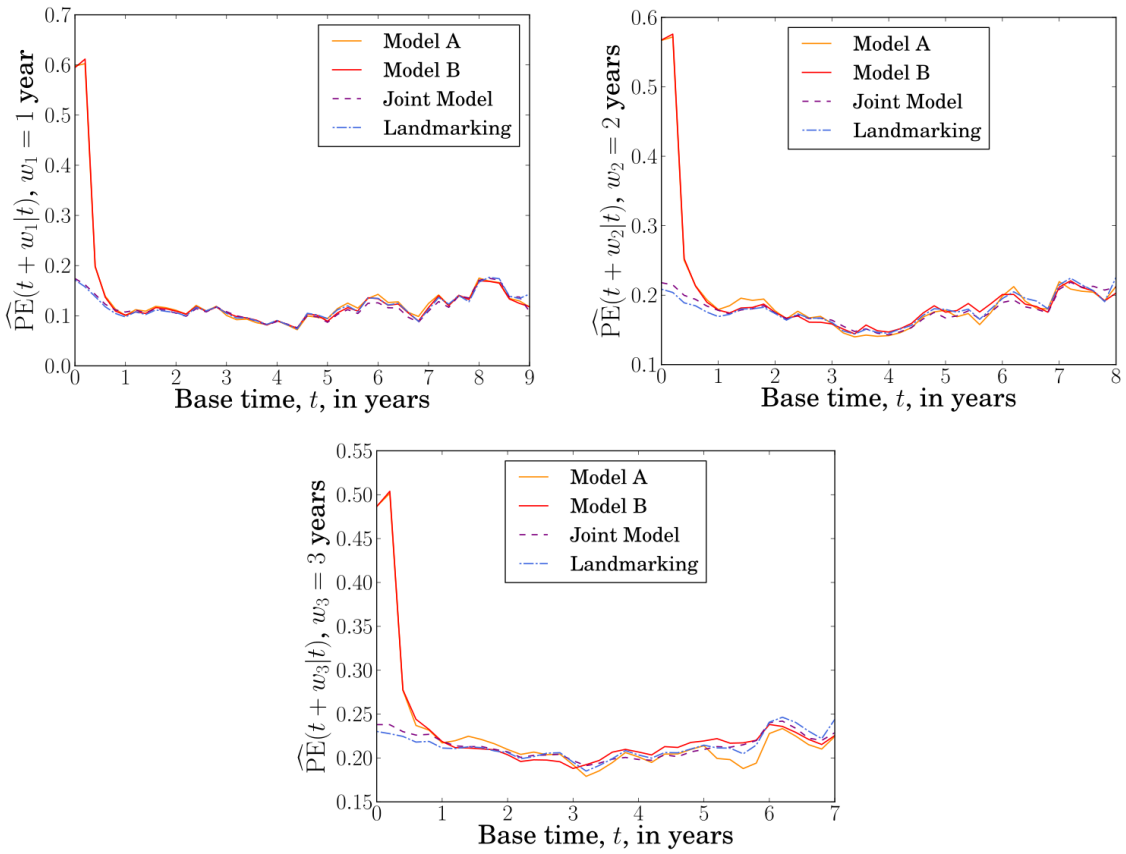


Figure 6.20: Fixed prediction window results for the Liver data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu e^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted against base time t (in years) for the Liver data with three fixed prediction windows, $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The error is calculated for t ranging from 0 to 9,8 or 7 years, for w_1, w_2 and w_3 respectively, with 0.2 year intervals. In Equation (6.26) in the main paper a squared loss function was used. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper.

Bibliography

[1] H. C. Van Houwelingen, "Dynamic prediction by landmarking in event history analysis", Scand. J. Stat. **34**, 70–85 (2007).

[2] D. Rizopoulos, G. Molenberghs, and E. M. E. H. Lesaffre, "Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking", Biometrical J. **59**, 1261–1276 (2017).

[3] D. R. Cox, "Regression models and life-tables", J. Roy. Stat. Soc. B Met. **34**, 187–202 (1972).

[4] L. Tian, D. Zucker, and L. J. Wei, "On the Cox model with time-varying regression coefficients", J. Am. Stat. Assoc. **100**, 172–183 (2005).

[5] S. Lee and H. Lim, "Review of statistical methods for survival analysis using genomic data", Genom. Inform. **17**, e41 (2019).

[6] D. G. Kleinbaum and M. Klein, *Survival analysis. A self-learning text* (Springer-Verlang, New York, NY, USA, 2005).

[7] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data* (Wiley, NJ, USA, 2012).

[8] M. Mills, "The Cox proportional-hazards regression model", in Introducing survival and event history analysis (2012), pp. 86–113.

[9] D. Rizopoulos, *Joint models for longitudinal and time-to-event data with applications in R*, CRC Biostatistics Series (Chapman and Hall, New York, NY, USA, 2012).

[10] D. Rizopoulos, "Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data", Biometrics **67**, 819–829 (2011).

[11] H. Van Houwelingen and H. Putter, *Dynamic prediction in clinical survival analysis* (Taylor & Francis Group, Boca Raton, FL, USA, 2011).

[12] Y. Zhu, L. Li, and X. Huang, "On the landmark survival model for dynamic prediction of event occurrence using longitudinal data", in New frontiers of biostatistics and bioinformatics, edited by Y. Zhao and D.-G. Chen, ICSA Book Series in Statistics (2018), pp. 387–401.

[13] J. Anderson, K. Cain, and R. Gelber, "Analysis of survival by tumor response", J. Clin. Oncol. **1**, 710–719 (1983).

[14] A. A. Tsiatis, A. Degruttola, and M. S. Wulfsohn, "Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS", J. Am. Stat. Assoc. **90**, 27–37 (1995).

[15] J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic concepts and methods for joint models of longitudinal and survival data", J. Clin. Oncol. **28**, 2796–2801 (2010).

[16] L. Wu, W. Liu, G. Y. Yi, and Y. Huang, "Analysis of longitudinal and survival data: joint modeling, inference methods, and issues", J. Probab. Stat. **2012**, 1–17 (2012).

[17] M. W. Arisido, L. Antolini, D. P. Bernasconi, M. G. Valsecchi, and P. Rebora, "Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint", BMC Med. Res. Methodol. **19**, 222 (2019).

[18] M. Moreno-Betancur, J. B. Carlin, S. L. Brilleman, S. K. Tanamas, A. Peeters, and R. Wolfe, "Survival analysis with time-dependent covariates subject to missing data or measurement error: multiple imputation for joint modeling (MIJM)", Biostatistics **19**, 479–496 (2018).

[19] A. L. Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois, "Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group.", Stat. Med. **34**, 2181–2195 (2015).

[20] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues", BMC Med. Res. Methodol. **16**, 117 (2016).

[21] K. Mauff, E. Steyerberg, I. Kardys, E. Boersma, and D. Rizopoulos, "Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach", Stat. Comput. **30**, 999–1014 (2020).

[22] L. Li, S. Luo, B. Hu, and T. Greene, "Dynamic prediction of renal failure using longitudinal biomarkers in a cohort study of chronic kidney disease", Stat. Biosci. **9**, 357–378 (2017).

[23] N. E. Breslow, J. Lubin, P. Marek, and B. Langholz, "Multiplicative models and cohort analysis", J. Am. Stat. Assoc. **78**, 1–12 (1983).

[24] D. C. Thomas, "Models for exposure-time-response relationships with applications to cancer epidemiology", Annu. Rev. Publ. Health. **9**, 451–482 (1988).

[25] D. Rizopoulos, "The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC", J. Stat. Softw. **72**, 1–46 (2016).

[26] P. Murtaugh, E. Dickson, G. Van Dam, M. Malinchoc, P. Grambsch, A. Langworthy, and C. Gips, "Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits", Hepatology **20**, 126–134 (1994).

[27] R. Schoop, E. Graf, and M. Schumacher, "Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates", Biometrics **64**, 603–610 (2008).

[28] D. Abrams, A. Goldman, C. Launer, J. Korvick, J. Neaton, L. Crane, M. Grodesky, S. Wakefield, K. Muth, S. Kornegay, D. Cohn, A. Harris, R. Luskin-Hawk, N. Markowitz, J. Sampson, M. Thompson, and L. Deyton, "A comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy", New. Engl. J. Med. **330**, 657–662 (1994).

[29] X. Guo and B. P. Carlin, "Separate and joint modeling of longitudinal and event time data using standard computer packages", Am. Stat. **58**, 16–24 (2004).

[30] P. Andersen, O. Borgan, R. Gill, and N. Keiding, *Statistical models based on counting processes* (Springer-Verlag, New York, NY, USA, 1993).

[31] R. Henderson, P. Diggle, and A. Dobson, "Identification and efficacy of longitudinal markers for survival", Biostatistics **3**, 33–50 (2002).

[32] S. Cekic, S. Aichele, A. M. Brandmaier, Y. Köhncke, and P. Ghisletta, "A tutorial for joint modeling of longitudinal and time-to-event data in R", Quantitative and Computational Methods in Behavioral Sciences **1**, e2979 (2021).

[33] K. Mauff, E. W. Steyerberg, G. Nijpels, A. A. W. A. Van Der Heijdene, and D. Rizopoulos, "Extension of the association structure in joint models to include weighted cumulative effects", Stat. Med. **36**, 3746–3759 (2017).

[34] M. Abrahamowicz, M. E. Beauchamp, and M. P. Sylvestre, "Comparison of alternative models for linking drug exposure with adverse effects", Stat. Med. **31**, 1014–1030 (2012).

[35] M. P. Sylvestre and M. Abrahamowicz, "Flexible modeling of the cumulative effects of time-dependent exposures on the hazard", Stat. Med. **28**, 3437–3453 (2009).

[36] N. E. Breslow, "Discussion on Professor Cox's paper", J. Roy. Stat. Soc. B Met. **34**, 216–217 (1972).

[37] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives", Comput. J. **7**, 155–162 (1964).

[38] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C. The art of scientific computing*, 2nd ed. (Cambridge University Press, Cambridge, UK, 1992).

[39] R. Van Den Boomgaard and R. Van Der Weij, "Gaussian convolutions. Numerical approximations based on interpolation", in Scale-space and morphology in computer vision, Vol. 2106, edited by M. Kerckhove, Scale-Space 2001. Lecture Notes in Computer Science (2001), pp. 205–214.

[40] R. Schoop, M. Schumacher, and E. Graf, "Measures of prediction error for survival data with longitudinal covariates", Biometrical J. **53**, 275–293 (2011).

[41] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid, "A review of spline function procedures in R", BMC Med. Res. Methodol. **19**, 46 (2019).

[42] D. M. Witten and R. Tibshirani, "Survival analysis with high-dimensional covariates", Stat. Methods. Med. Res. **19**, 29–51 (2010).

[43] J. Huang and D. Harrington, "Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter", Biometrics **58**, 781–791 (2002).

[44] M. Sheikh and A. C. C. Coolen, "Analysis of overfitting in the regularized Cox model", J. Phys. A: Math. Theor. **52**, 384002 (2019).

[45] A. C. C. Coolen, J. E. Barrett, P. Paga, and C. J. Perez-Vicente, "Replica analysis of overfitting in regression models for time-to-event data", J. Phys. A: Math. Theor. **50** (2017).

[46] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement", BMC Med. **13** (2015).

[47] R. L. Prentice, "Covariate measurement errors and parameter estimation in a failure time regression model", Biometrika **69**, 331–342 (1982).

[48] A. A. Tsiatis and M. Davidian, "A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error", Biometrika **88**, 447–458 (2001).

[49] Y. Zhu, X. Huang, and L. Li, "Dynamic prediction of time to a clinical event with sparse and irregularly measured longitudinal biomarkers", Biometrical J. **62**, 1371–1393 (2020).

[50] M. E. Miller, S. L. Hui, and W. M. Tierney, "Validation techniques for logistic regression models.", Stat. Med. **10**, 1213–1226 (1991).

[51] I. Tsamardinos, E. Greasidou, and G. Borboudakis, "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation", Mach. Learn. **107**, 1895–1922 (2018).

[52] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for traditional and novel measures", Epidemiology **21**, 128–138 (2010).

[53] M. P. Sylvestre and M. Abrahamowicz, "Comparison of algorithms to generate event times conditional on time-dependent covariates", Stat. Med. **27**, 2618–2634 (2008).

[54] D. Rizopoulos, *Multivariate joint models vignette*, Online, Available from http://www.drizopoulos.com/vignettes/multivariate%20joint%20models. Accessed July 2021, 2018.

Chapter 7

Conclusions

7.1 Summary of results

We begin by briefly reviewing the findings from Chapters 3, 4 and 6. Following this, we present potential avenues for future research and discuss the wider contributions of the work in this thesis.

Chapter 3: Degree irregularity and rank probability bias in network meta-analysis.

In this chapter, we performed a simulation study to investigate how network topology affects the precision and accuracy of outcomes from a Bayesian NMA. The results from the study can be categorised two-fold into (i) comparisons between treatments within a particular network, and (ii) comparisons between networks with different topologies.

In the first instance, we found that disparity in the number of trials involving different treatments is associated with variation in the precision of treatment effect estimates and that this, in turn, is correlated with a systematic bias in the estimated rank probabilities. In simulations of networks with four treatments, the probability that a treatment ranked best was overestimated for the treatment involved in the fewest trials and underestimated for the treatment involved in the most trials. The same pattern was observed for the probability of being ranked worst. The probabilities of being second and third best were subject to a systematic bias in the opposite direction. These trends in rank probability bias were associated with an increase in standard deviation of effect estimates for treatments included in fewer studies. Based

on this observation, we then offered an explanation for how variation in the precision of treatment effect estimates could generate a bias in rank probabilities.

In order to make comparisons between networks, we defined a measure of 'degree irregularity' that quantifies asymmetry in the way trials are distributed between the treatments. We set the degree of a treatment node to be the number of trials involving that treatment. The irregularity of the network was then the variance in this degree between the nodes. In simulations of networks with different irregularities, we found that more regular networks had more precise treatment effect estimates and smaller bias on rank probabilities. Using this result, we showed that by considering the effect of a new trial on the irregularity of the network, we can propose potential candidates for future trials that will best compliment an existing network of evidence.

Chapter 4: Network meta-analysis and random walks.

Motivated by the well-established analogy between electrical networks and random walks [1] on the one hand, and that of electrical networks and NMA on the other [2], in Chapter 4 we presented a new analogy between network meta-analysis and random walks. The transition matrix describing the random walk was constructed by taking the transition probabilities to be inversely proportional to the variance associated with each edge. Therefore, a walker is more likely to travel across edges corresponding to more precise measurements, and well-connected treatments are visited more often. By analysing the average movement of the random walker, we were able to obtain information about the propagation of evidence through the network. In particular, for a walk that starts at node a and ends at node b, we found that the expected net number of times the walker crosses each edge can be used to construct the evidence flow network for the comparison ab.

We then defined a second transition matrix, this time for a random walker moving on the evidence flow network. For the evidence flow network of comparison ab, the walker starts its journey at node a and stops once it reaches b. This network is directed and acyclic, meaning walkers are restricted to move in a specified direction along each edge, that is, in the direction of evidence flow. The walker can then only take a finite number of routes from a to b, moving along paths of direct and indirect evidence. We interpreted the proportion of walkers flowing through each of these paths as the flow of

evidence through that path. An analytical expression for this quantity was obtained from the product of transition probabilities along the edges that make up that path. This, in turn, led to an analytical expression for the proportion contribution of each direct treatment comparison to each network treatment effect estimate.

In applications to synthetic and real-world data sets, we demonstrated that the random-walk method for constructing proportion contributions offers a number of advantages over the algorithm currently used for this purpose [3]. In some scenarios, the existing algorithm only selects a subset of paths on the evidence flow network meaning paths that potentially contribute a risk of bias are missed. Furthermore, the paths identified by the algorithm in [3] depend on the order in which they are selected. This makes the results of the algorithm ambiguous. The method developed in this chapter overcomes these limitations. In particular, the random-walk approach identifies all paths of evidence and assigns them a value of flow that reflects the properties of the evidence flow network. The resulting proportion contributions are unambiguous and, in addition, the method is able to handle networks with multi-arm trials.

Chapter 6: Retarded kernels for longitudinal survival analysis and dynamic prediction.

In Chapter 6 we developed an approach to the dynamic prediction of patient survival probabilities based on time-varying covariates. We modelled the probability of survival by conditioning hazard rates on the observed covariates measured from baseline up to some subject-specific final observation time. The hazard rates were specified via a time-dependent association kernel that describes the impact of covariate changes at earlier times on the patient's hazard rate at later times. By requiring that our model maintained the well-established features of the Cox model, we derived two kernel parameterisations. In particular, these kernels fulfilled the criteria that our model (i) reduces to the standard Cox model for covariates that are observed to be fixed over time, and (ii) contains the instantaneous Cox model as a special case. We assumed that the impact of a covariate measured in the past decays exponentially over time. In doing so, our models assign more weight to more recent measurements.

In constructing our model, we aimed to overcome some of the limitations of standard dynamic prediction methods. For example, joint modelling is based on

high-dimensional parameterisations of the longitudinal and survival sub-models which makes it conceptually and computationally demanding. For instance, when the time series of covariate measurements from different patients exhibit varied and complicated characteristics (e.g they are non-linear as a function of time), correctly modelling the longitudinal covariate trajectories is challenging and can be prone to misspecification. This then increases the risk of introducing bias. Furthermore, the number of model parameters rapidly increases with the inclusion of covariates meaning fitting the models quickly becomes computationally intensive. The landmarking approach, on the other hand, is much simpler but this comes with its own drawbacks. In particular, landmarking does not make efficient use of the available data. Predictions made at a certain 'landmark time' use only the at risk data set, meaning data from patients who have experienced an event before this time is discarded. In addition, standard landmarking approaches use only the most recent covariate measurements rather than the full history of the longitudinal trajectories. The retarded kernel approach developed in this chapter therefore sits somewhere in between the two standard approaches in terms of complexity. The model accounts for the full history of longitudinal covariates, but is more parsimonious than joint modelling.

To assess the performance of our models, we applied the retarded kernel approach to three clinical data sets. Using an established measure of prediction error, we compared our results with those obtained from the two standard approaches. In the different scenarios we tested, no one model was found to be consistently superior or inferior. Therefore, the retarded kernel approach exhibited similar predictive accuracy as the more established approaches.

7.2 Avenues for future work

7.2.1 Network meta-analysis

Chapters 3 and 4 relate to the broader idea that studying the properties of the NMA network leads to a better understanding of the mechanics of the network meta-analysis process. The work in these chapters prompts further questions on this theme.

Network characteristics. In Chapter 3 we made use of a topological index (degree irregularity, h^2) that describes the network structure in terms of the number

of trials involving each treatment. We controlled for factors such as the number of treatments, the number of participants in each trial arm, and the relative effectiveness of the treatments. In doing so, we were able to isolate the characteristic we were interested in and understand the mechanism by which it affects the analysis. The reality of networks of treatments and trials is, of course, much more complicated than the scenarios we simulated. It would therefore be interesting to investigate the impact of other network characteristics. In particular, our finding that increasing the regularity of a network improves the precision and accuracy of outcomes does not tell the whole story.

For example, for a network of four treatments one could construct a loop structure (Figure 3.2 (b) in Chapter 3) where the two horizontal edges represent a strong connection (multiple trials) but the vertical connections are weak (they correspond to a single trial). If we require that parallel edges represent the same number of trials (e.g. the two horizontal edges represent 10 trials and the two vertical connections represent one trial) then, because of the symmetry of the network structure, each treatment would be involved in an equal number of trials. In this scenario, as long as we continued to add the same number of trials to the parallel connections, the irregularity of the network would remain zero (see Figure 7.1). However, intuitively one would imagine that it would be more beneficial to add trials to the weaker vertical connections than to the already strong horizontal ones. This example illustrates that irregularity is likely not the only topological feature that affects NMA outcomes. In future research, one could investigate characteristics such as network connectivity and potentially construct a more general topological index that accounts for a number of different features.

Another network feature that we have so-far neglected in our investigations is the number of participants. For example, there may be interest in investigating the optimal distribution of participants between the trials. In other words, is it more beneficial (for the accuracy and precision of NMA outcomes) to have many small trials, or fewer trials with a larger number of participants? The number of participants in a trial is related to the sampling variability associated with that trial. Therefore, we could rephrase the question more generally: to obtain accurate and precise estimates of the mean and variance of a distribution (for example), is it preferable to have many noisy measurements or a small number of precise measurements? The question could also

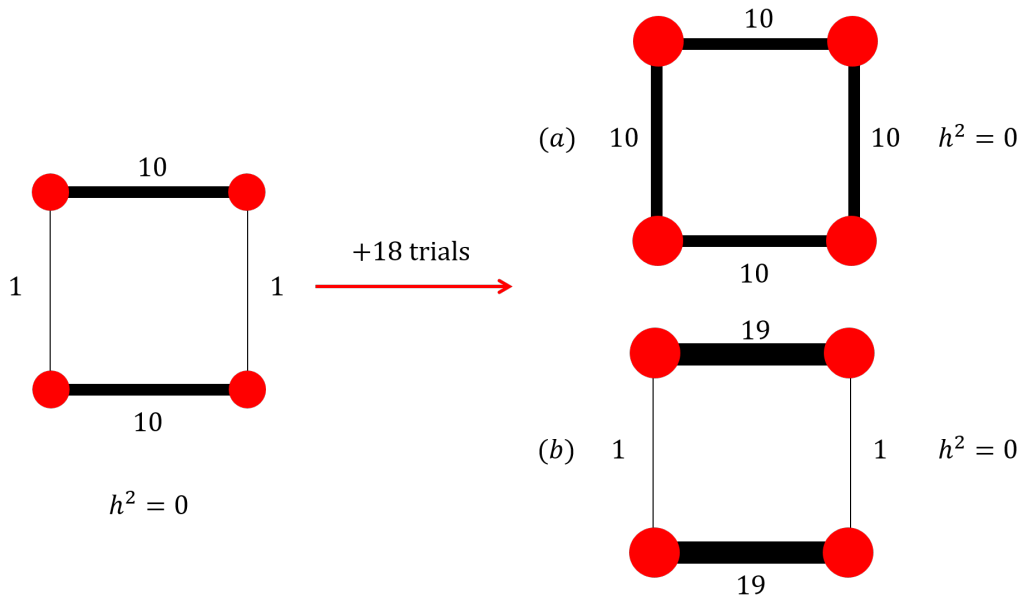


Figure 7.1: An example of a regular network ($h^2 = 0$) constructed from a loop structure with equal parallel edges. Edges are labelled with the number of trials they represent. Networks (a) and (b) show two possible ways of distributing an additional 18 trials. Both versions maintain the regularity of the network ($h^2 = 0$) but network (a) also has equality in edge thickness (for the existing edges). It is hypothesised that option (a) would produce more accurate and precise outcomes than option (b) even though both networks have the same irregularity. This example was not explored in Chapter 3 and illustrates that the irregularity metric introduced in this chapter does not tell the whole story about network topology.

be phrased subject to constraints, for example, given a maximum number of total participants, what is the best way to distribute them between trials? Consideration of the cost of conducting multiple trials could then also be taken into account.

The work in Chapter 3 identified a mechanism by which a systematic bias enters into the network. One then asks whether this bias can be adjusted for in the methodology. Due to the complicated nature of real meta-analytic networks, this is likely to be a difficult task but, nonetheless, could lead to a valuable result.

NMA and random walks. Random walks are used in a wide range of applications to study systems that can be described by networks. The random walk analogy presented in Chapter 4 therefore provides the catalyst for NMA to exploit the plethora of existing research on this topic. For example, outside of NMA literature, random walks have been used for ranking nodes in a network [4, 5]. By defining transition probabilities that reflect certain properties of the treatments and trials, one could make use of these methods to rank treatment options. Indeed, such a method has previously been developed [6]. In this application the transition probabilities were specified in terms of

the probability that one treatment is better than another. Extensions to this approach could include applying additional constraints to the walker so that the treatment ranking reflects a range of desirable properties.

In Chapter 4, we defined transition probabilities in terms of the variance associated with different comparisons. We could have instead used other parameterisations such as the treatment effect estimates themselves. Other features of the random walk are also worthy of investigation. For example, the expected number of steps taken to get from one node to another, or variation in the paths traversed by the walkers. One could also remove the requirement that the random walker makes a transition at every step. There may then be interest in the time the walker spends at each node.

Finally, as discussed in more detail in Section 2.7 of Chapter 2, other techniques from statistical mechanics have the potential to be useful in network meta-analysis. For example, it might be possible to frame certain questions in NMA as constrained optimisation or constraint satisfaction problems and to make use of related methodologies such as message passing and the cavity method. In future work, it would be interesting to explore these applications.

7.2.2 Dynamic prediction

Chapter 6 can be viewed as a 'proof of concept' for the retarded kernel approach to dynamic prediction. There is, therefore, more work to be done in evaluating and extending the method. For example, future research could involve performing simulation studies to check the internal consistency of the model. This would require the development of appropriate data generation techniques. The fact that our model conditions on the time period of observation means that generating data that is consistent with this model is not straightforward. In particular, any appropriate generation method must account for the mutual dependencies between event times, final observation times and time-varying covariates. For example, specifying the final observation time of a patient necessarily requires that they survive to at least that time. In future work it would also be useful to create an R package to implement the retarded kernel method for a given data set. This would make it accessible for others to use, evaluate and further develop the approach.

'First hitting time' (FHT) or 'first passage time' (FPT) models describe the time to an event in terms of the time taken for a stochastic process to reach some threshold value. The models have a range of applications in fields such as physics, biology, engineering and economics [7–9]. In statistical physics, the most notable is perhaps the description of Brownian motion, the random movement of a particle in a liquid or gas [8]. As one may predict, FHT models have been used in survival analysis [7, 10]. Here, the underlying stochastic process represents the health of the individual. This process can be parameterised in terms of the covariates leading to a so-called 'threshold regression' model. In Chapter 6 we focussed on extending Cox's proportional hazards (PH) model so that our approach was a natural addition for analysts in the field. In fact, a (parametric) PH model can be obtained as a special case of threshold regression models [11]. The FHT approach therefore offers an appealing connection between statistical physics and survival analysis.

Future work could investigate how FHT models compare to the retarded kernel approach, and how these ideas could be incorporated. For example, it would be interesting to see if one could justify the form of association kernels from mechanistic models of the disease at a lower level using first hitting time ideas. FHT models also provide a natural way of simulating time-to-event data. There may then be interest in investigating how different dynamic prediction models perform on this data.

7.3 Central themes, comments and concluding remarks

In this thesis we have explored applications of statistical physics in medical statistics. We have demonstrated that ideas from statistical mechanics, and a physicist's approach to understanding physical systems have relevance to medical statistics problems. In doing this we have contributed to ongoing discussions in the field with the aim of furthering the understanding and development of medical statistics methodology.

7.3.1 Interdisciplinary application of statistical mechanics to medical statistics

In Chapter 2 we provided an introduction to network meta-analysis that aimed to serve as a starting point for researchers with a background in physics to enter the field. While the work in this chapter does not contain any original results, we believe that it represents the first concerted effort to bring together the two disciplines. To the best of our knowledge, the interdisciplinary connections and analogies between physics and NMA have not previously been compiled into a single discussion. We also present new ideas for how statistical mechanics may contribute to NMA in the future. We hope that by writing such an article, we will encourage and facilitate other physicists to enter this field.

Chapters 3, 4 and 6 all serve as examples of how medical statistics can benefit from a statistical physics perspective. While the interdisciplinary application in Chapter 4 is evident, the influence of statistical mechanics in Chapters 3 and 6 is more subtle, and is explained below.

In this thesis, we took an exploratory approach to research and looked at a range of different problems in medical statistics. The selection of topics was naturally influenced by our background in statistical physics. In particular, we were drawn to problems that had connections to familiar concepts such as network structures and stochastic processes. Based on our experience in analysing physical systems, we asked questions that were most pertinent to our own understanding of the topics and tried to tackle problems using techniques from our statistical mechanics toolkit.

In this way, the choice to investigate the effect of network topology in Chapter 3, and to explore models of time evolutionary stochastic processes in Chapter 6 was influenced by our interest in, and familiarity with, similar topics encountered in statistical physics. Furthermore, in carrying out these projects we benefited from skills acquired from a background in statistical mechanics. Examples include familiarity with simulation and optimisation techniques, understanding of Bayesian models and Markov chain Monte Carlo methods, and experience in modelling probability.

The applicability of statistical mechanics to medical statistics is, however, best demonstrated in Chapter 4. Random walks are a central topic in statistical mechanics,

most notably used for modelling the diffusion of particles in liquids and gases [12]. In Chapter 4 we were able to demonstrate an analogy between random walks and network meta-analysis. In doing so we contributed to a better understanding of the flow of evidence and developed a more reliable method of constructing the proportion contribution matrix.

7.3.2 Contributions to medical statistics methodology

The motivation for research into medical statistics is a pragmatic one. In order for clinicians to provide the best quality of care for their patients they require the highest quality information. Methods of analysis must, therefore, provide the best possible representation of the data, be practical to implement and produce meaningful results that inform medical decision making.

In our interactions and collaborations with statisticians, we have learned to think about the impact of our research in relation to clinical practice. This has motivated us to investigate practical applications of our work and has ultimately resulted in projects that we believe have contributed to the understanding and development of methods in medical statistics.

Our simulation study in Chapter 3 demonstrates how biases in ranking metrics might originate from the structure of the network of treatment and trials. This contributes to a better understanding of the limitations of NMA methodology and provides guidance on the use of ranking statistics in practice. Furthermore, in Chapter 4, by studying the properties of a random walk on the network, we gain insight into how evidence flows in NMA. These projects serve as examples of how our work has helped to improve understanding of medical statistics methods.

Other work in this thesis has contributed methodology that builds and improves upon existing methods. For example, in Chapter 6 we developed an approach to dynamic prediction by extending the standard Cox model. This methodology offers advantages over the two standard methods without compromising predictive accuracy. As discussed in Chapter 4, the random-walk approach to constructing evidence streams is more reliable than existing algorithms. The methods we have developed based on this approach are now implemented in the widely used software `netmeta` [13]. As a result, other software tools such as CINeMA [14] and ROB-MEN [15] can now be made

more reliable. This will be useful in clinical practice for the evaluation of confidence and risk-of-bias in the treatment effect estimates from an NMA.

7.3.3 General outlook

In recent years, there has been a rapidly growing interest in using statistical mechanics as a framework to study phenomena outside the realm of traditional physics [16–18]. These applications have led to fruitful developments in a diverse range of interdisciplinary fields including biology, economics, and sociology. I believe that there is still a way to go before the same can be said of medical statistics. My hope is that the work in this thesis has highlighted the potential for medical statistics to benefit from an interdisciplinary application of statistical physics, and that it has contributed some initial steps towards achieving this aim.

Bibliography

[1] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, arXiv:math/0001057, 2000.

[2] G. Rücker, "Network meta-analysis, electrical networks and graph theory", Res. Synth. Meth. **3**, 312–324 (2012).

[3] T. Papakonstantinou, A. Nikolakopoulou, G. Rücker, A. Chaimani, G. Schwarzer, M. Egger, and G. Salanti, "Estimating the contribution of studies in network meta-analysis: paths, flows and streams", F1000Res. **7**, 610 (2018).

[4] C. Daniłowicz and J. Baliński, "Document ranking based upon Markov chains", Inform. Process. Manag. **37**, 623–637 (2001).

[5] J. Blanchet, G. Gallego, and V. Goyal, "A Markov chain approximation to choice modeling", Oper. Res. **64**, 886–905 (2016).

[6] A. Chaimani, R. Porcher, É. Sbidian, and D. Mavridis, "A Markov chain approach for ranking treatments in network meta-analysis", Stat. Med. **40**, 451–464 (2021).

[7] M.-L. T. Lee and G. A. Whitmore, "Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary", Stat. Sci. **21**, 501–513 (2006).

[8] G. Klein, "Mean first-passage times of Brownian motion and related problems", Proc. R. Soc. A **211**, 431–443 (1952).

[9] S. Redner, *A guide to first-passage processes* (Cambridge University Press, Cambridge, UK, 2001).

[10] D. Stogiannis, C. Caroni, C. E. Anagnostopoulos, and I. K. Toumpoulis, "Comparing first hitting time and proportional hazards regression models", J. Appl. Stat. **38**, 1483–1492 (2011).

[11] M.-L. T. Lee and G. A. Whitmore, "Proportional hazards and threshold regression: their theoretical and practical connections", Lifetime Data Anal. **16**, 196–214 (2010).

[12] N. Davidson, *Statistical mechanics* (Dover Publications Inc, Mineola, NY, USA, 2003).

[13] G. Rücker, U. Krahn, J. König, O. Efthimiou, A. Davies, T. Papakonstantinou, and G. Schwarzer, *Netmeta: network meta-analysis using frequentist methods*, R package version 2.0-0. https://CRAN.R-project.org/package=netmeta, R Foundation for Statistical Computing (Vienna, Austria, 2021).

[14] A. Nikolakopoulou, J. P. T. Higgins, T. Papakonstantinou, A. Chaimani, C. Del Giovane, M. Egger, and G. Salanti, "CINeMA: an approach for assessing confidence in the results of a network meta-analysis", PLOS Med. **17**, e1003082 (2020).

[15] V. Chiocchia, A. Nikolakopoulou, J. P. T. Higgins, M. J. Page, T. Papakonstantinou, A. Cipriani, T. A. Furukawa, G. C. M. Siontis, M. Egger, and G. Salanti, "ROB-MEN: a tool to assess risk of bias due to missing evidence in network meta-analysis", BMC Med. **19**, 304 (2021).

[16] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics", Rev. Mod. Phys. **81**, 591–646 (2009).

[17] G. Parisi, "Statistical physics and biology", Phys. World. **6**, 42–47 (1993).

[18] D. Stauffer, "Introduction to statistical physics outside physics", Physica A **336**, Proceedings of the XVIII Max Born Symposium "Statistical Physics Outside Physics", 1–5 (2004).

Chapter 8

Supplementary material for 'Degree irregularity and rank probability bias in network meta-analysis'

Preface

This chapter contains supplementary simulations and figures for the paper 'Degree irregularity and rank probability bias in network meta-analysis' published by Research Synthesis Methods[1] and presented in Chapter 3.

These sections correspond to Sections S3-S11 in the published online Supplementary Material. We have separated this from the main body of the thesis because it contains a large number of figures that would otherwise interrupt the flow of the text.

[1] A. L. Davies and T. Galla, "Degree irregularity and rank probability bias in network meta-analysis", Res. Synth. Meth. **12**, 316-332 (2021). 10.1002/jrsm.1454

8.1 Within network plots: The effect of the number of studies per treatment for equally effective treatments

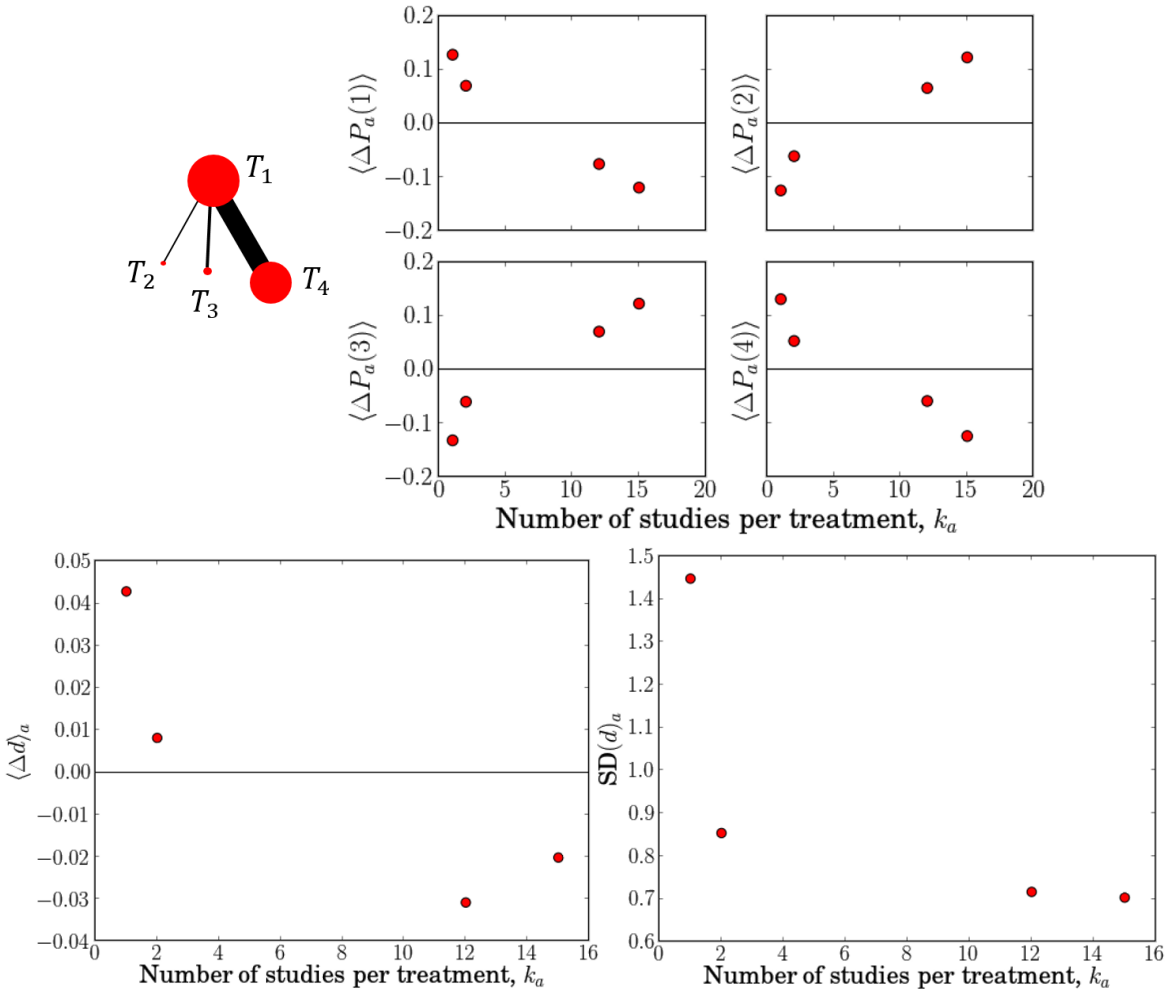


Figure 8.1: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a star network with $\boldsymbol{K} = (1, 2, 12, 0, 0, 0)$.

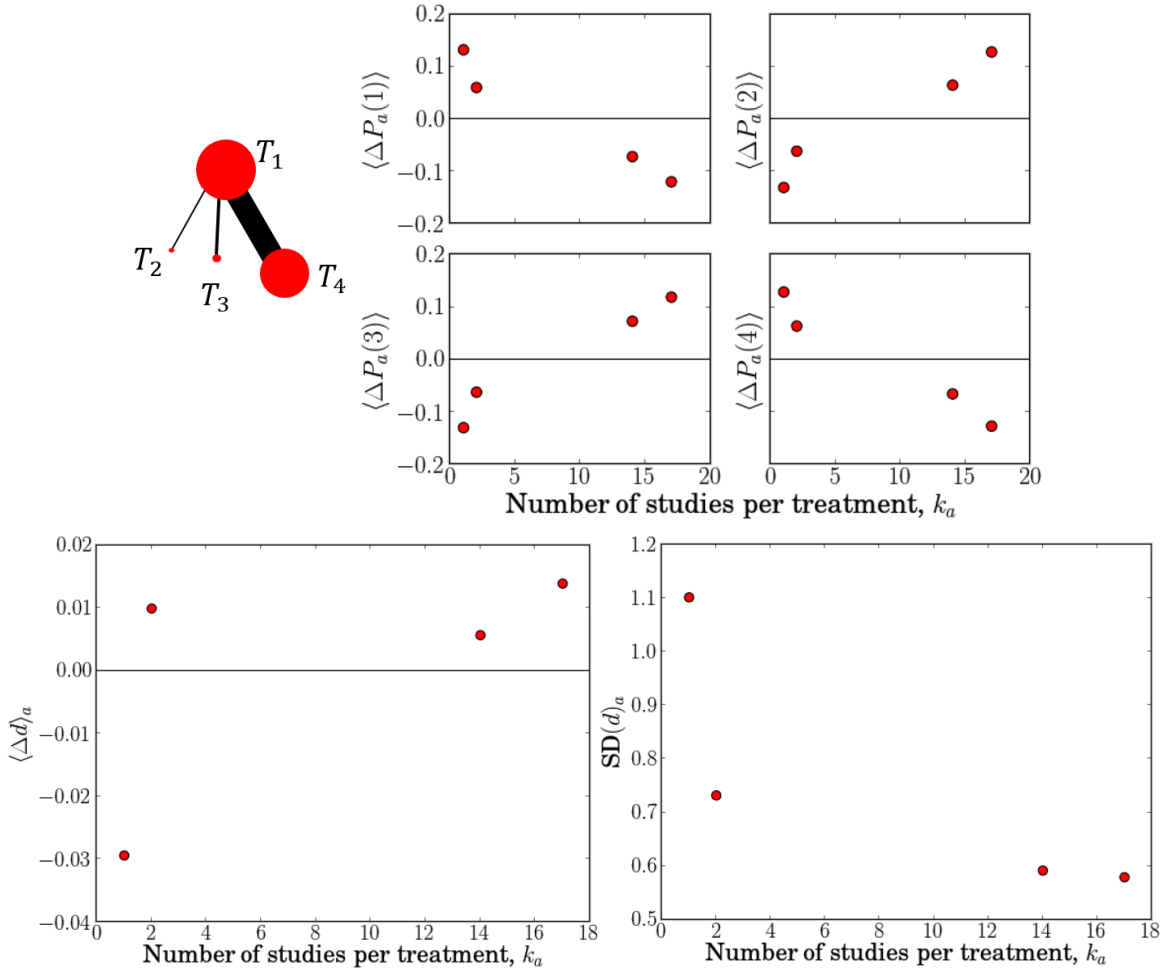


Figure 8.2: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a star network with $\boldsymbol{K} = (1, 2, 14, 0, 0, 0)$.

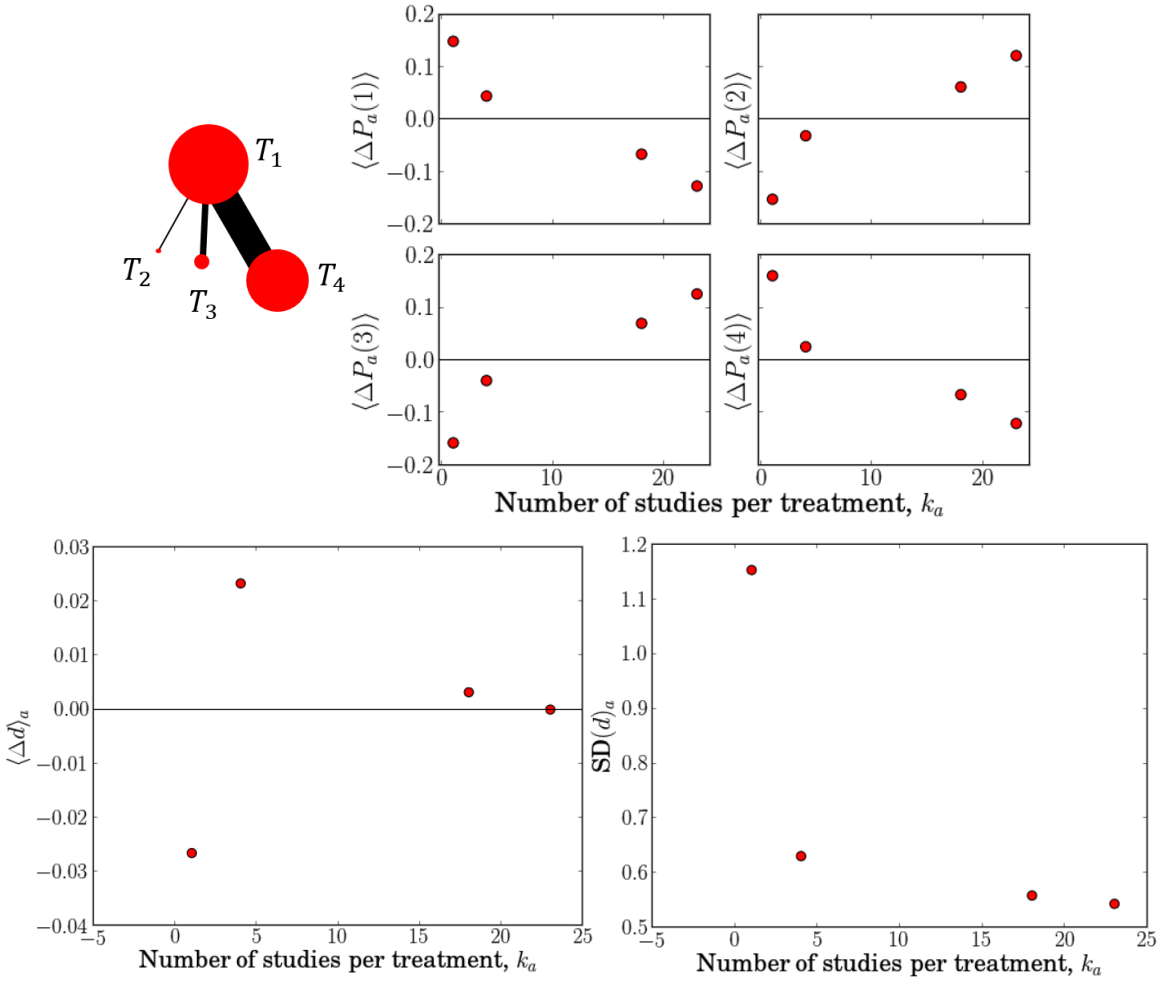


Figure 8.3: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a star network with $\boldsymbol{K} = (1, 4, 18, 0, 0, 0)$.

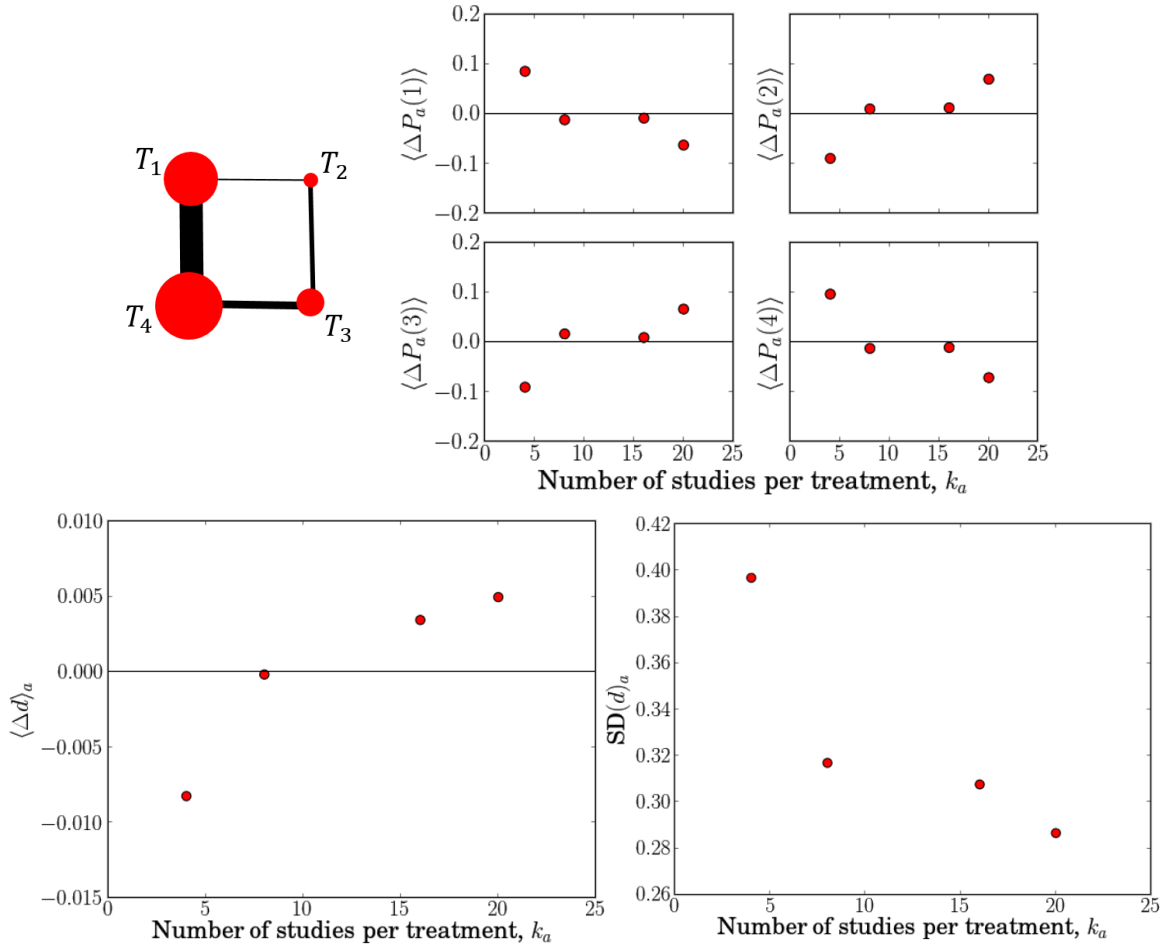


Figure 8.4: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a loop network with $\boldsymbol{K} = (1, 0, 15, 3, 0, 5)$.

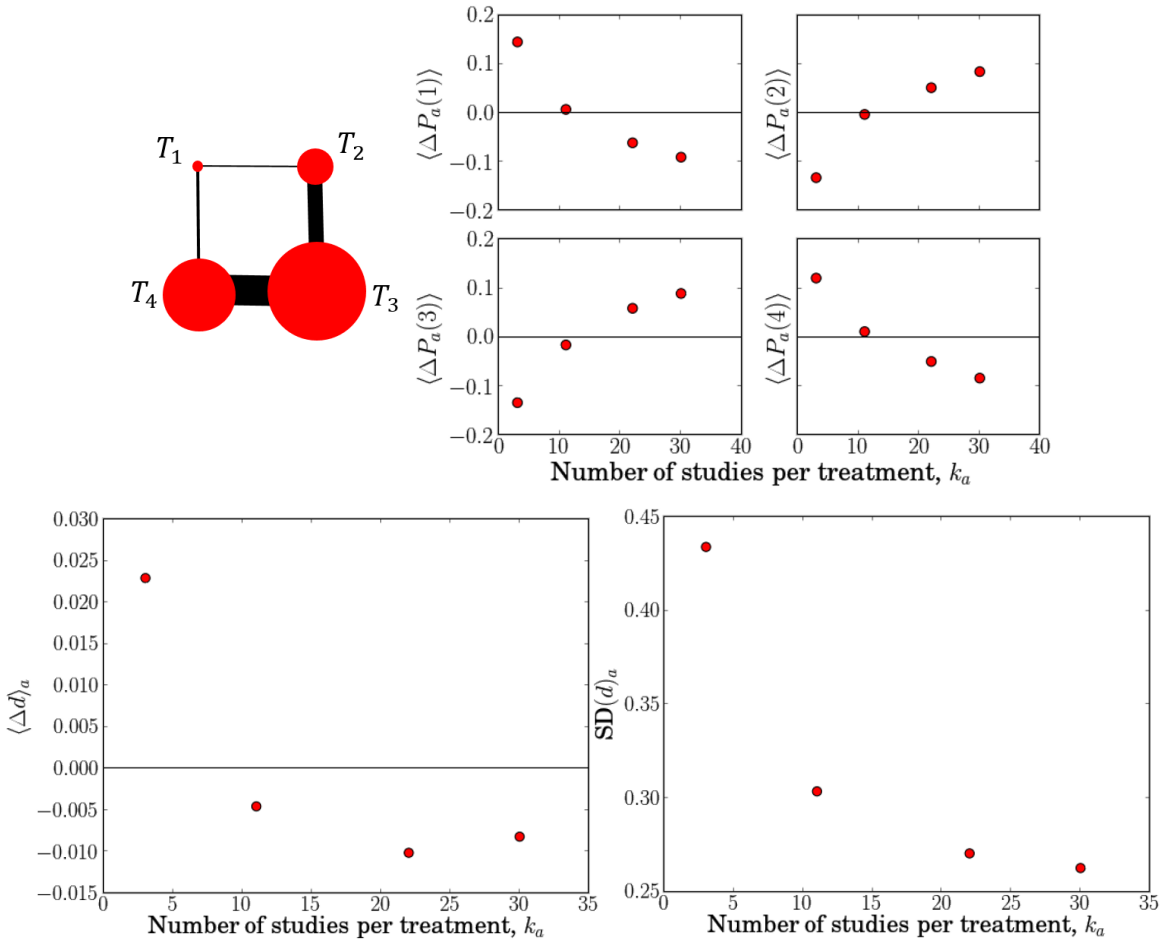


Figure 8.5: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a loop network with $\boldsymbol{K} = (1, 0, 2, 10, 0, 20)$.

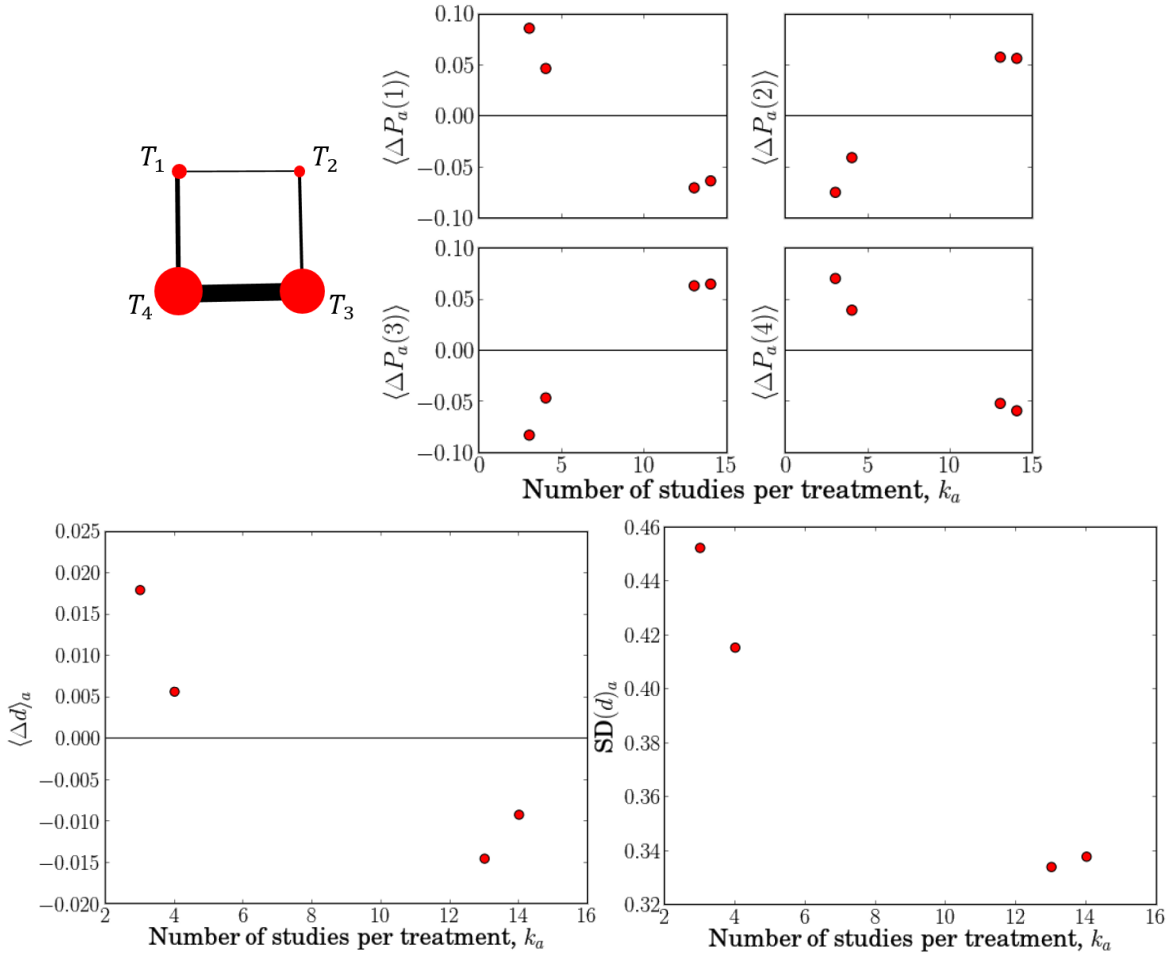


Figure 8.6: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a loop network with $\boldsymbol{K} = (1, 0, 3, 2, 0, 11)$.

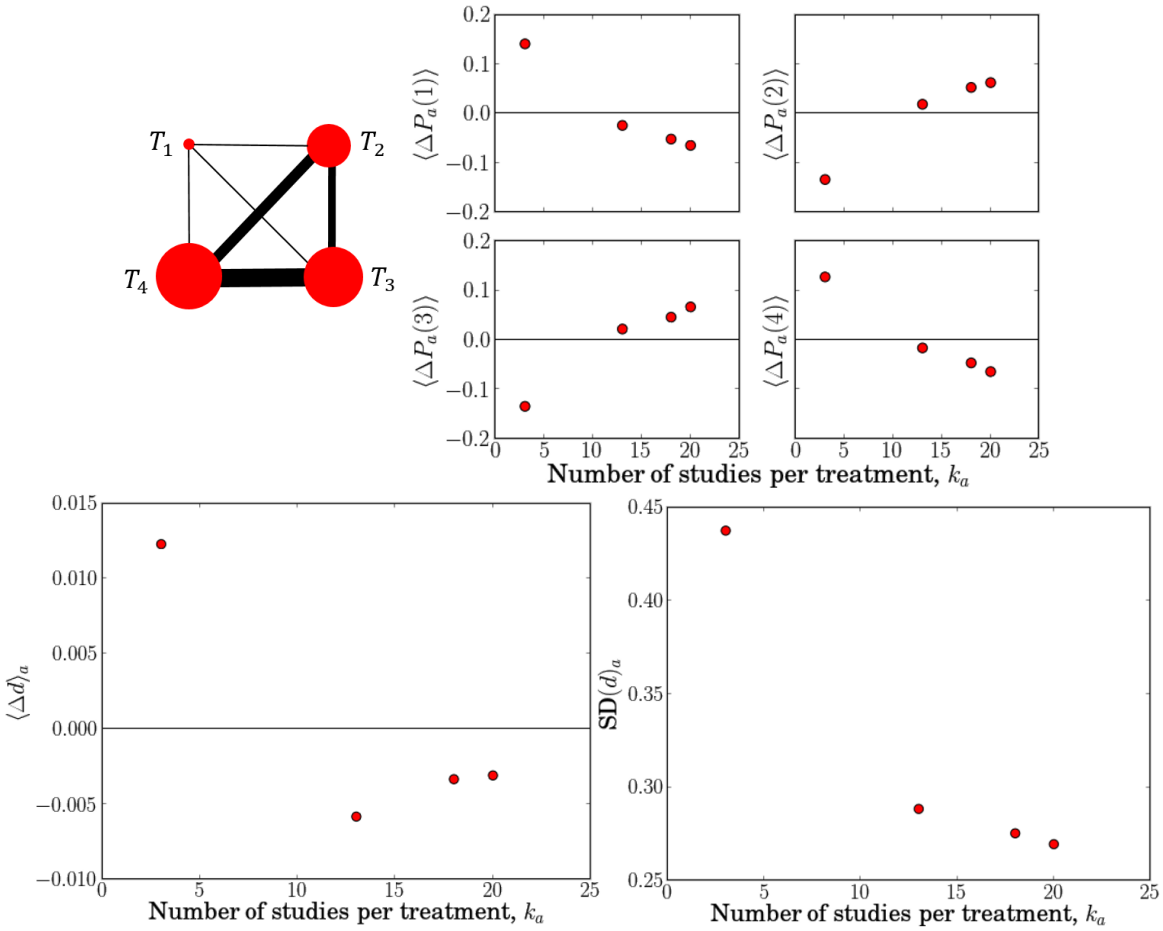


Figure 8.7: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (1, 1, 1, 5, 7, 12)$.

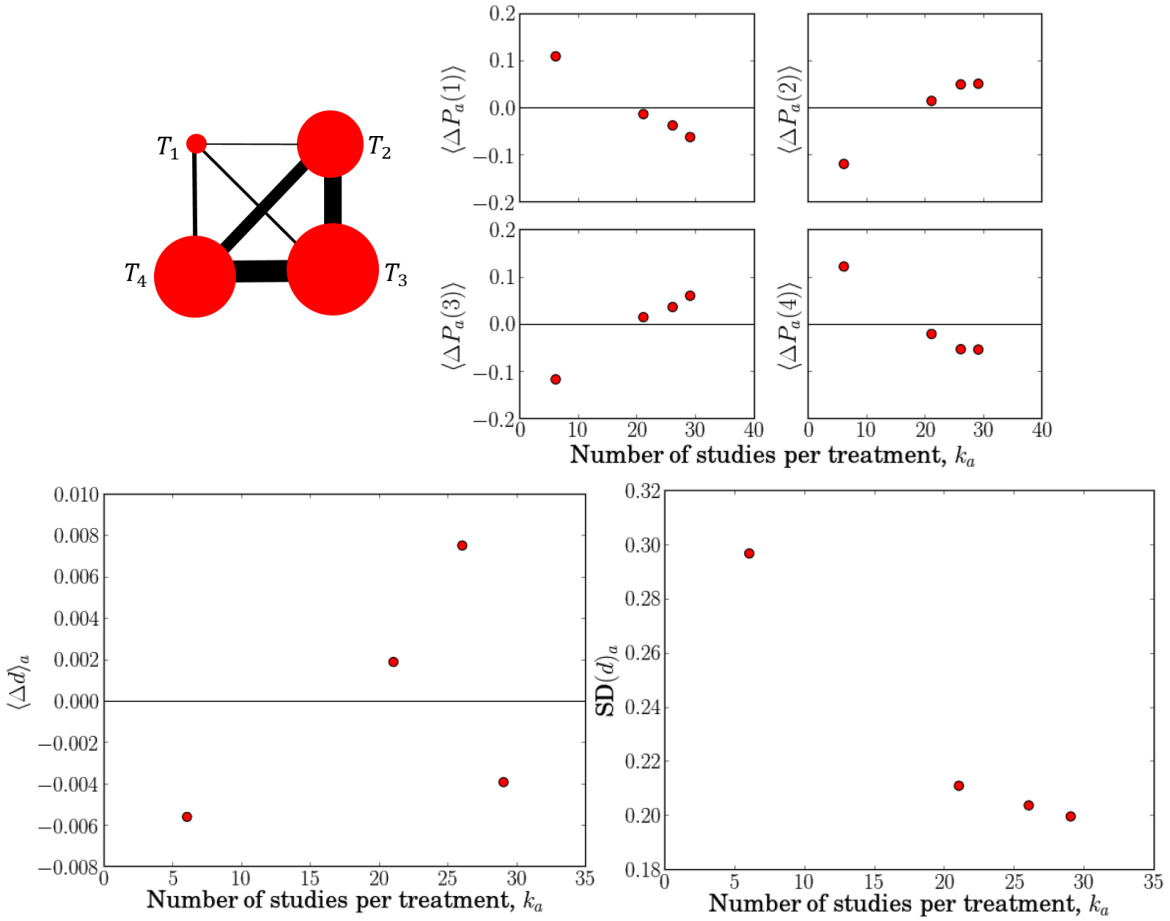


Figure 8.8: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (1, 2, 3, 12, 8, 15)$.

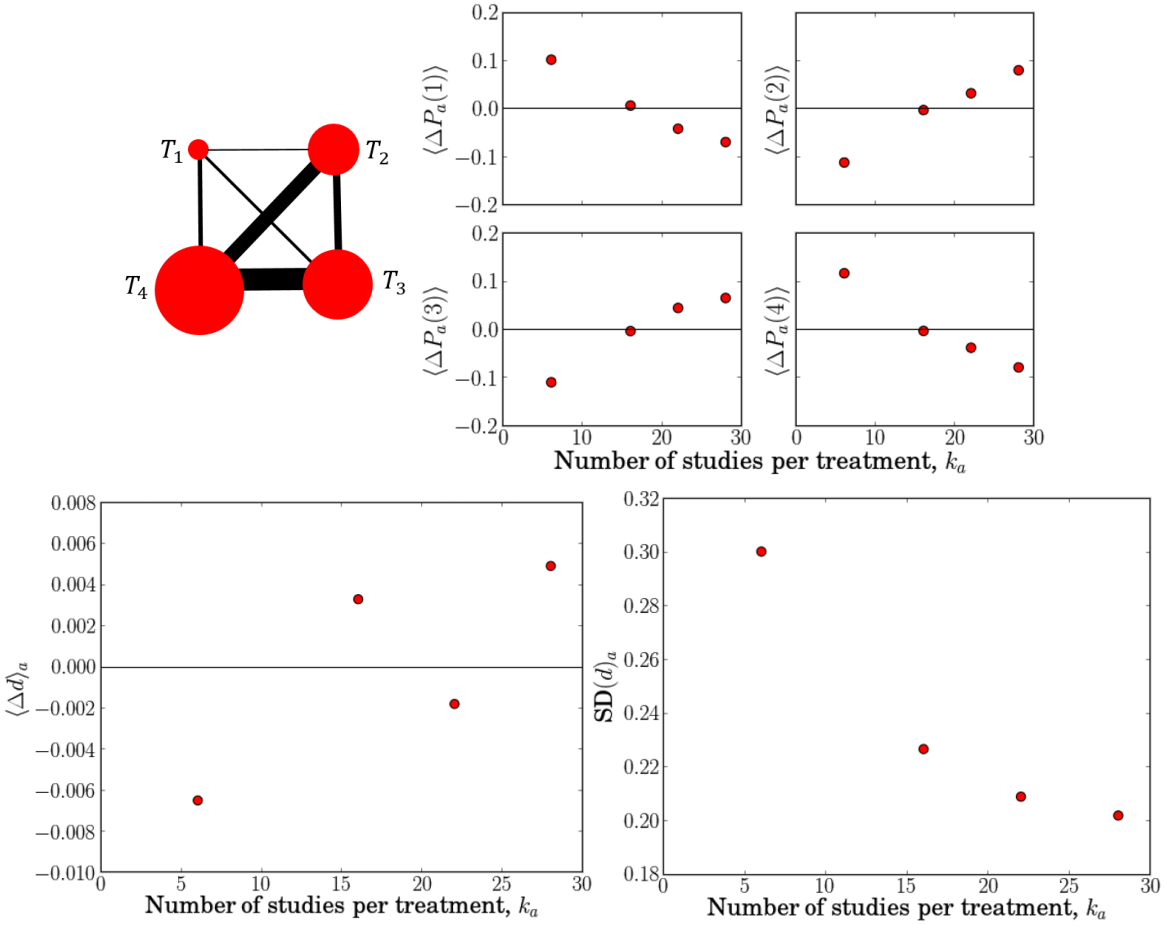


Figure 8.9: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (1, 2, 3, 5, 10, 15)$.

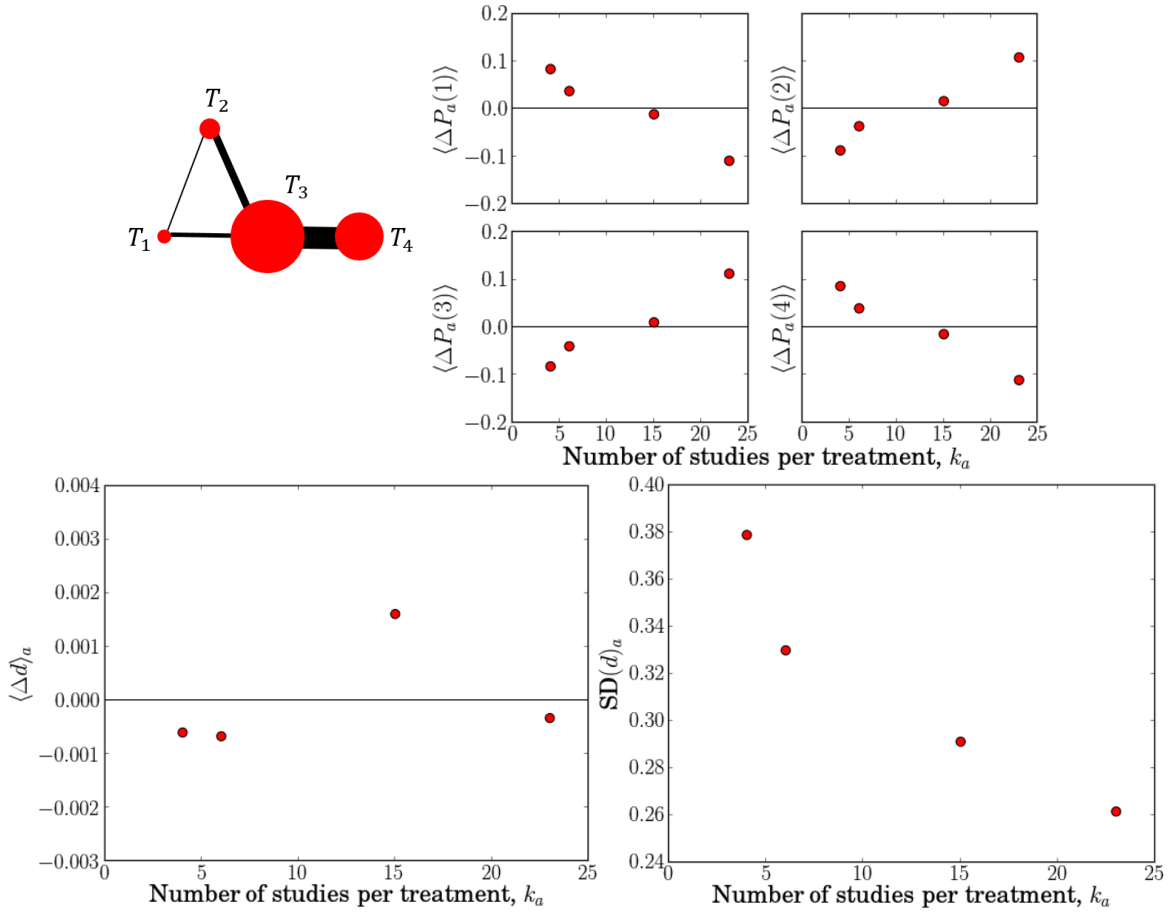


Figure 8.10: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a tadpole network with $\boldsymbol{K} = (1, 3, 0, 5, 0, 15)$.

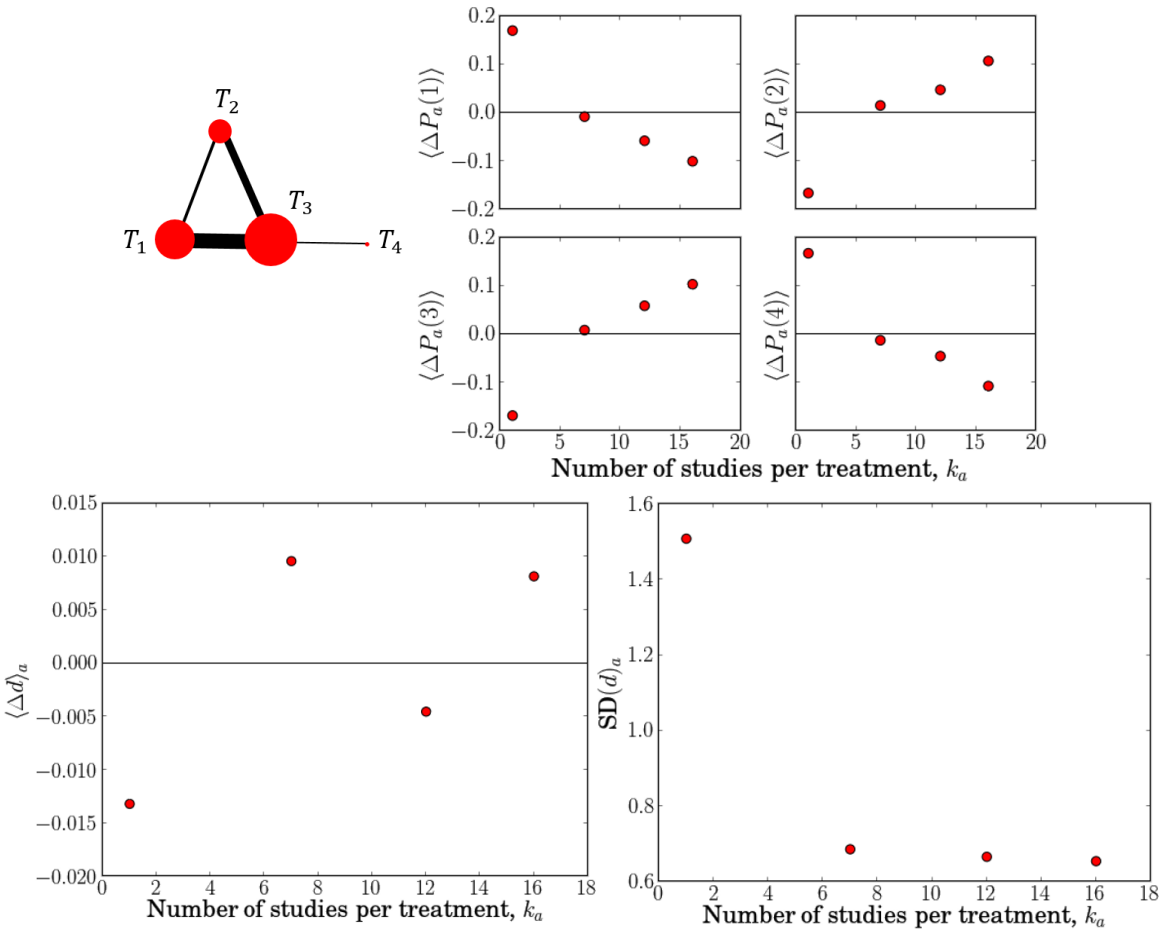


Figure 8.11: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a tadpole network with $\boldsymbol{K} = (2, 10, 0, 5, 0, 1)$.

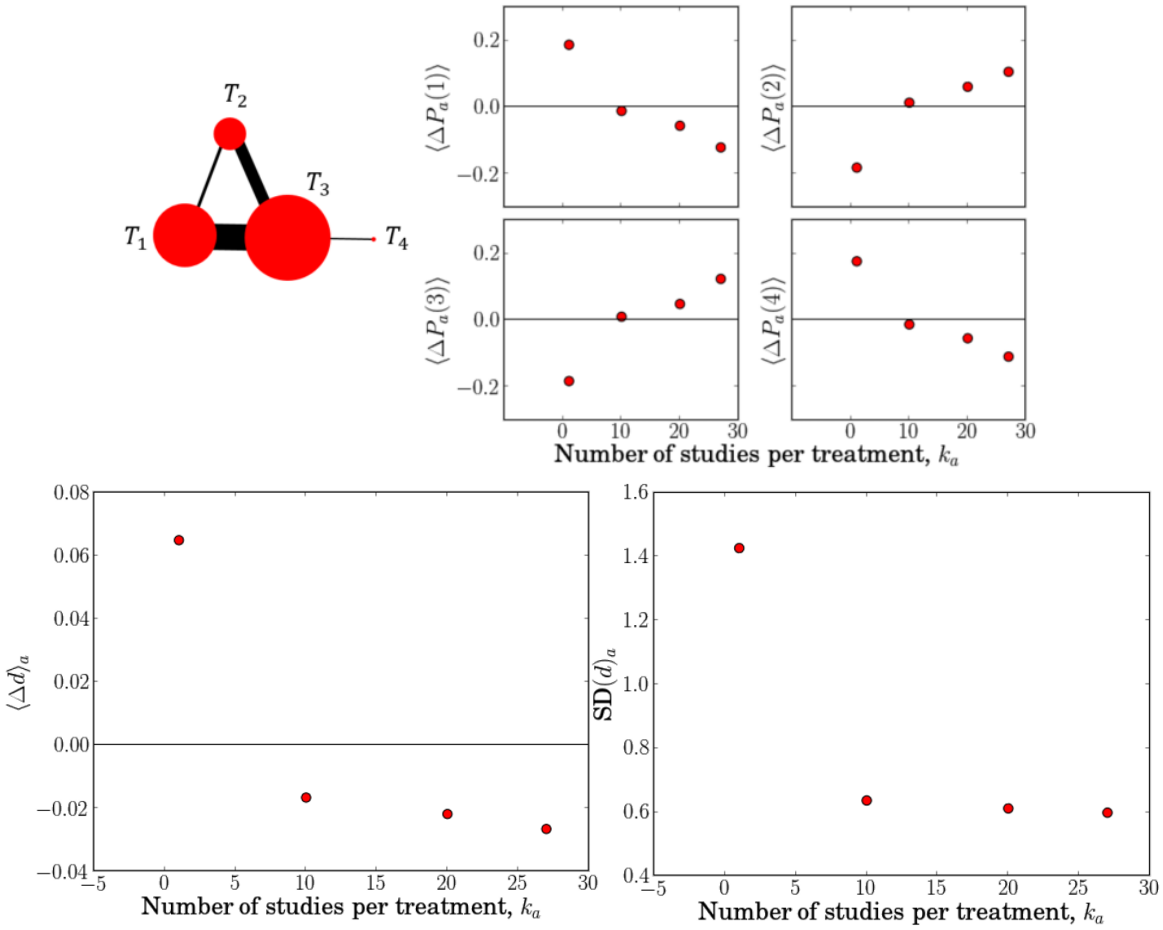


Figure 8.12: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a tadpole network with $\boldsymbol{K} = (2, 18, 0, 8, 0, 1)$.

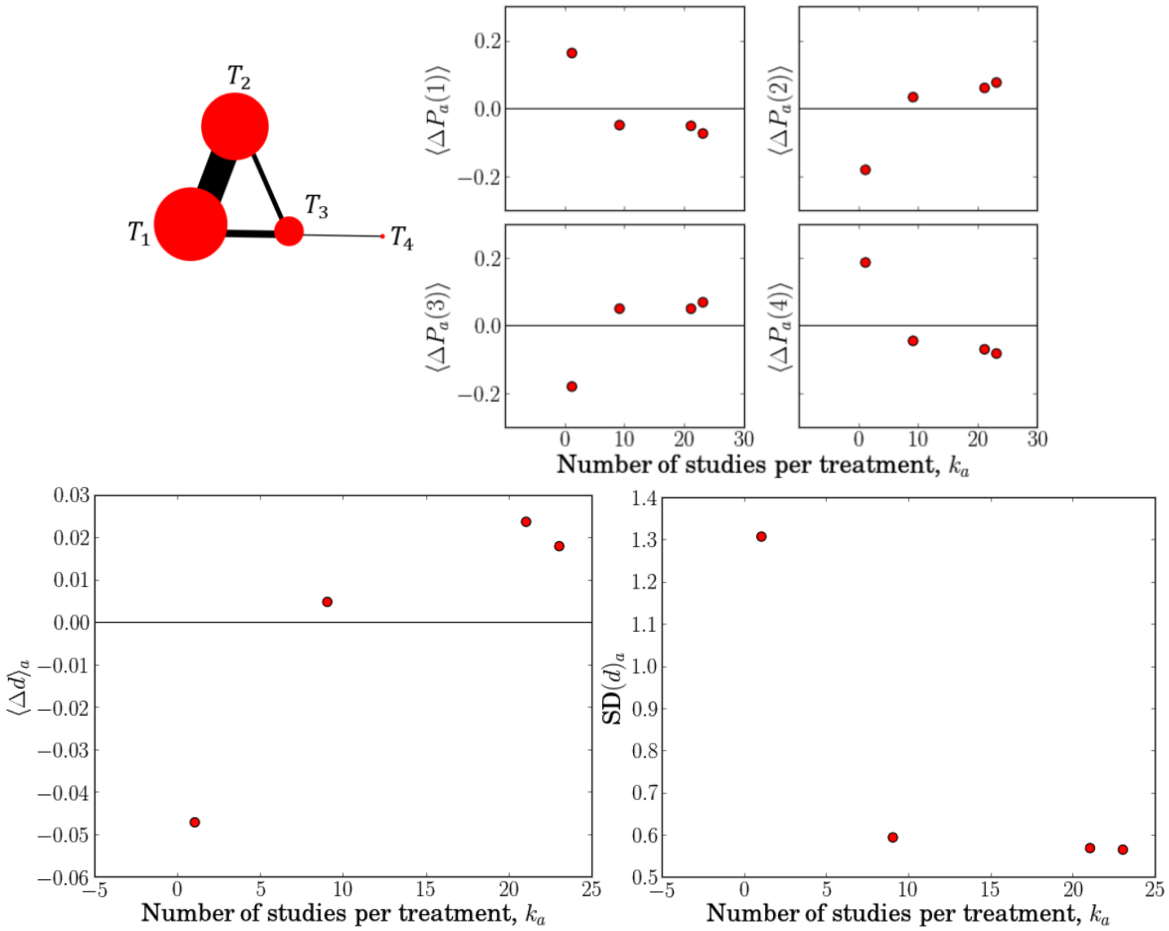


Figure 8.13: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a tadpole network with $\boldsymbol{K} = (18, 5, 0, 3, 0, 1)$.

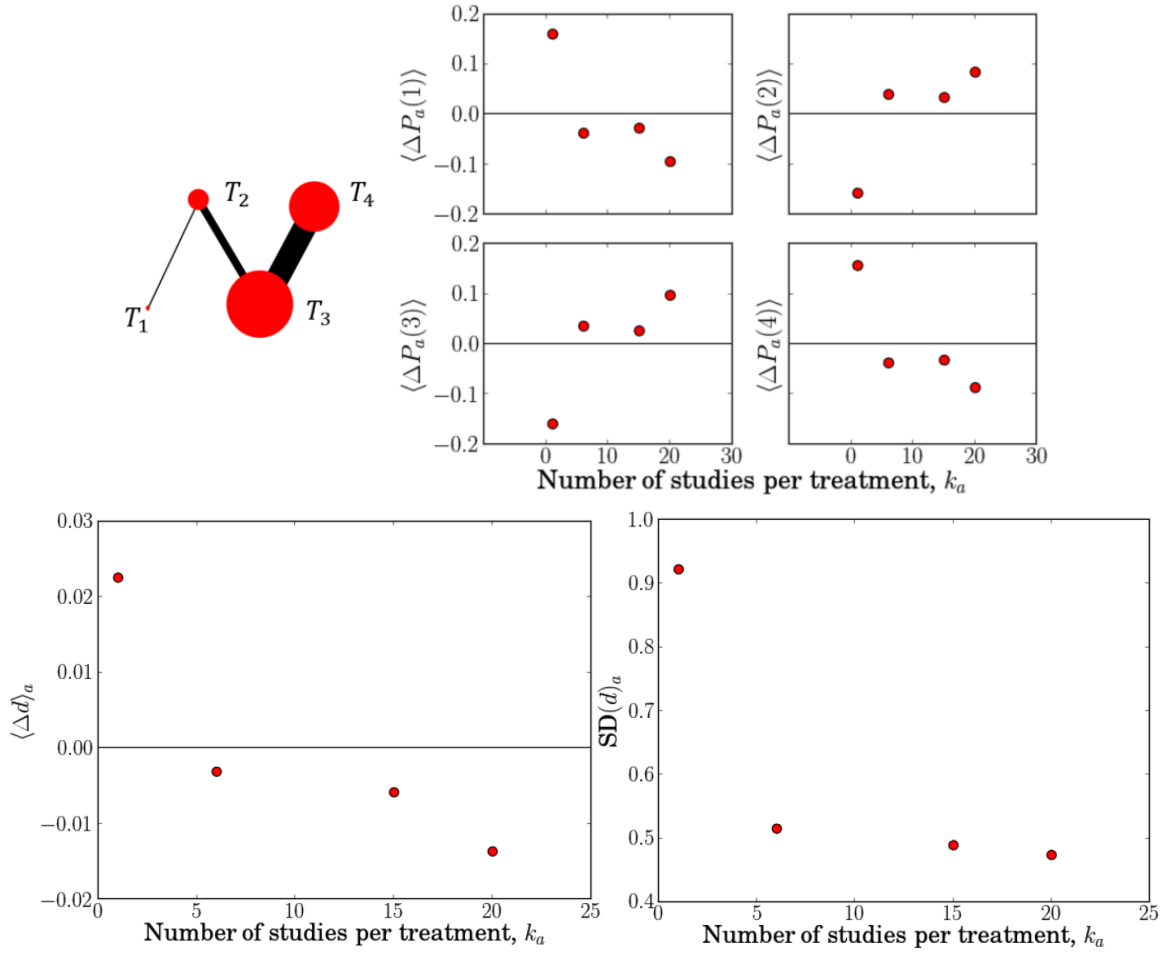


Figure 8.14: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a ladder network with $\boldsymbol{K} = (1, 0, 0, 5, 0, 15)$.

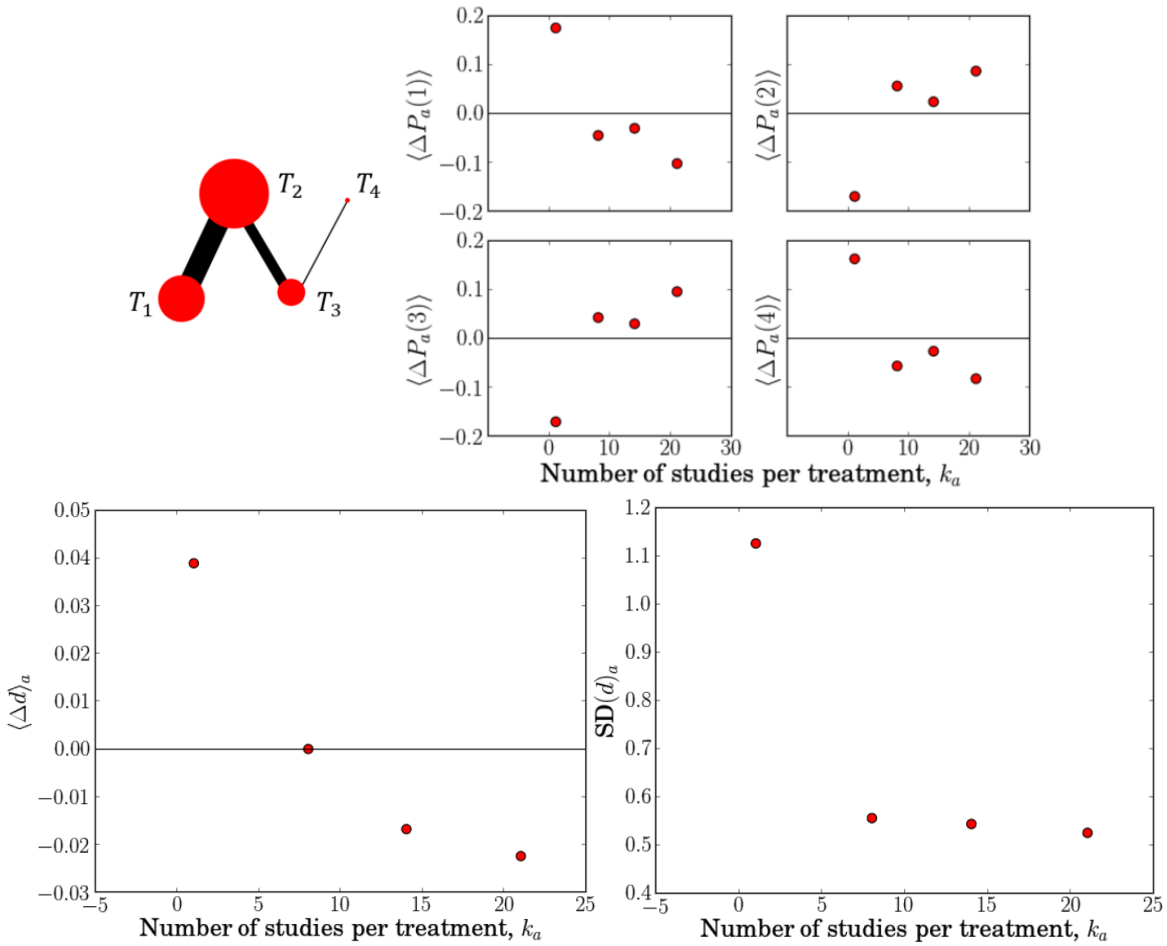


Figure 8.15: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a ladder network with $\boldsymbol{K} = (14, 0, 0, 7, 0, 1)$.

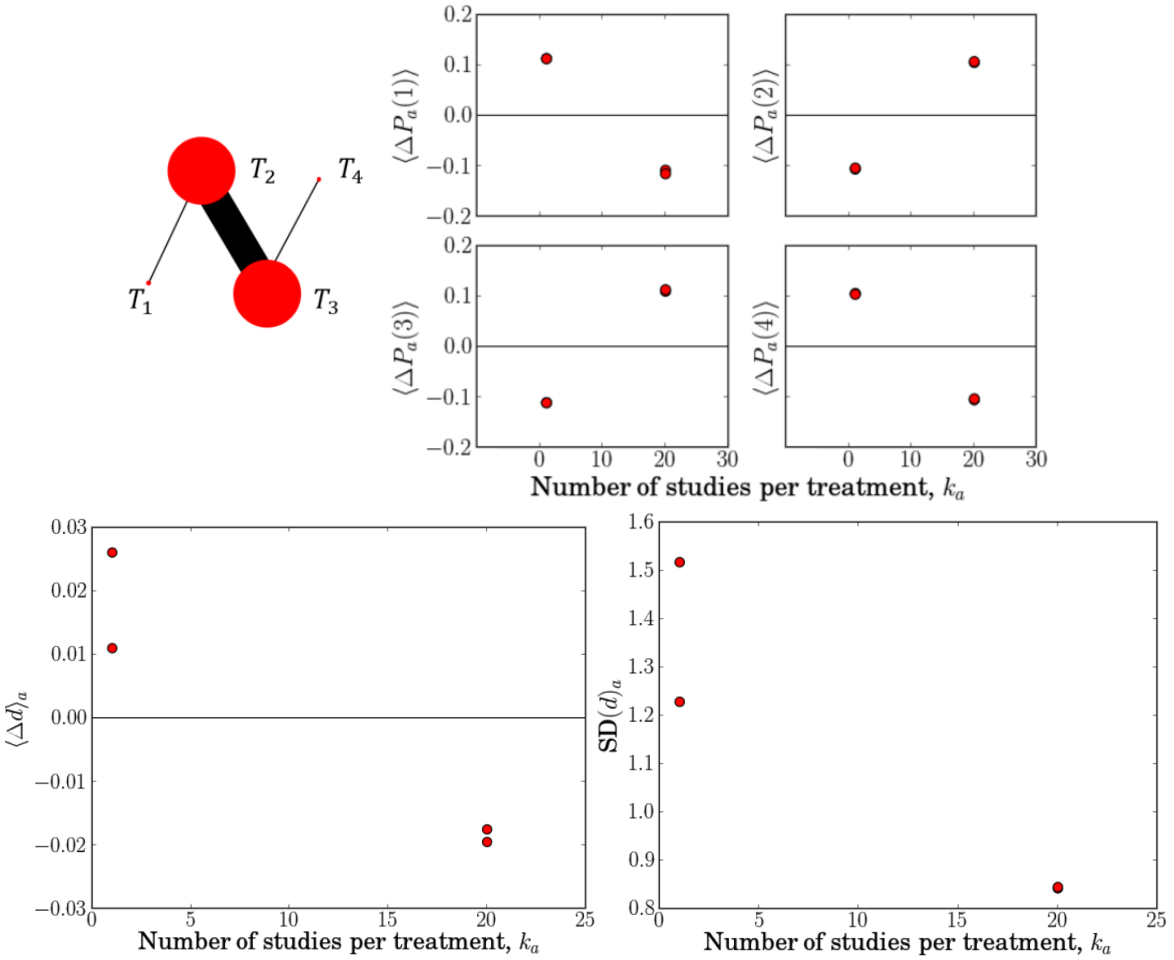


Figure 8.16: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a ladder network with $\boldsymbol{K} = (1, 0, 0, 19, 0, 1)$.

8.2 Within network plots: Bias on SUCRA

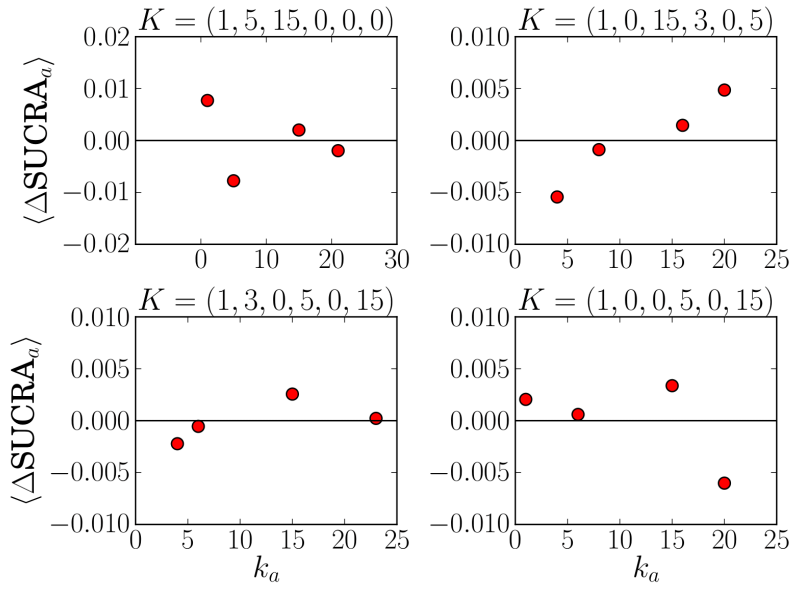


Figure 8.17: Some examples of the effect of the number of studies per treatment on SUCRA_a for different network geometries. For these examples $\boldsymbol{d} = (0, 0, 0)$ and the networks are made up of exclusively 2-arm trials.

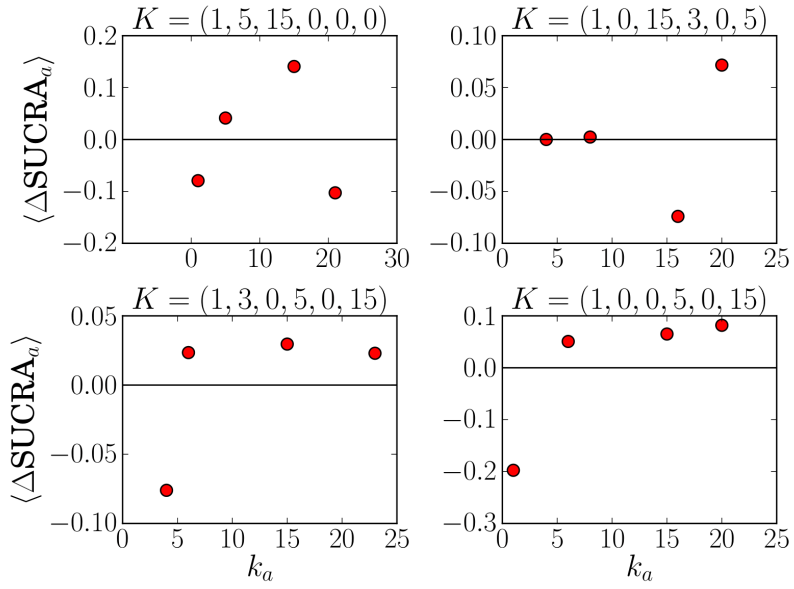


Figure 8.18: Some examples of the effect of the number of studies per treatment on SUCRA_a for different network geometries. For these examples $\boldsymbol{d} = (0.5, 1.0, 1.4)$ and the networks are made up of exclusively 2-arm trials.

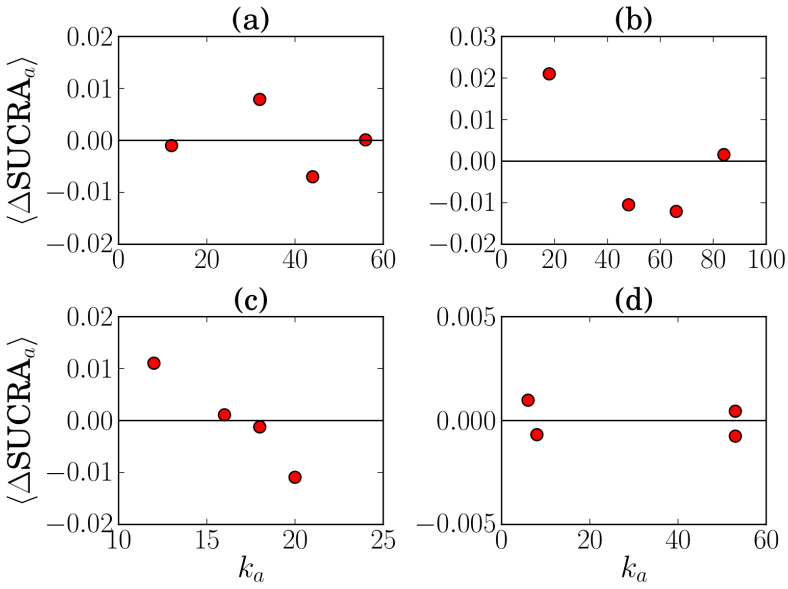


Figure 8.19: Some examples of the effect of the number of studies per treatment on SUCRA_a for different network geometries. For these examples $\boldsymbol{d} = (0, 0, 0)$ and the networks contain multi-arm trials. We use n_m to indicate the number of m-arm trials. Figure (a): $\boldsymbol{K} = (2, 4, 6, 10, 20, 30)$ $(n_2, n_3, n_4) = (66, 0, 1)$, Figure (b): $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$ $(n_2, n_3, n_4) = (90, 4, 1)$, Figure (c): $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$ $(n_2, n_3, n_4) = (21, 4, 0)$, Figure (d): $\boldsymbol{K} = (2, 2, 2, 3, 3, 48)$ $(n_2, n_3, n_4) = (48, 0, 2)$.

8.3 Within network plots: Rank probability for non-equally effective treatments

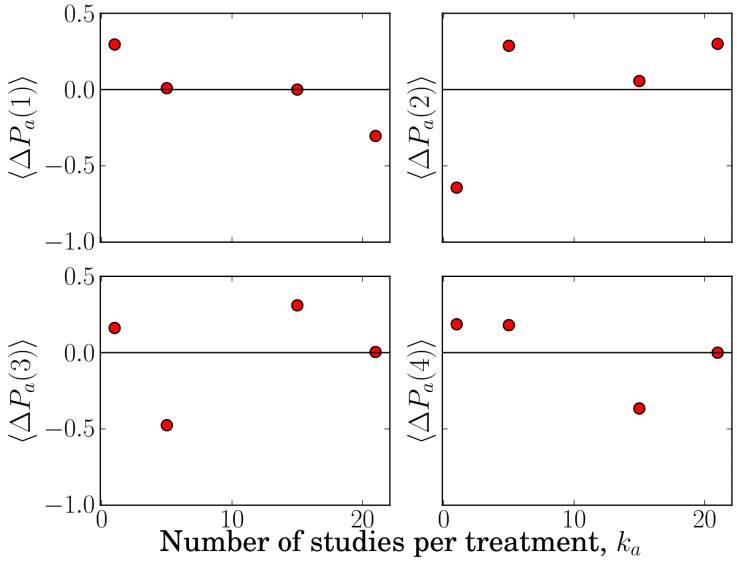


Figure 8.20: Bias of rank probability against the number of studies per treatment for a star network with $\boldsymbol{K} = (1, 5, 15, 0, 0, 0)$ and non-equally effective treatments, $\boldsymbol{d} = (0.5, 1.0, 1.4)$.

8.4 Between network plots: Treatment effect bias and irregularity

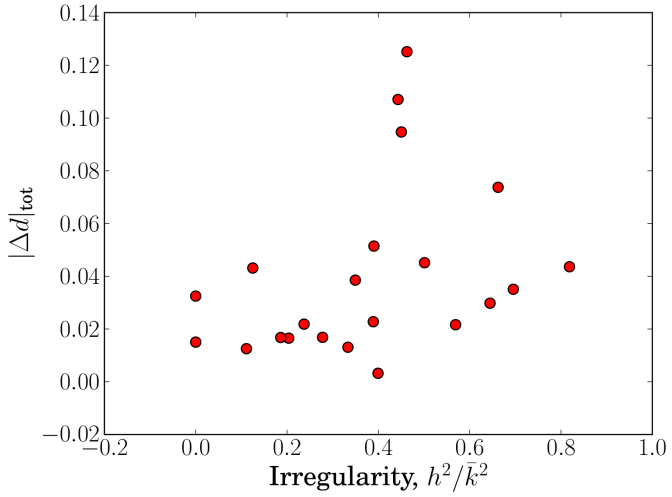


Figure 8.21: The effect of irregularity on the total bias of treatment effects.

8.5 Between network plots: The effect of the total number of studies

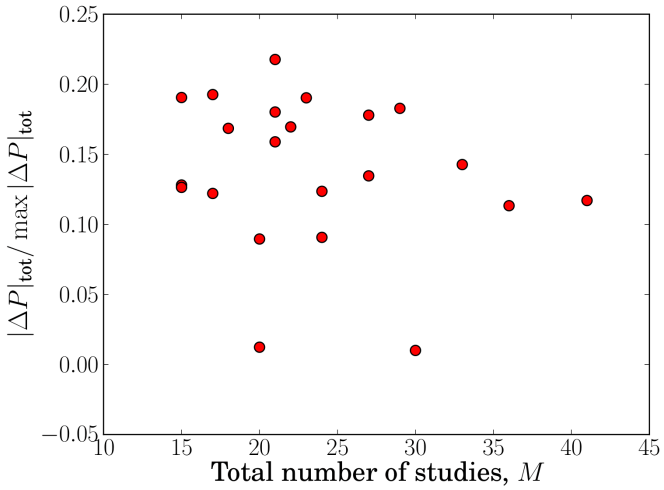


Figure 8.22: The effect of the total number of studies in the network on total rank probability bias. Total bias is plotted as a proportion of the maximum total rank probability bias.

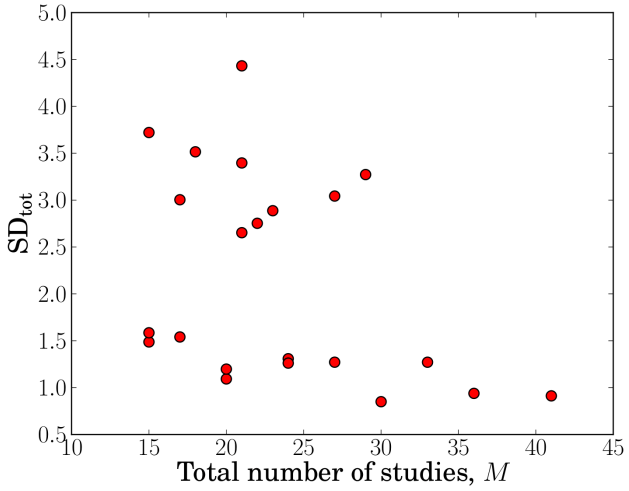


Figure 8.23: The effect of the total number of studies in the network on the network's total standard deviation on treatment effect estimates.

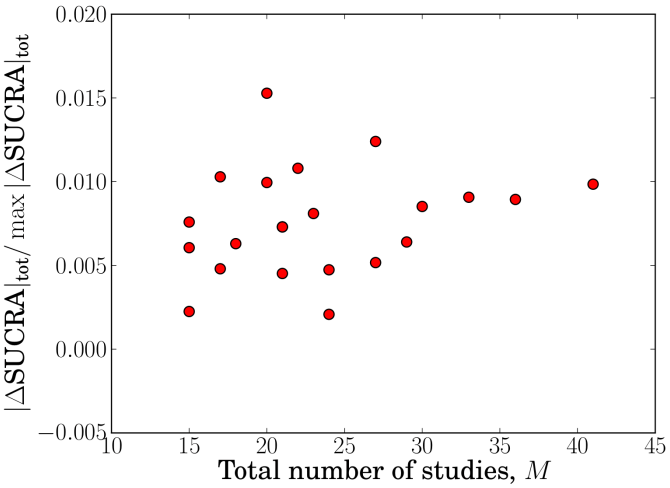


Figure 8.24: The effect of the total number of studies in the network on a network's total bias on SUCRA values. Total bias is plotted as a proportion of the maximum total SUCRA bias.

8.6 Multi-arm studies: Within network plots

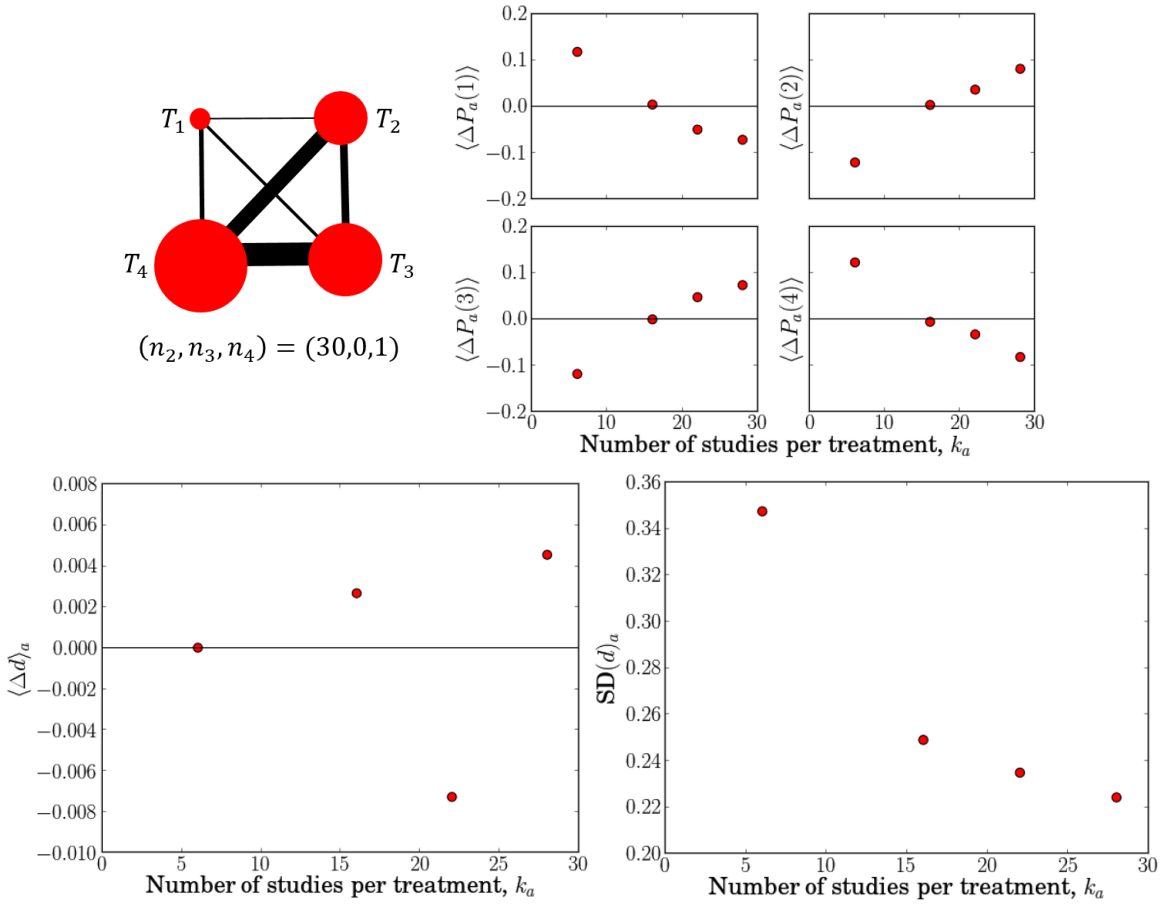


Figure 8.25: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (1, 2, 3, 5, 10, 15)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (30, 0, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

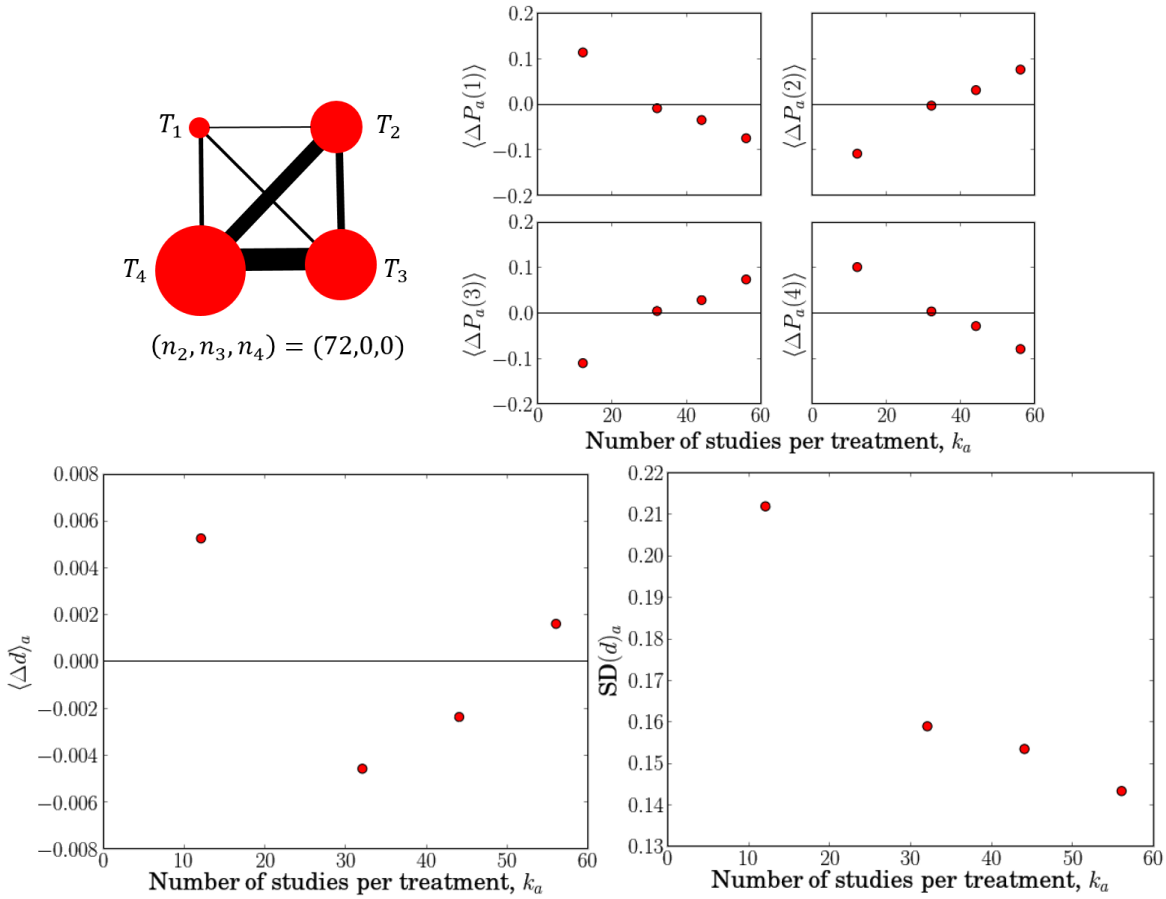


Figure 8.26: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 4, 6, 10, 20, 30)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (72, 0, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

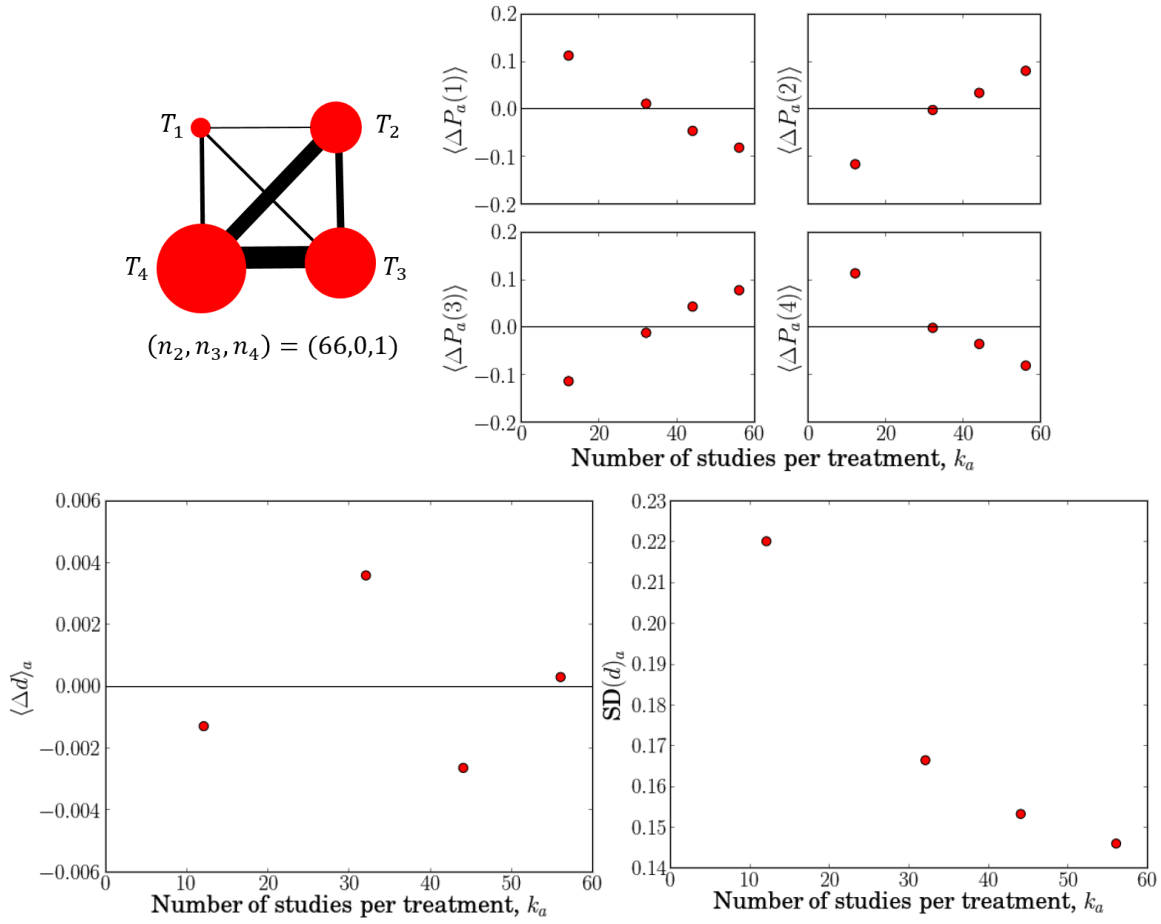


Figure 8.27: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 4, 6, 10, 20, 30)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (66, 0, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

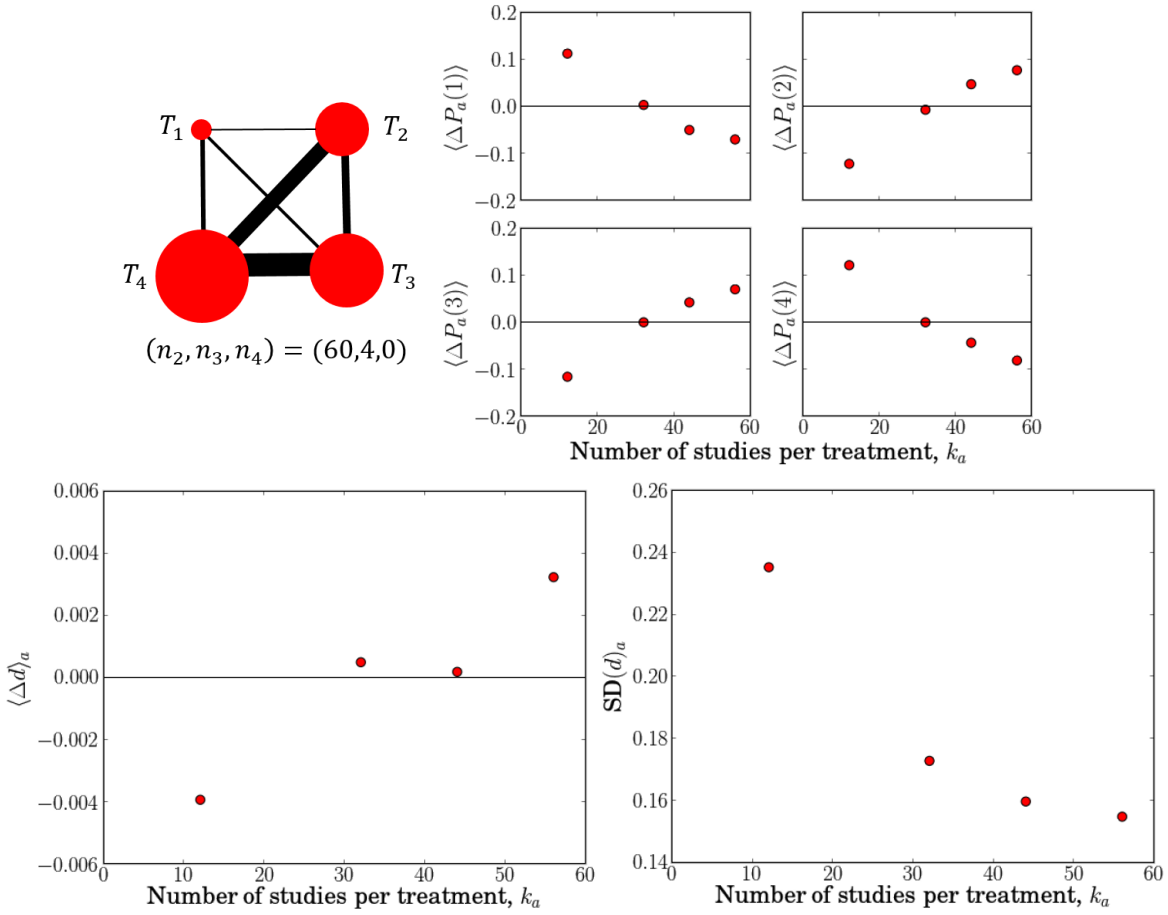


Figure 8.28: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 4, 6, 10, 20, 30)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (60, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

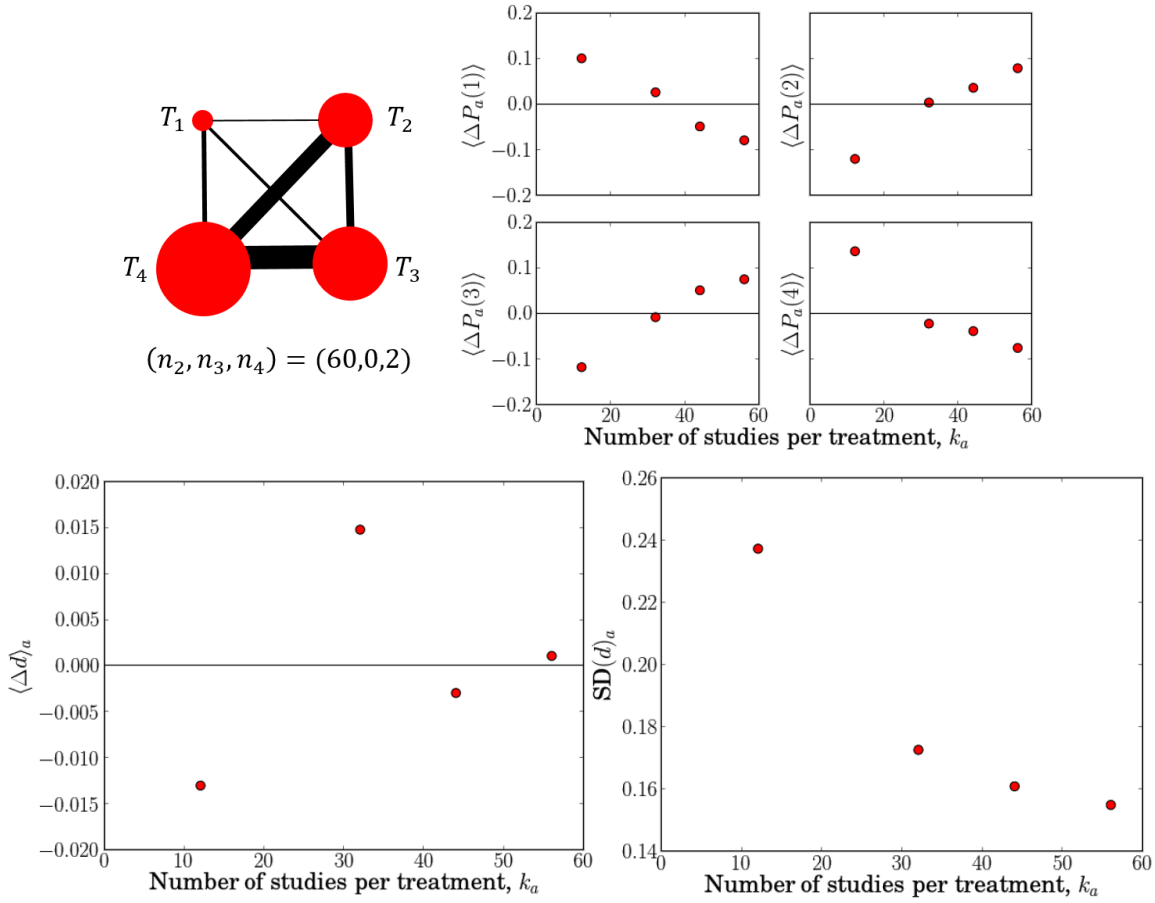


Figure 8.29: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 4, 6, 10, 20, 30)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (60, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

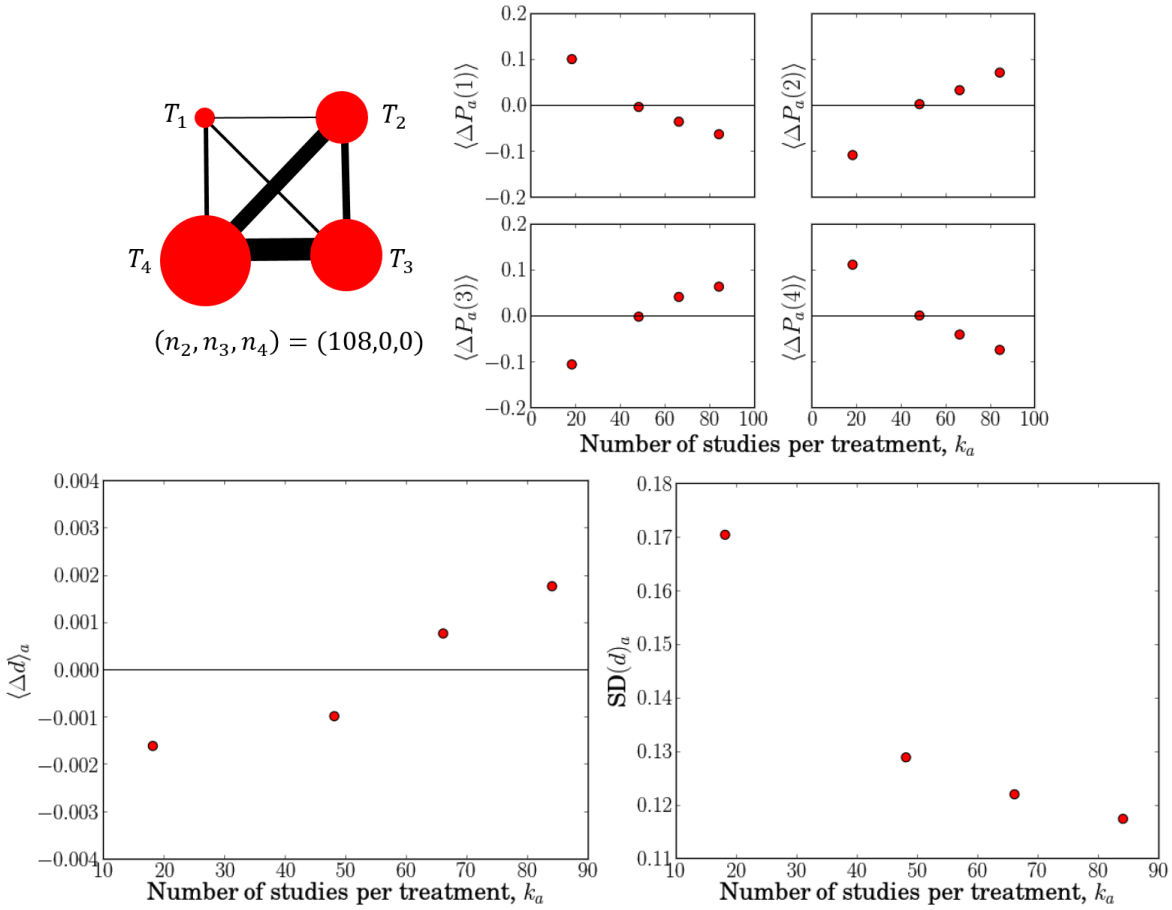


Figure 8.30: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (108, 0, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

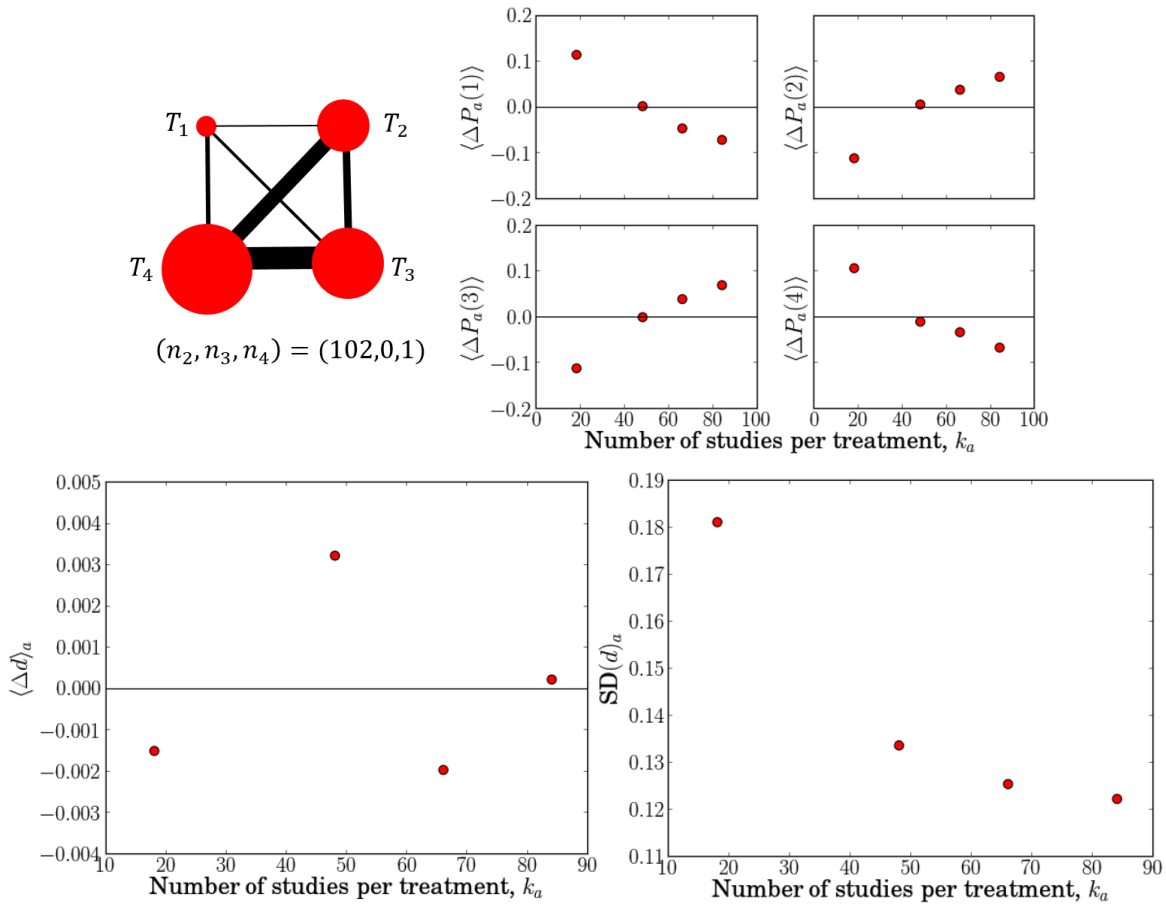


Figure 8.31: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (102, 0, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

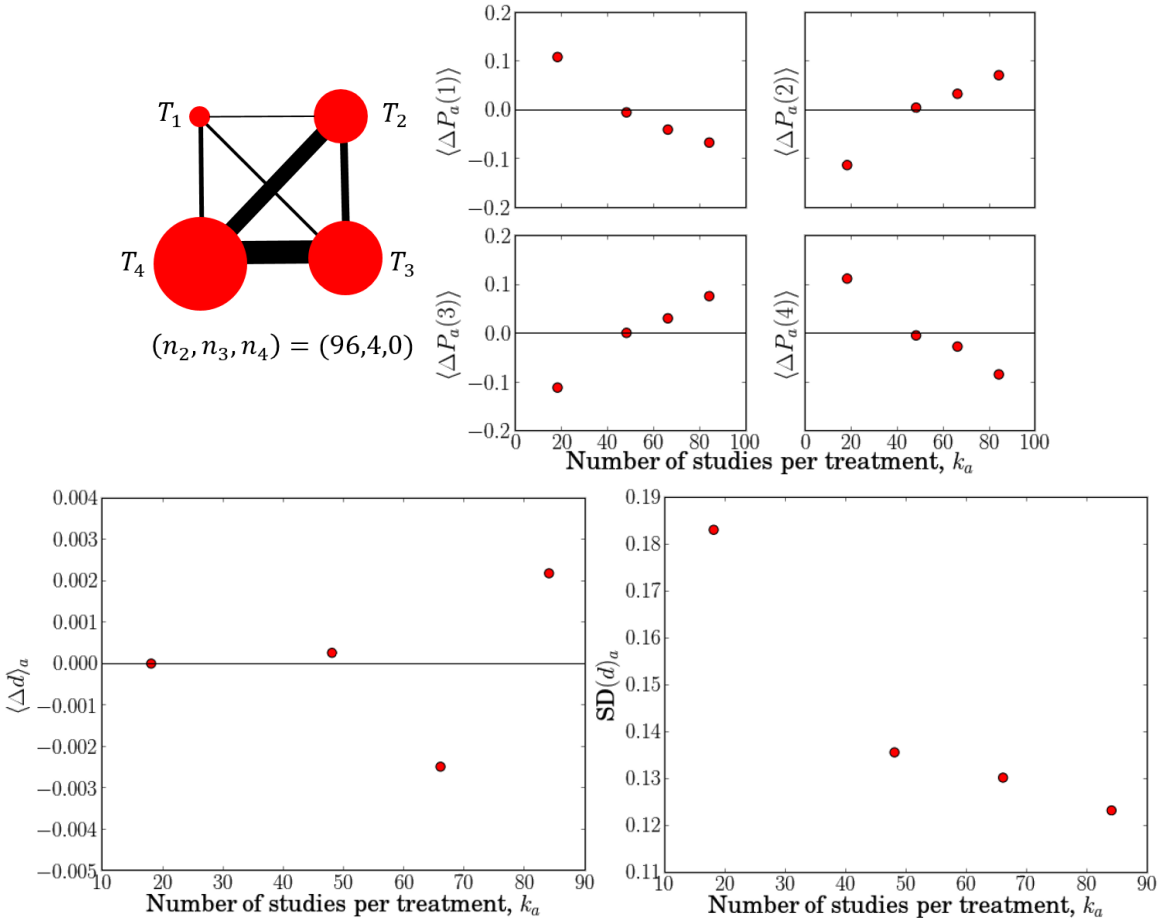


Figure 8.32: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (96, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

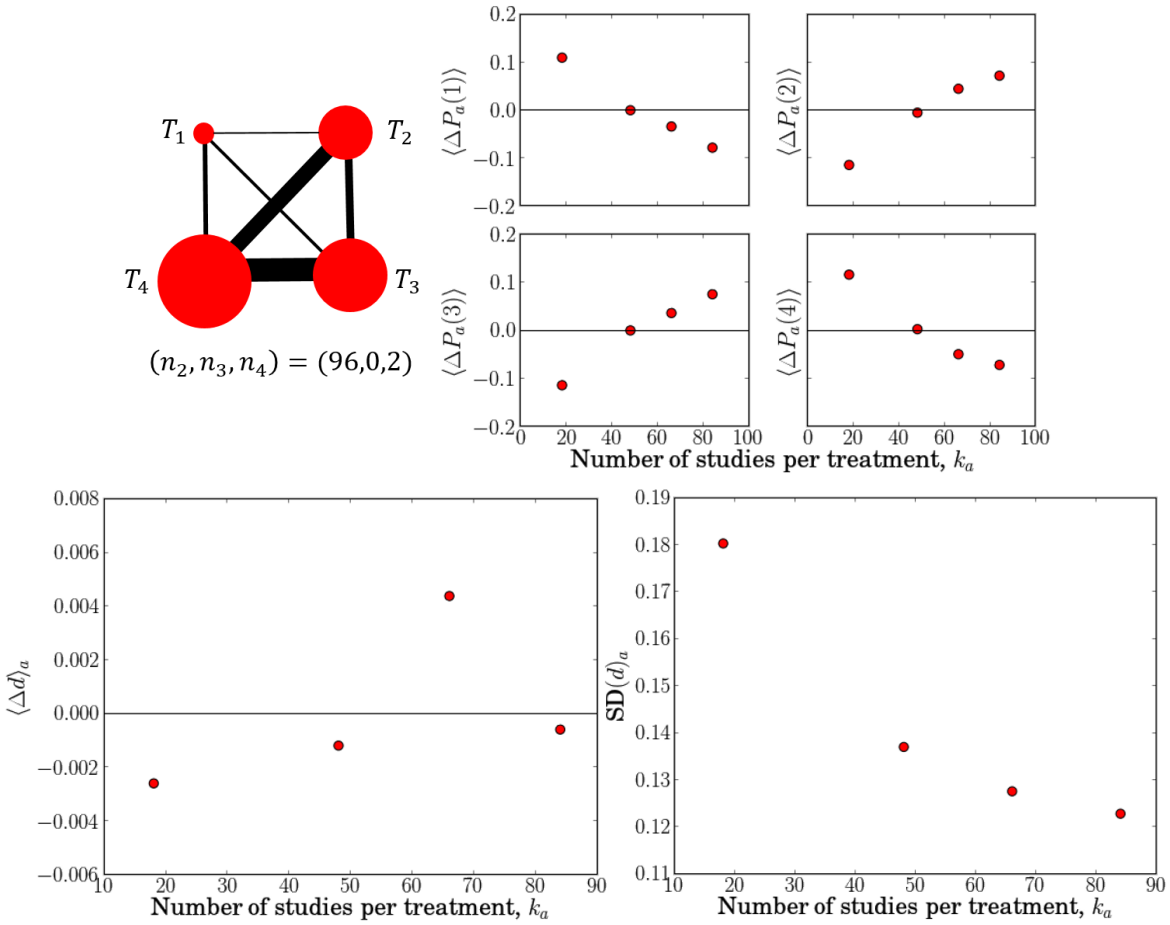


Figure 8.33: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (96, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

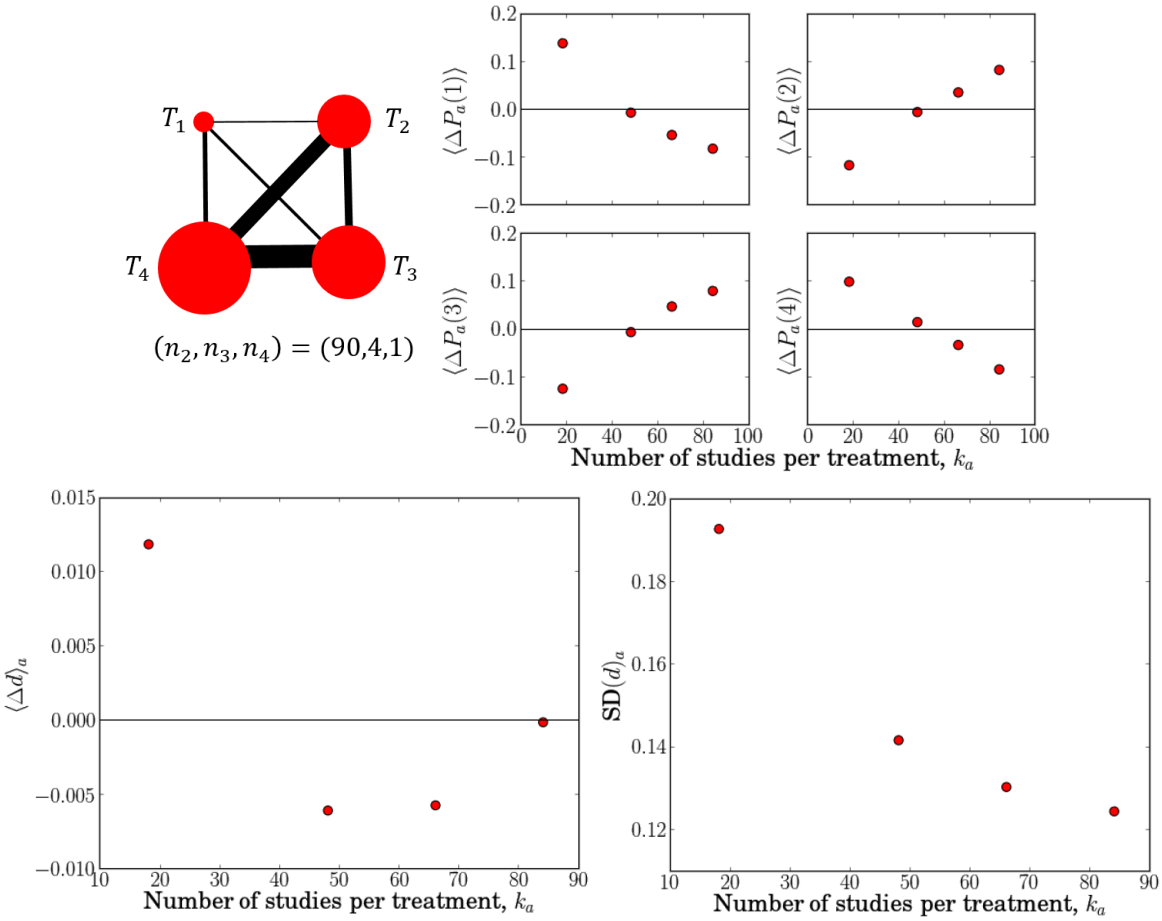


Figure 8.34: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (90, 4, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

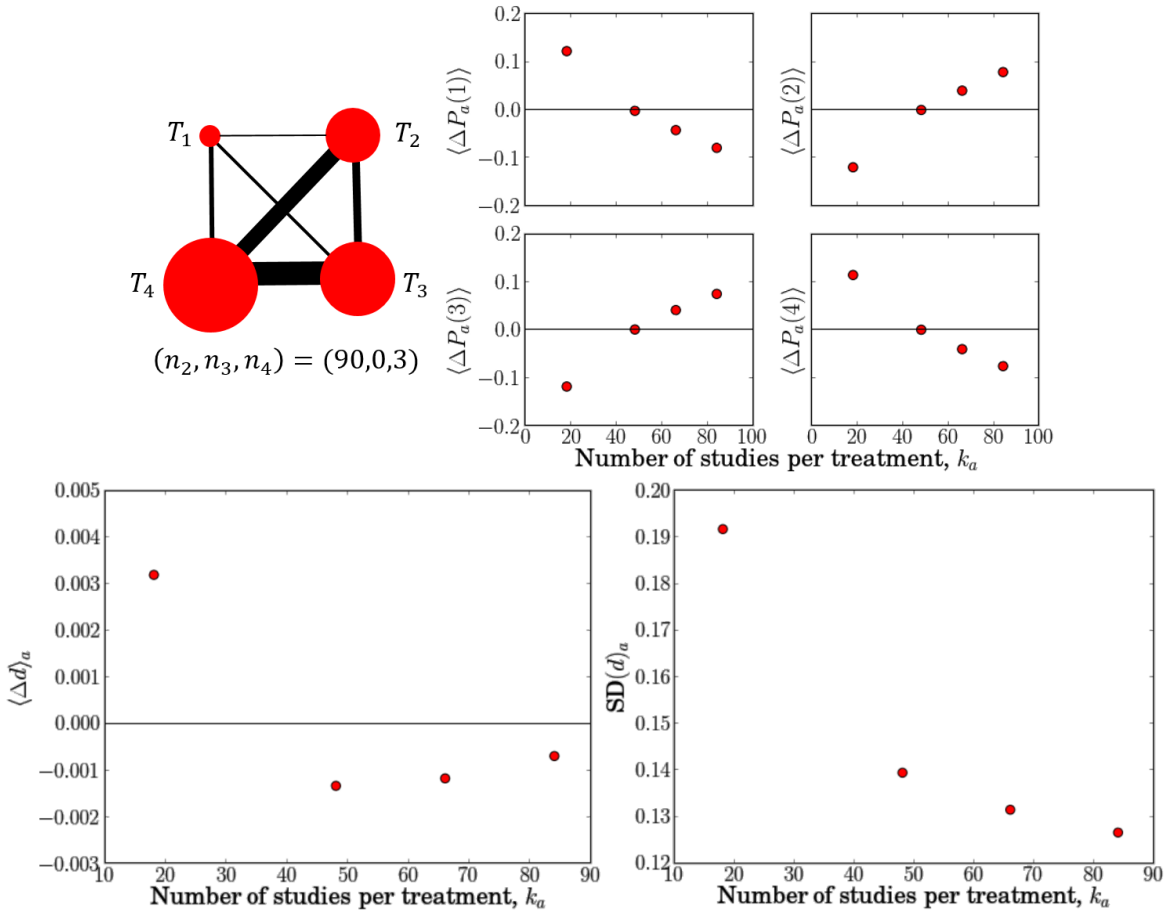


Figure 8.35: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 6, 9, 15, 30, 45)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (90, 0, 3)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.203704$.

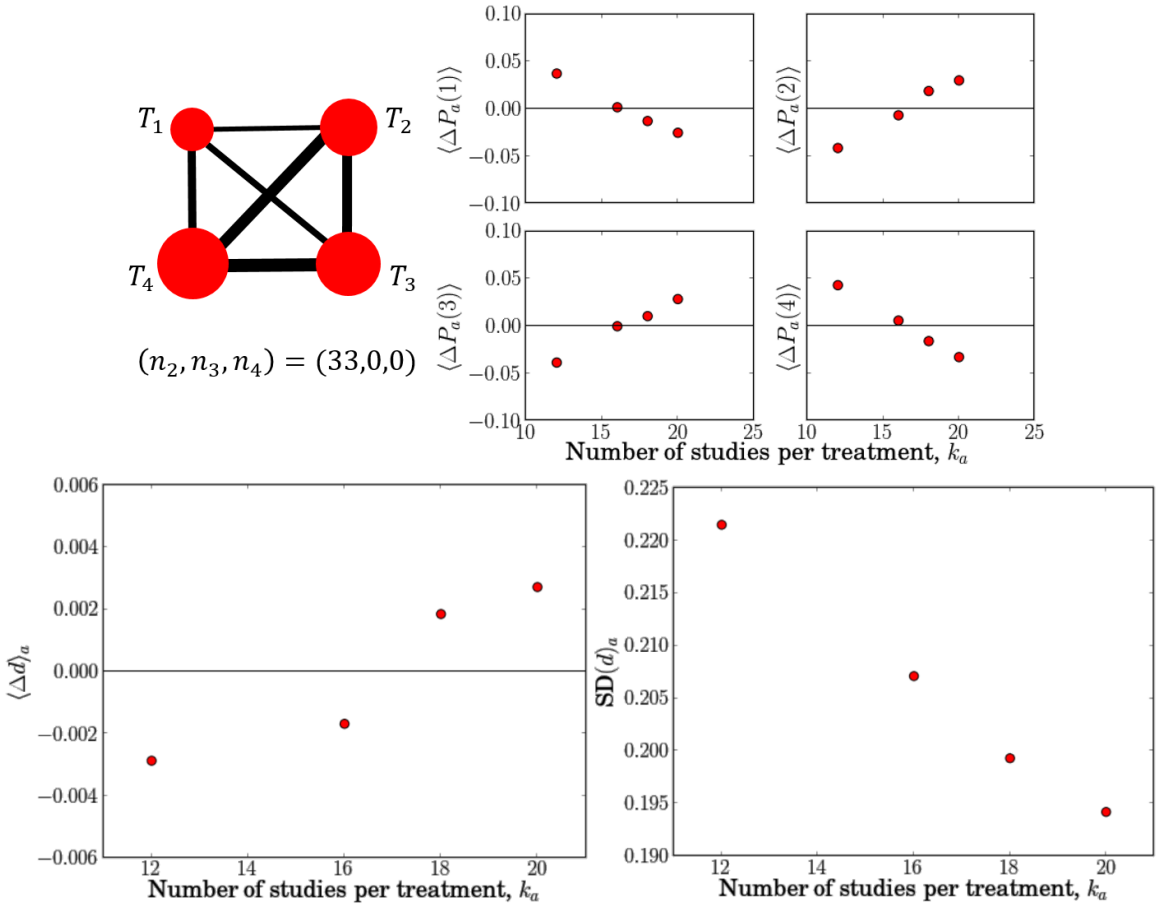


Figure 8.36: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (33, 0, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

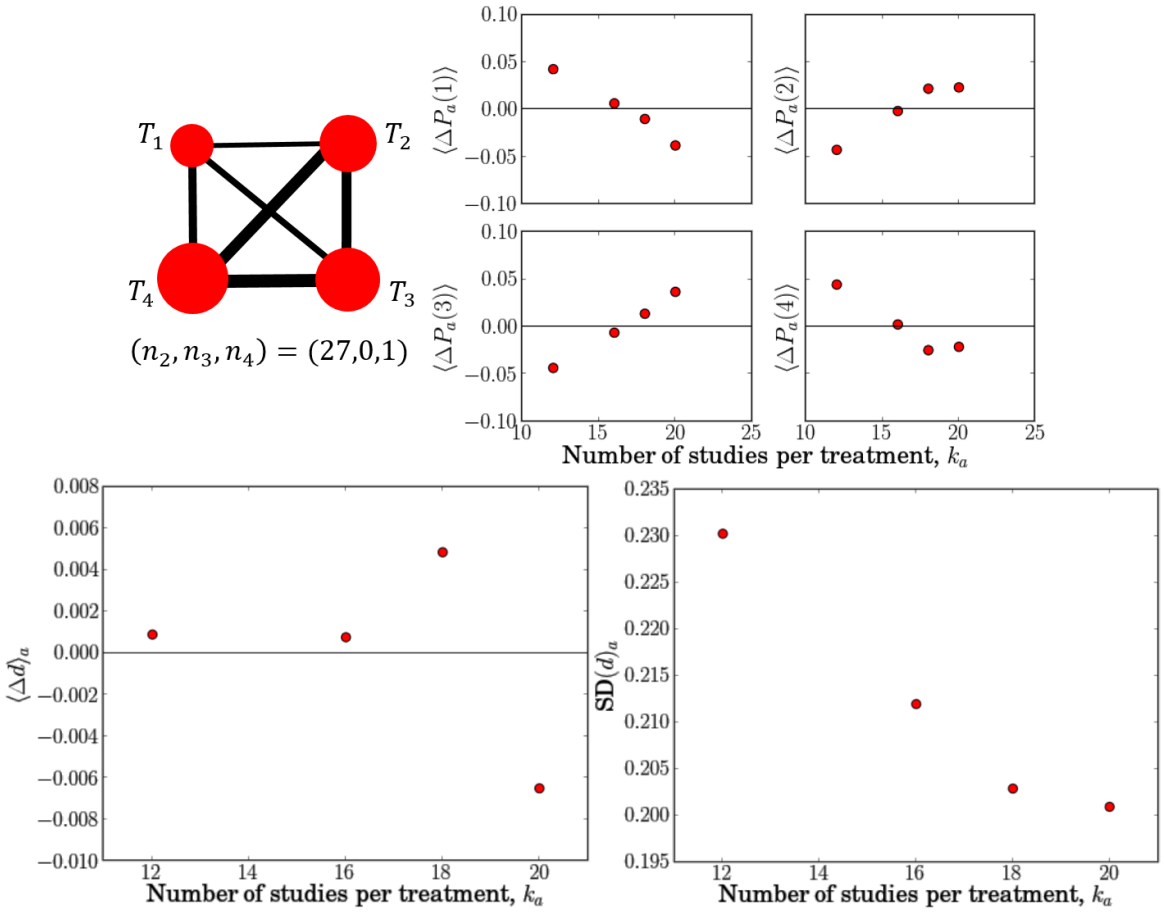


Figure 8.37: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (27, 0, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

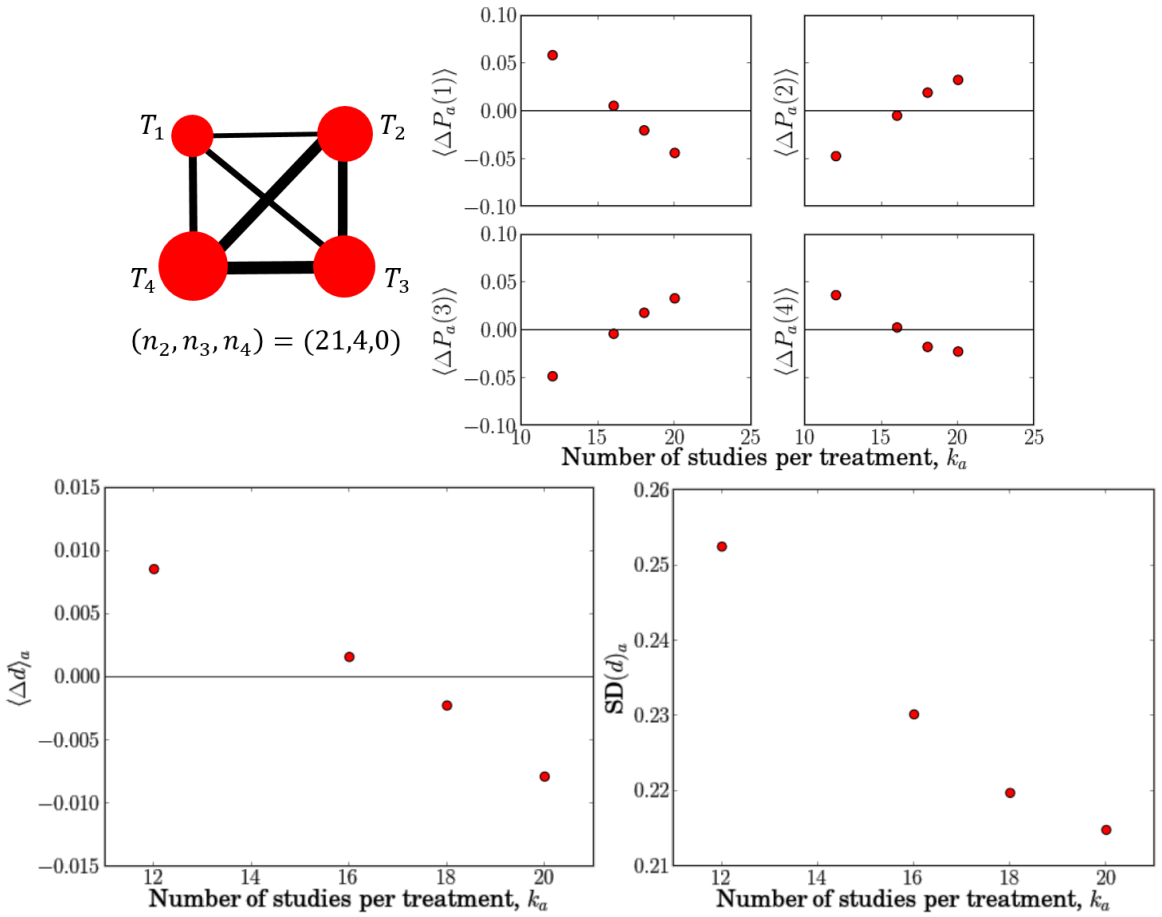


Figure 8.38: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (21, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

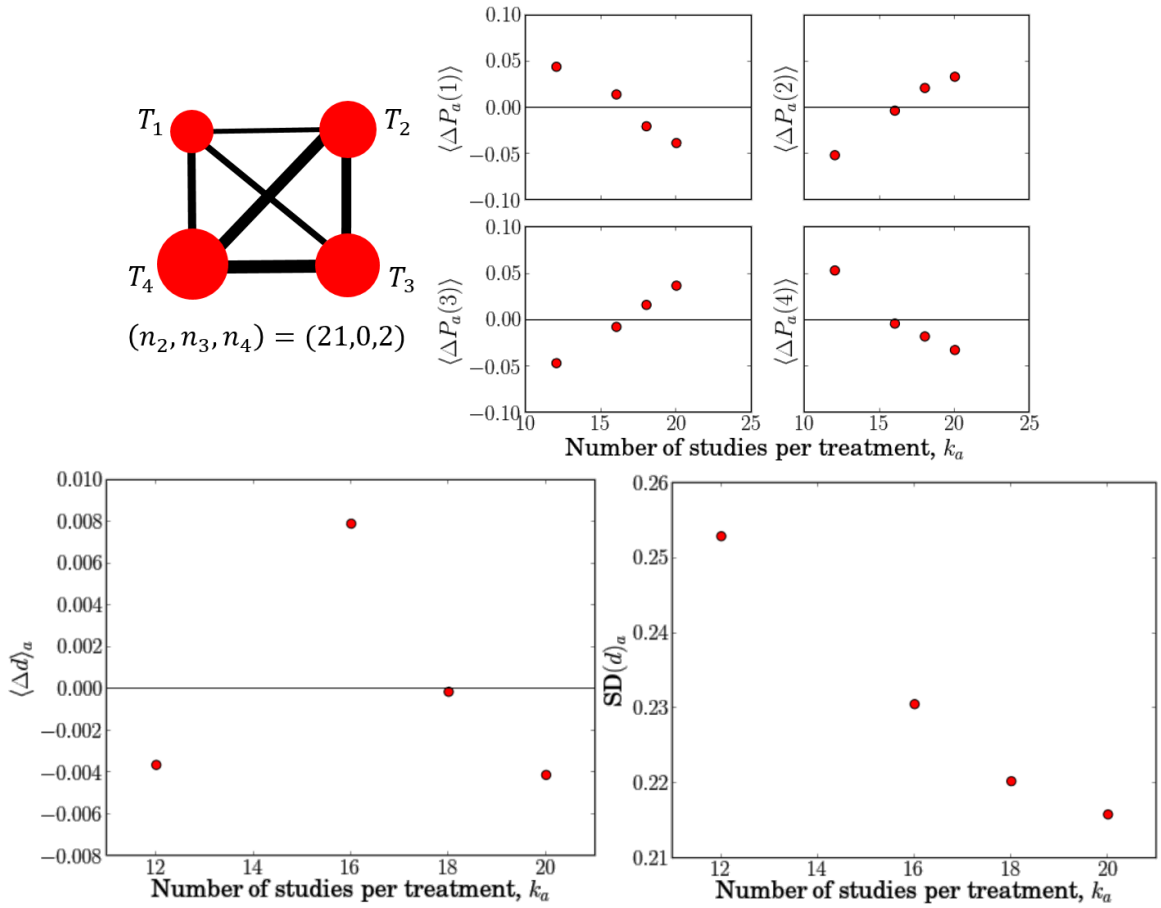


Figure 8.39: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (21, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

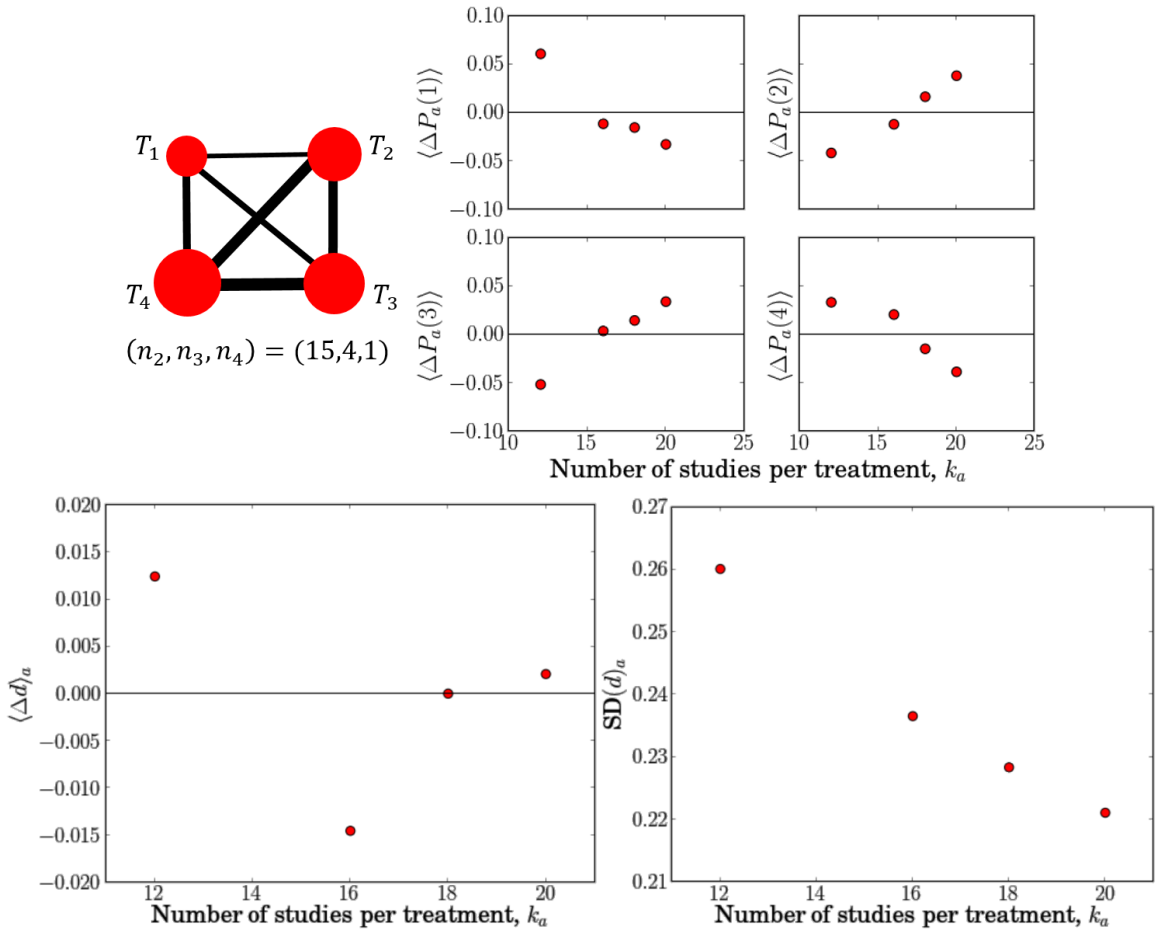


Figure 8.40: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (15, 4, 1)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

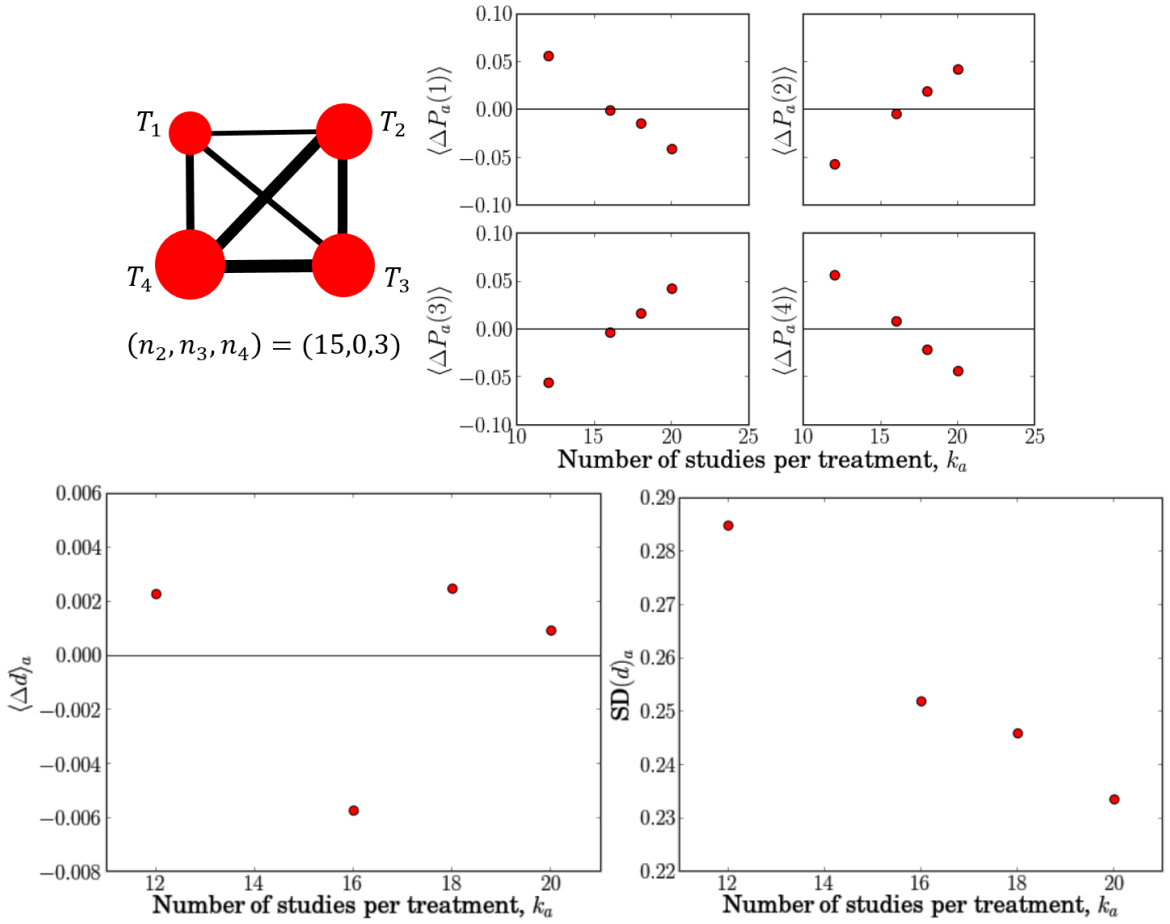


Figure 8.41: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (3, 4, 5, 6, 7, 8)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (15, 0, 3)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.032140$.

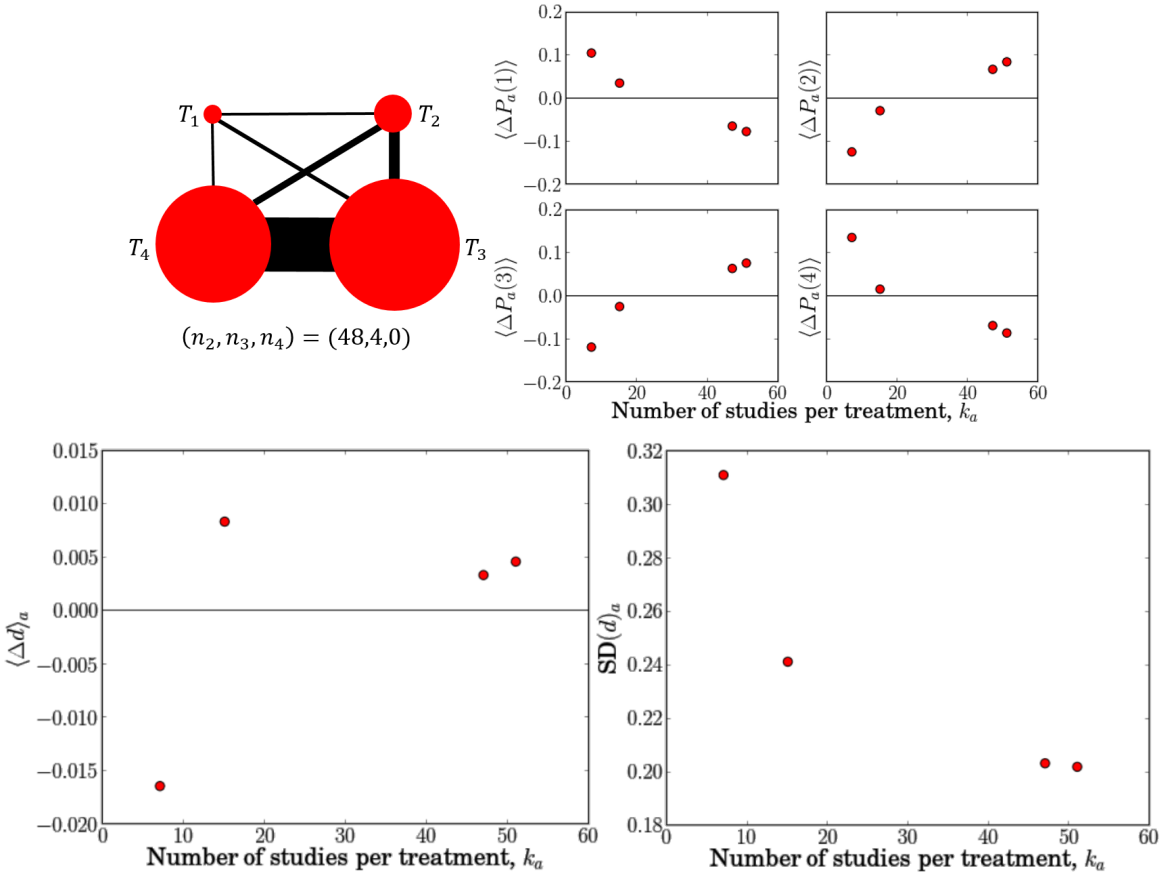


Figure 8.42: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 3, 2, 8, 5, 40)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (48, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.412222$.

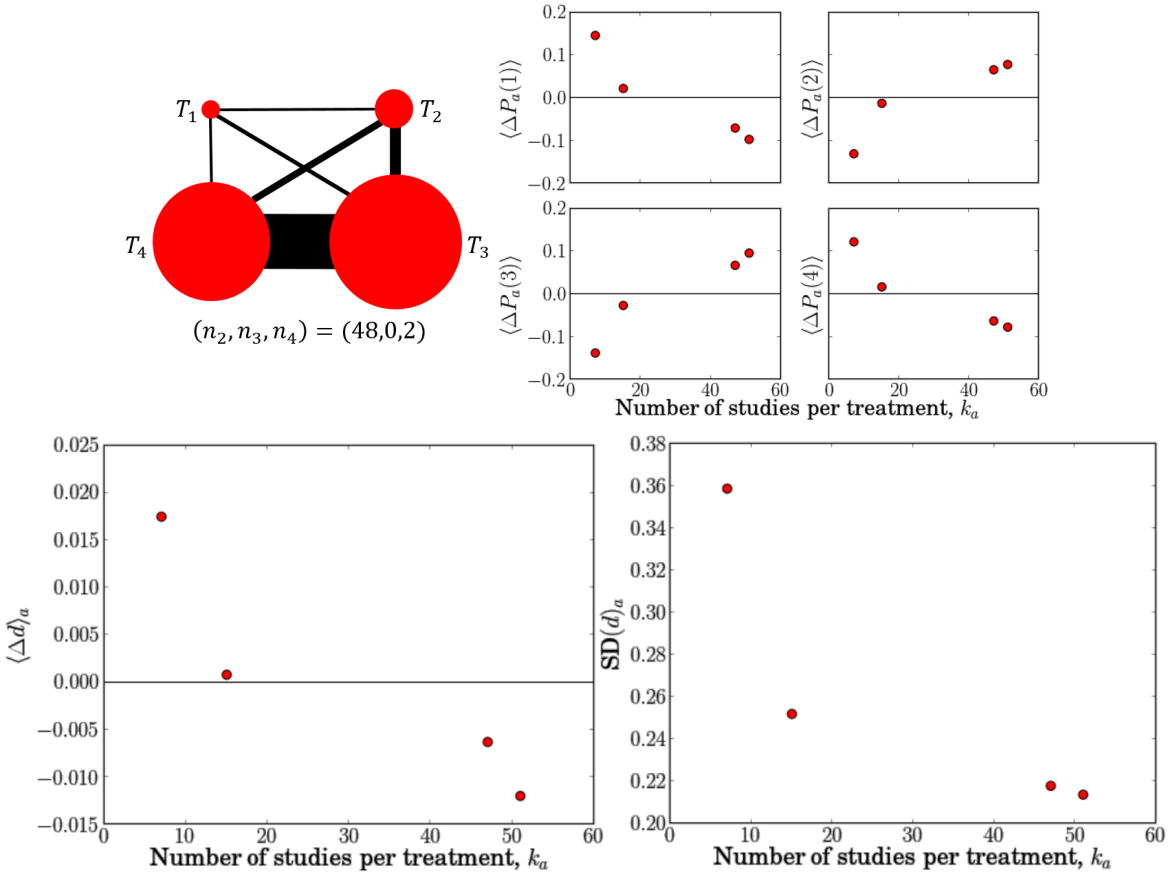


Figure 8.43: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 3, 2, 8, 5, 40)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (48, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.412222$.

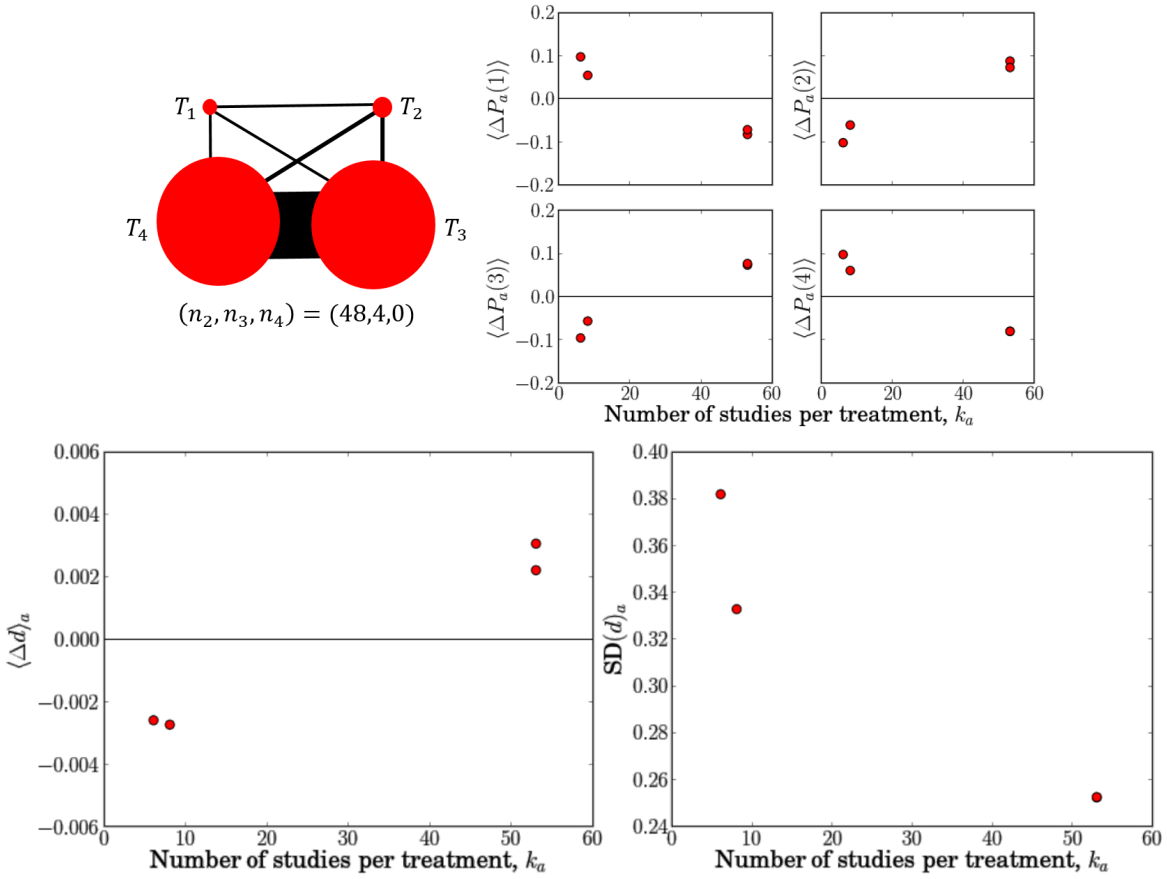


Figure 8.44: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 2, 2, 3, 3, 48)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (48, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.588333$.

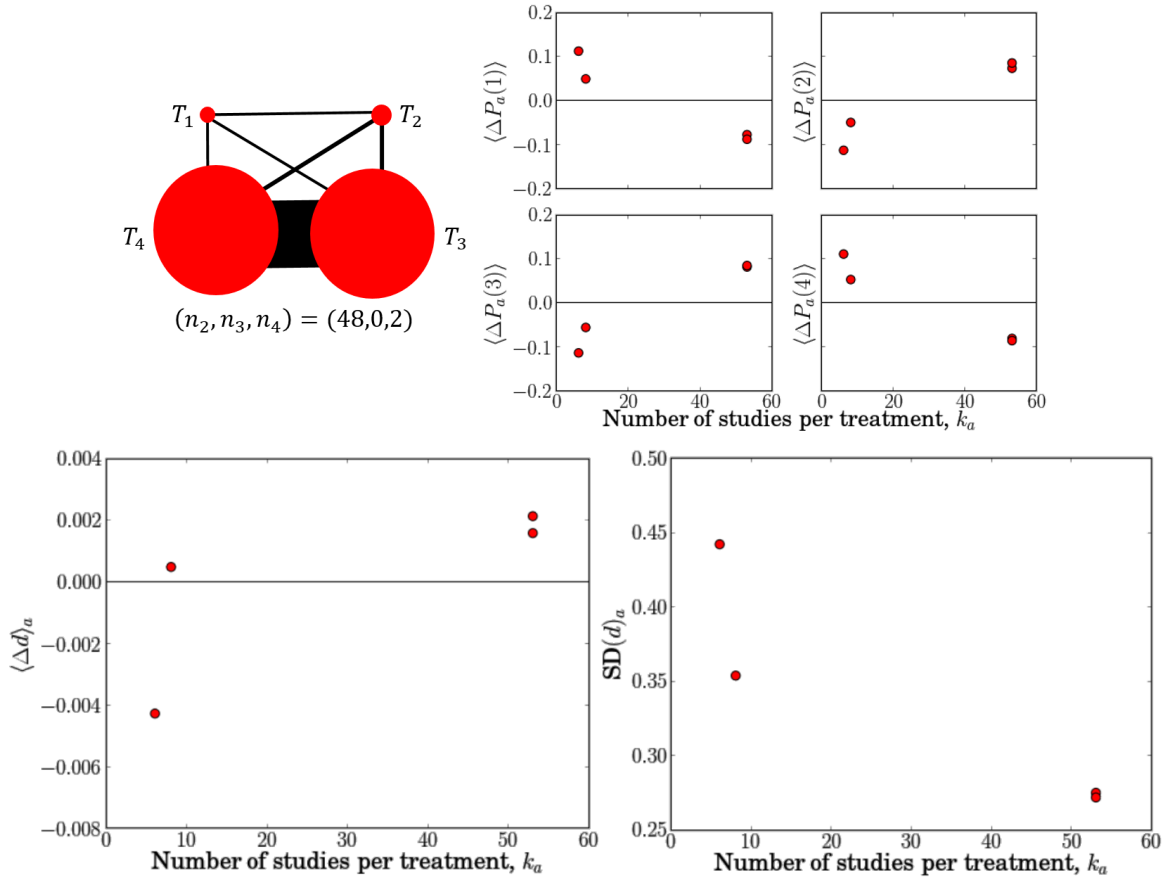


Figure 8.45: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 2, 2, 3, 3, 48)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (48, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.588333$.

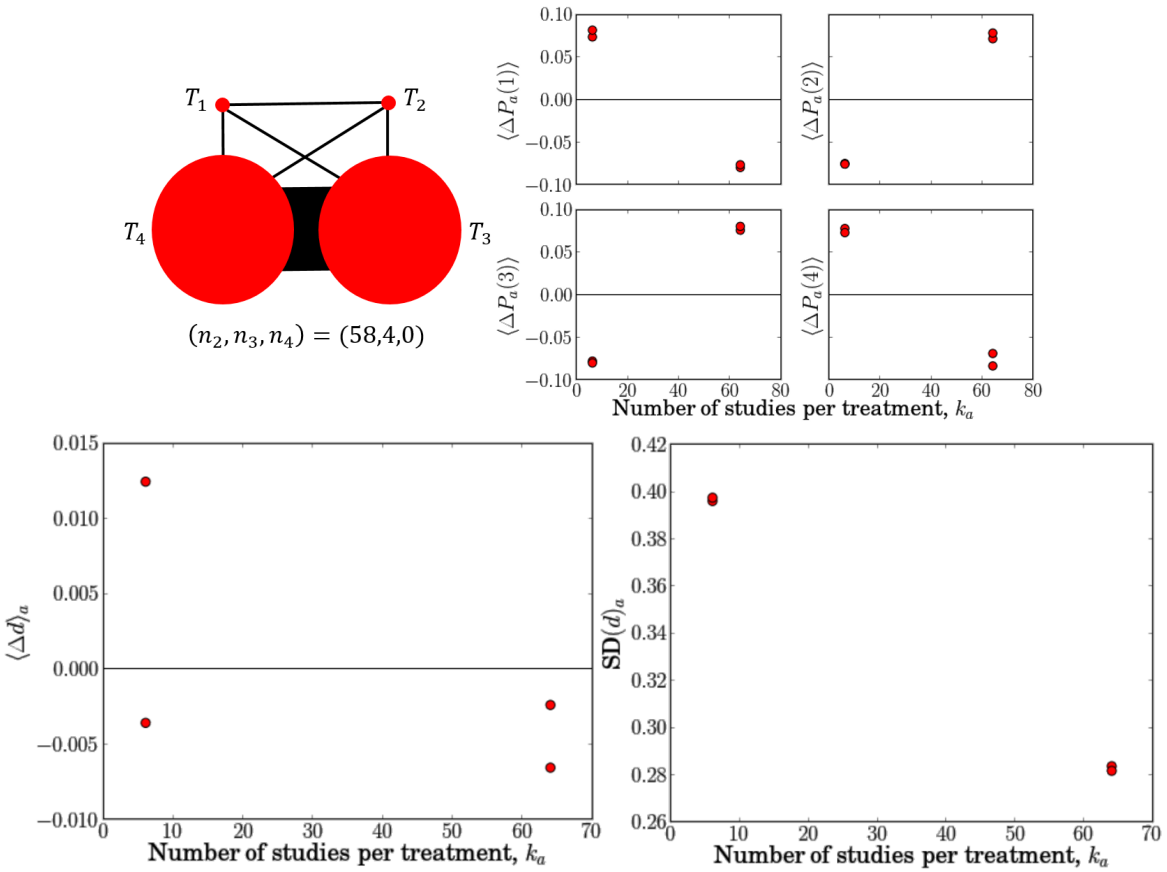


Figure 8.46: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 2, 2, 2, 2, 60)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (58, 4, 0)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.686531$.

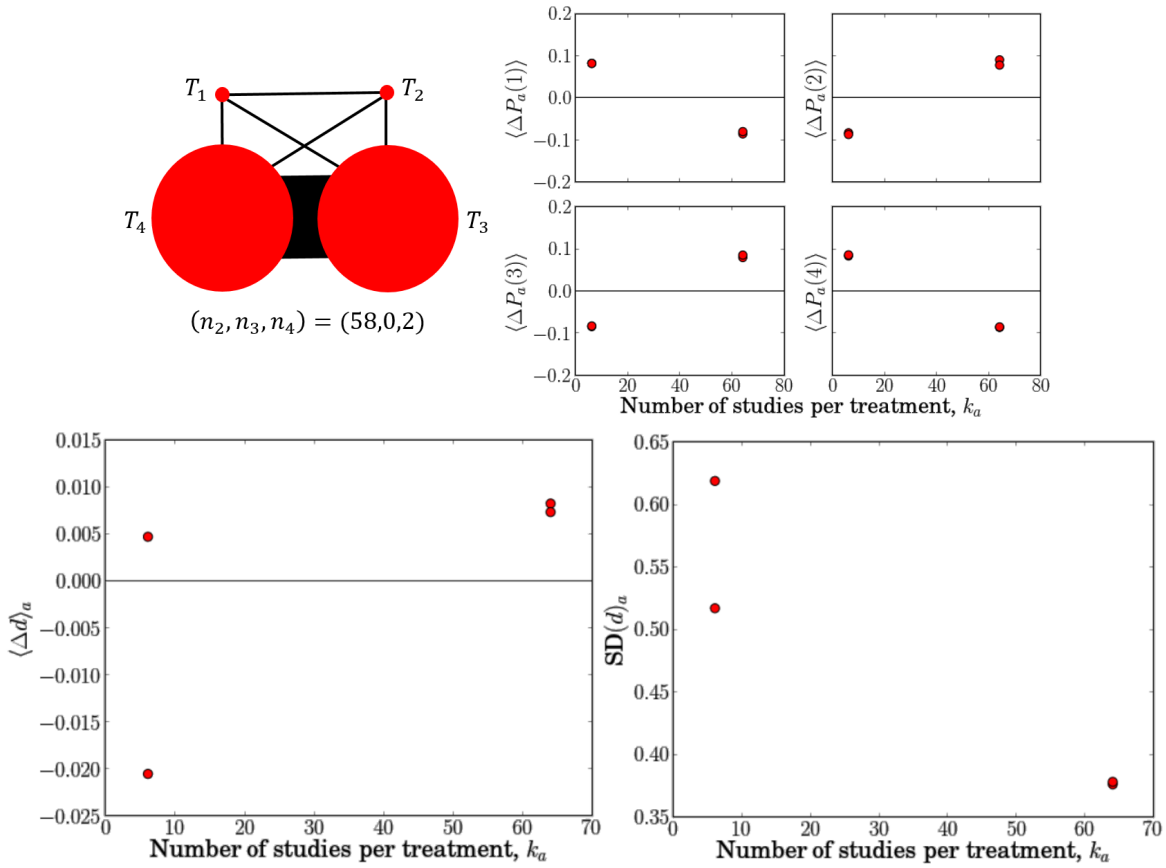


Figure 8.47: The number of studies per treatment versus bias of rank probabilities, and treatment-specific bias and standard deviation of treatment effects for a complete loop network with $\boldsymbol{K} = (2, 2, 2, 2, 2, 60)$. The number of two, three and four arm studies included is $(n_2, n_3, n_4) = (58, 0, 2)$. The irregularity of the network is $h^2/\bar{k}^2 = 0.686531$.

8.7 Comparing data-generating models

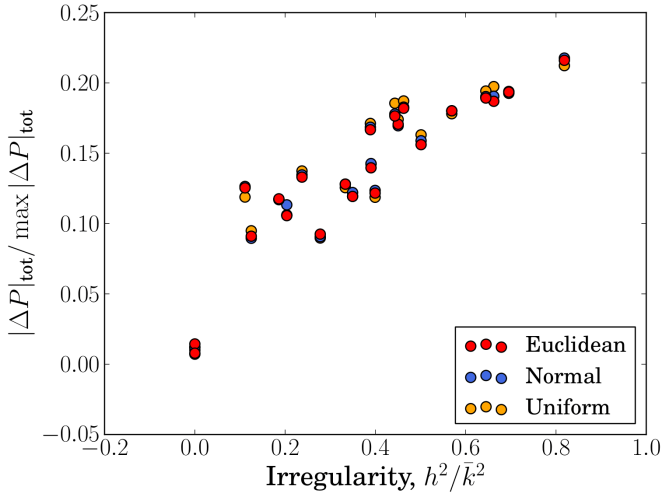


Figure 8.48: Comparing plots of network irregularity versus total rank probability bias for different data-generating models.

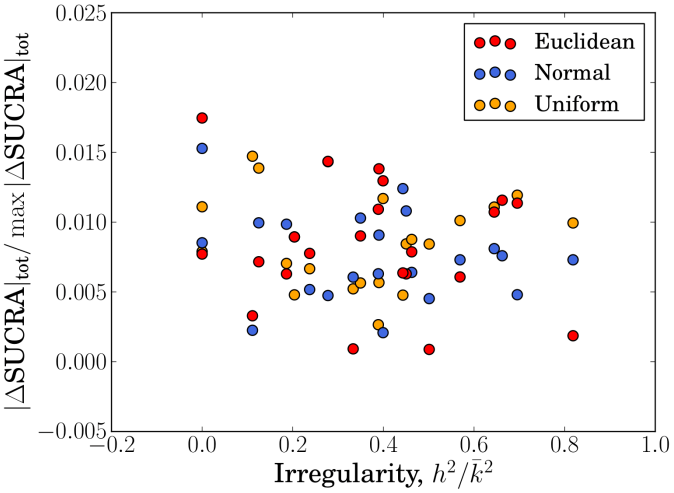


Figure 8.49: Comparing plots of network irregularity versus total SUCRA bias for different data-generating models.

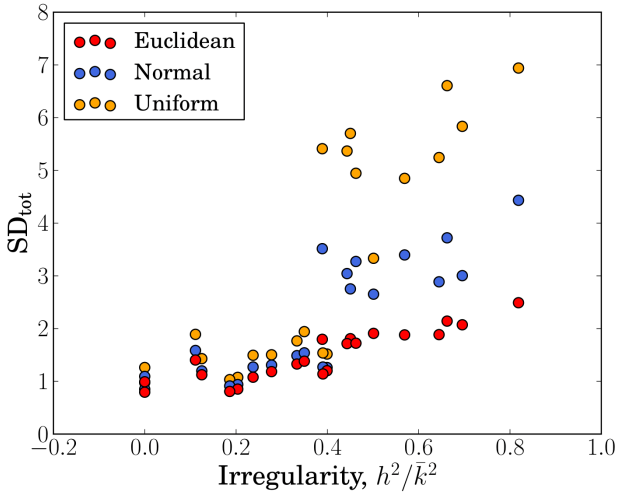


Figure 8.50: Comparing plots of network irregularity versus total standard deviation for different data-generating models. Standard deviation is lowest for the 'Euclidean' method as this DGM is the most restrictive in the variation of binomial probabilities sampled. Uniform has the greatest standard deviation because it is the least restrictive.

8.8 Bias of between-trial variance

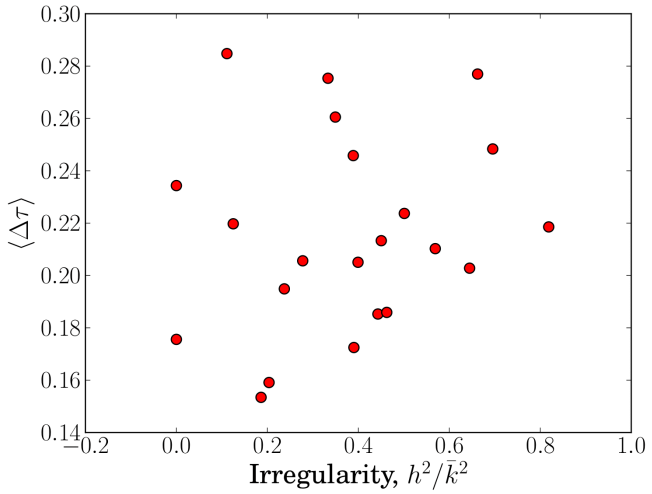


Figure 8.51: The effect of network irregularity on the accuracy of τ estimation. This is for networks with $\boldsymbol{d} = (0, 0, 0)$ and made up of exclusively 2-arm trials.

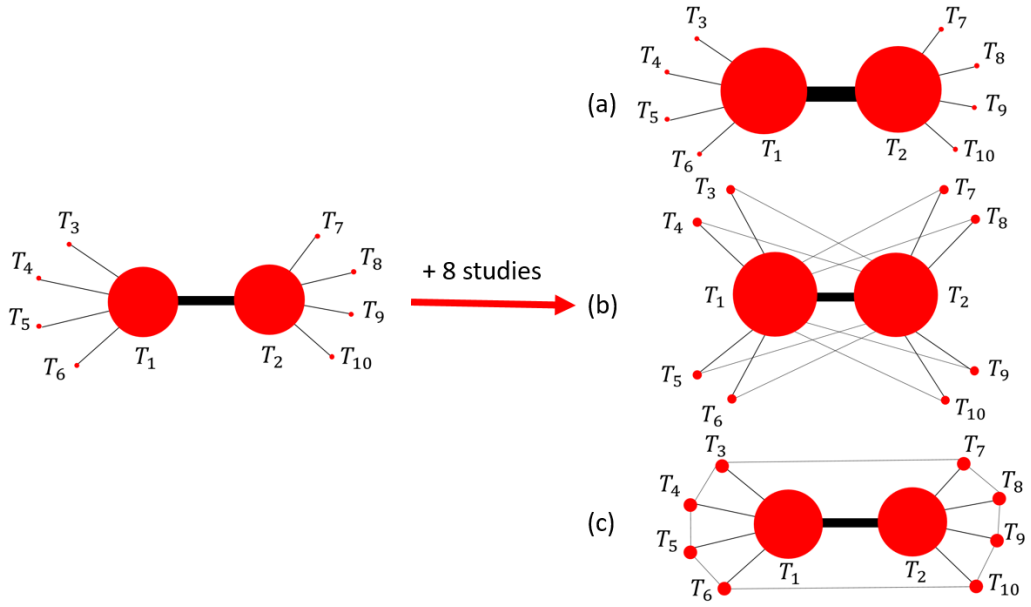


Figure 8.52: Network diagrams showing the networks simulated with $N = 10$ treatments. The original network has 12 $(T_1 - T_2)$ studies, and 1 study comparing the other connected treatments to T_1 or T_2. Networks (a), (b) and (c) have 8 studies added to them. In (a) all 8 are added to the $(T_1 - T_2)$ comparison, and in (b) and (c) each new connecting line represents one study. The results of these simulations can be found in Table 8.1.

8.9 Testing robustness

8.9.1 More than four treatments

To test if our results generalised to networks with more than four treatments we simulated four networks of ten treatments, each with different irregularity. The first network was the 'original' and is shown on the left in Figure 8.52. This network contains 20 trials in total; 12 trials comparing treatments T_1 and T_2, one trial connecting each of T_3, T_4, T_5 and T_6 to T_1 and one trial connecting each of T_7, T_8, T_9 and T_{10} to T_2. The other three networks (networks (a) to (c) in Figure 8.52) are made from the original network and eight extra studies. Network (a) has the highest irregularity and network (c) has the lowest irregularity. Table 8.1 summarises the results from these simulations and shows that high degree irregularity is associated with high rank probability bias and high SD_{tot}. Therefore we find that our results hold for networks with more than four treatments.

Table 8.1: Degree irregularity and quality of NMA outcome for the networks in Figure 8.52.

Network	M	$h^2/\bar{k}^2$	$\mathrm{SD}_{\mathrm{tot}}$	$\lvert\Delta P\rvert_{\mathrm{tot}}$
Original:	20	2.25	16.76	2.94
(a):	28	2.70	16.72	3.06
(b):	28	1.65	6.82	2.76
(c):	28	0.86	5.64	1.79

8.9.2 Unequal participants per arm

In our second robustness test, simulations were done for six networks where the number of participants per arm, rather than being assigned a fixed value, was randomly generated from a uniform distribution between 20 and 100. Figures 8.53 and 8.54 show that the results of these simulations are consistent with our previous findings.

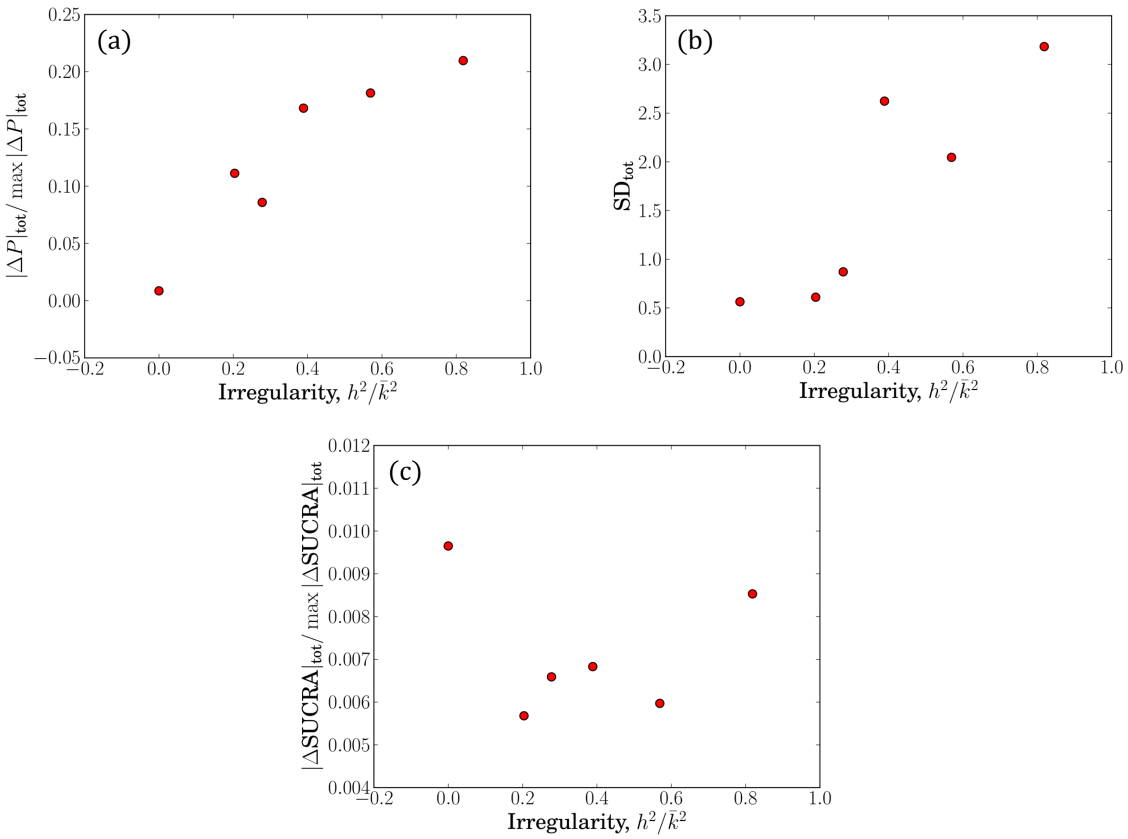


Figure 8.53: The effect of degree irregularity on a network's (a) total rank probability bias, (b) total standard deviation of treatment effect estimates, and (c) total SUCRA bias for networks with an unequal number of participants per arm. These networks have equally effective treatments and contain only 2-arm trials.

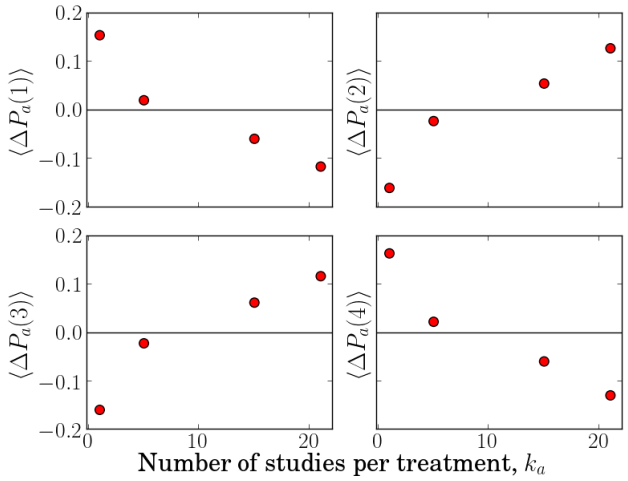


Figure 8.54: The effect of the number of studies per treatment on the bias on rank probabilities, $\Delta P_a(r)$, for $r = 1, 2, 3, 4$. These plots are for a star network with $\boldsymbol{K} = (1, 5, 15, 0, 0, 0)$ and for networks with an unequal number of participants per arm.