# Proposed Summary Measures for Ranking Treatments in Network Meta-Analysis

# PROPOSED SUMMARY MEASURES FOR RANKING TREATMENTS IN NETWORK META-ANALYSIS

BY

DANIELLE RICHER, M.A.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2014) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

| | |
|---|---|
| TITLE: | Proposed Summary Measures for Ranking Treatments in Network Meta-Analysis |
| AUTHOR: | Danielle Richer<br>M.Sc., (Statistics)<br>McMaster University, Canada |
| SUPERVISOR: | Dr. Joseph Beyene |
| NUMBER OF PAGES: | xii, 72 |

I dedicate this thesis to Margaret.

# Abstract

Network meta-analysis (NMA) is a process by which several treatments can be simultaneously compared for relative effectiveness. When conducted in a Bayesian framework, the probability that each treatment is ranked 1st, 2nd and so on can be calculated. A square matrix of these probabilities, referred to as the rank probability matrix, can be structured with rows representing treatments and columns representing ranks.

In this thesis, a simulation study was conducted to explore properties of five proposed rank probability matrix summary measures: determinant, Frobenius norm, trace, diagonal maximum and diagonal minimum. Each measure is standardized to approach 1 for absolute certainty. The goal of this simulation is to identify strengths and weaknesses of these measures for varying networks. The measures are applied to previously published NMA data for further investigation.

The simulation study and real data analysis revealed pros and cons of each summary measure; the Frobenius norm was found most effective. All summary measures yielded higher values with increases in symmetry, relative effect size and number of studies in the network.

If the rank probability matrix is used as the primary output of a network meta-analysis (as is often the case), a simple measure of the overall confidence in the rankings is beneficial. Future research will require exploration into the distributions of these measures.

# Acknowledgements

Sincerest thanks to my supervisor, Dr. Joseph Beyene, for his guidance and support throughout this process, my thesis examining committee members Prof. Román Viveros-Aguilera and Dr. Gregory Pond for their feedback, and the financial support provided by the Department of Mathematics and Statistics, McMaster University and NSERC (through Dr. Beyene).

I would also like to acknowledge the supportive and stimulating environment of Dr. Beyene's Statistics for Integrative Genomics and Methods Advancement (SIGMA) research group and would like to thank my fellow lab members and graduate students for their helpful feedback, encouragement and friendship throughout my graduate studies.

Finally, my best friend Shaun, who provided endless support through my successes and struggles, I will be forever grateful.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and Background

## 1.1 History

Science is a cumulative endeavour, and the process of synthesizing research should be done as scientifically as the individual contributions. Chalmers et al. (2002) in their review of research synthesis methods, state: "Although the need to synthesize research evidence has been recognized for well over two centuries, explicit methods for this form of research were not developed until the 20th century." The same article provides detailed accounts of the first cases of research synthesis: Karl Pearson in 1904 summarized the results of 11 studies looking at the effectiveness of vaccines against typhoid, while Joseph Goldberger published an analysis of bacteriuria in typhoid fever in 1907. Goldberger's research outlined four steps that are considered essential in standard practice of modern research synthesis: literature review, selection of appropriate studies, abstraction of data, and statistical analysis.

A systematic review is the collection and summary of similar research done in a comprehensive and transparent way. When quantitative methods are appropriate for combining results, the process is called a meta-analysis. Broadly speaking, a meta-analysis can be defined as a systematic literature review supported by statistical methods where the goal is to aggregate and contrast the findings from several related studies (Glass, 1976).

These processes have been applied to fields including ecology, education, psychology, sociology, and economics, but their importance is most notable in health care. Randomized control trials (RCTs) are the gold standard of health care research, and meta-analyses of RCTs are widely conducted. Systematic reviews and meta-analyses are often found at the top of hierarchies of evidence, and serve as an important link between research and practice, as these are the best-read publications (Chalmers, 1993).

## 1.2 Standards of Meta-Analysis

Despite growing popularity of systematic reviews and meta-analyses over the 20th century, their reliability was limited by the quality and frequency of reviews, which led to two important developments. First, standards were needed to ensure the quality of the reviews. To address the suboptimal reporting of meta-analyses, a guideline called the QUOROM Statement (Quality Of Reporting Of Meta-analyses) was developed in 1996, which focused on the reporting of meta-analyses of randomized controlled trials. This document was later revised to become the PRISMA

Statement, which stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses. The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram to help ensure the quality of research syntheses (Moher et al., 2009).

The second important development was the creation of the Cochrane Collaboration. The task of the Cochrane Collaboration is to prepare, maintain and disseminate systematic, up-to-date reviews of RCTs of health care research, and, when RCTs are not available, reviews of the most reliable evidence from other sources (Chalmers, 1993).

Systematic reviews and meta-analyses can assess whether or not there is conclusive evidence about a specific treatment. They are often used as a starting point for developing clinical practice guidelines, they can help to identify whether further research is needed, and they can highlight inconsistencies in research that require investigation.

## 1.3 Comparative Effectiveness Research

By design, meta-analyses amalgamate studies with similar participants, interventions and designs with the goal of comparing the effectiveness of two treatments. In reality, health care research is a much more complicated endeavour. Often more than two treatment options are available, and beyond consideration of efficacy are issues of harm and cost. The broad study of these issues is called comparative effectiveness research (CER). CER is defined as the generation and synthesis of evidence that

compares the benefits and harms of alternate methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care (Spine, 2010). Systematic reviews and meta-analyses are methods that fall under the umbrella of CER. Methods of CER focus on generating credible and relevant information as quickly and inexpensively as possible. The rigor demanded of meta-analyses is not present in all types of CER.

## 1.4 Network Meta-Analysis (NMA)

Network meta-analysis (NMA) can be viewed as a combination of the strict requirements of meta-analyses and the broader investigations of CER. A network meta-analysis uses a statistical process to synthesize the results of RCTs over a network of research that compares multiple treatments. Statisticians have only introduced methods of NMA in the final quarter of the 20th century, and those methods continue to be developed today. A NMA produces a set of estimates of the efficacy of any treatment in the network relative to any other (Dias et al., 2011c). The estimates produced through the NMA allow an optimal treatment to be determined, and a set of ranks to be assigned to remaining treatments.

## 1.5 Current State of NMA Research

Research applying the methods of NMA is limited, but this is changing quickly. In a review of published literature before 2007 by Salanti et al. (2008b), the authors

found 18 English-language articles had applied methods of NMA to treatment networks with at least four treatments, where treatments are broadly defined as any kind of intervention including no treatment or placebo. The earliest publication found was from 1999. The topics included treatments of epilepsy, rheumatoid arthritis, smoking cessation, prevention of fractures, hypertension, cancer, stroke and myocardial infarction. Networks of only three treatments were not considered as these have been studied extensively.

While it is clear that network meta-analyses are becoming more popular, the methods for conducting them continue to be developed. To ensure validity of findings and minimize error, these studies must be carefully designed and conducted. As the application of NMA statistical techniques becomes more widespread, the suitability and shortcomings of different methods must be investigated. Li et al. (2011) write, "Evaluating the performance of the different methods, through simulations and empirical studies, is critical before they become widely available." Several published papers detail statistical models for NMA, but very few authors have used simulations in their research (Song et al., 2012; Jonas et al., 2013; Thorlund and Mills, 2012; Mills et al., 2011). Further exploration of these methods to varying scenarios through simulation is an important next step in the development of NMA.

## 1.6 Aims

This thesis aims to explore one current model used widely in NMA - the Bayesian hierarchical model - to investigate the following two questions through simulation:

- What measures can be used to summarize the overall accuracy of the ranks yielded in a NMA?
- How will these measures be affected by changes to the network involving effect sizes, the network's geometry, and the number of studies contributing to the network?

These questions are motivated by applications of NMA to health research.

Investigators are not solely interested in accurately identifying the treatment most likely to be best. Determining a ranking of treatments with some credibility is also useful. The treatments that are second best, third best, and so on become important when access is limited due, for example, to cost or drug interactions. Ranking many possible alternatives may be particularly important for policy makers from the developing world, where the best available treatment may not be affordable (Salanti et al., 2008a). For this reason, it is helpful to quantify the overall accuracy of the ranks generated from the model. Further, networks are not fixed, but change over time with updated studies and new treatments, so investigating how the accuracy of ranks is affected by changes to the network is also important.

# Chapter 2

# Methods

## 2.1 Meta-Analysis

Meta-analysis is a much larger endeavour than the statistical calculations required. While this section will discuss the statistical methods for meta-analysis, in order to yield meaningful results, the entire research process must be done to the standards discussed in the PRISMA statement.

Meta-analysis can be broadly defined as the quantitative review and synthesis of the results of related but independent studies (Normand, 1999). This section provides an introduction to the methods used in meta-analysis and how they were extended and developed to create models for network meta-analysis.

### 2.1.1 Goals of Meta-Analysis

Two important objectives can be met through meta-analysis. First, parameter estimates can be made with more precision and statistical power. Second, an assessment can be made of the variability between studies, and study characteristics associated with this variability can be identified. Heterogeneity is defined as the existence of differing outcomes in studies that is not attributable to chance (Normand, 1999).

### 2.1.2 Notations

Traditional meta-analysis is used in the synthesis of studies comparing two treatments. To begin, the items of comparison (drugs, surgical techniques, rehab programs) will be labelled A and B, with treatment A considered the baseline treatment. In practice, this might be a placebo or a standard treatment. To investigate the relative effects of treatments A and B, all studies directly comparing these two treatments are collected and reviewed. Those studies deemed to be adequately randomized and similar, up to the standards of the investigators, have their results combined through a weighted process to yield a (hopefully) more accurate estimate of the relative effects of treatments A and B. Let the number of studies included that compare A and B be labelled $M$. Studies 1 through $M$ will have parameter estimates of the difference between treatments, $y_i$, where the subscript denotes the study from which that estimate came. Differences between treatments can be measured in absolute or relative terms depending on the available data. Relative risks and odds ratios are used to measure relative effects for binary data, taken on the log scale. In the case of continuous variables, absolute effects are taken typically using a mean

difference, but might also be a standardized mean difference. The odds ratio will be the outcome measure of interest considered in the simulations found in Chapter 3.

### 2.1.3 Models

A brief summary of meta-analysis models is provided here, using Normand (1999) and Fleiss (1993) as references.

Beginning with $M$ separate studies with random allocation to two treatments of interest, each study effect $y_i$, with $i = 1 \dots M$ is assumed to be independent with within-study variance $\sigma_i^2$.

There are two commonly-used models in meta-analysis which differ in their assumptions: fixed effects and random effects.

The fixed-effects model assumes that each study is yielding an estimate of the same underlying true effect size, and the different results between studies are only attributable to chance. In other words, fixed effects assumes homogeneity of studies.

However, heterogeneity will often be present between the individual studies making up the meta-analysis. Both participant demographics and study design will vary over the collection, if minimally, which leads to different observed study effects. The random effects model addresses study differences by assuming that each observed study effect size is drawn from a study-specific distribution. To determine which model is appropriate, it is possible to test for the presence of heterogeneity. It can be done using Cochran's Q-test, and quantifying the amount of heterogeneity present in a meta-analysis can be done using the $I^2$ statistic (Higgins and Thompson, 2002).

Detailed models and justifications for their use are included below.

### Random Effects Model

From Normand (1999), "It is almost always reasonable to believe that there is some between-study variation and few reasons to believe it is zero." The existence of heterogeneity in meta-analysis is generally assumed, and the standard model used under the assumption of the presence of unexplained heterogeneity is the random effects (RE) model. In this model, each study included in the meta-analysis is considered to be taken from a hypothetical set of all possible studies. As such, conclusions of the meta-analysis allow for inference on the true mean treatment effect over the population of interest.

A random effects model assumes that heterogeneity among study effects is purely random, and that a single true effect size exists around which the true individual study effects are distributed normally.

Then, each observed study effect size, $y_i$ corresponds to a true study effect size, $\theta_i$, through the equation:

$$y_i = \theta_i + e_i \quad i = 1, \ldots, M$$

where

$$e_i \sim N(0, \sigma_i^2).$$

It is then assumed that the study-specific means are distributed about a true population mean, $\mu$ with:

$$\theta_i = \mu + u_i \quad i = 1, \ldots, M$$

where

$$u_i \sim N(0, \tau^2)$$

$\tau^2$ is the between-study variance. The aim of the meta-analysis is to estimate $\mu$ and the true amount of heterogeneity between individual study effect sizes, $\tau^2$.

The estimate of $\mu$ is found through a weighted average, because contributing studies may vary greatly in terms of size and within-study variability. With the weights of study-level effects given by $w_i = 1/(\sigma_i^2 + \tau^2)$, the estimate is calculated:

$$\hat{\mu} = \frac{\Sigma w_i y_i}{\Sigma w_i}$$

.

The standard error is $SE(\hat{\mu}) = 1/\sqrt{\Sigma w_i}$, which can be used to conduct hypothesis tests or create confidence intervals. The between-study variance is not known in practice and can be estimated using the method of moments, maximum likelihood, or reduced maximum likelihood.

### Fixed Effects Model

A fixed effect analysis assumes that each study generates an estimate of the same parameter $\mu$, subject only to sampling error. The fixed effect model makes the assumption of homogeneity amongst studies and sets $\tau^2 = 0$.

$$y_i = \mu + e_i$$

where

$$e_i \sim N(0, \sigma_i^2)$$

In the fixed effects model, each study estimates the true effect size. It is once again estimated through a weighted average of the individual study effects, though the weights are calculated differently:

$$\hat{\mu} = \frac{\Sigma w_i y_i}{\Sigma w_i}$$

where $w_i = 1/\sigma_i^2$.

### 2.1.4 Heterogeneity and Meta-Regression

Modeling heterogeneity using the random-effects model is not appropriate if there is a covariate responsible for the differences between effect sizes. Researchers may suppose that the differences in study-level effect size are related to the demographic characteristics of the participants or the study design. When heterogeneity is detected and investigated, there is the potential for determining study-level covariate and treatment interactions. A meta-regression can be conducted using aggregate covariate measures to determine these interactions. A detailed discussion of meta-regression can be found in van Houwelingen et al. (2002).

## 2.2 Network Meta-Analysis

There are several excellent resources for understanding the basics of network meta-analysis. Fundamental papers including Lu and Ades (2004), Bucher et al. (1997), Lu and Ades (2006) and Lumley (2002) are valuable references. Three series of publications are also very thorough resources, the ISPOR documents (Jansen et al., 2011; Hoaglin et al., 2011), the NICE documents (Dias et al., 2011b,c,a,d), and a special issue in the journal *Research Synthesis Methods* (Salanti and Schmid, 2012).

The following summary of network meta-analysis is drawn from these references.

### 2.2.1 Introduction

A network meta-analysis is a framework for quantitatively synthesizing research of multiple interventions. It is alternatively called a multiple treatment meta-analysis (MTM) or a mixed treatment comparison (MTC). Rather than pooling information on trials comparing treatments A and B, network meta-analysis combines data from randomised comparisons, A vs B, A vs C, A vs D, B vs D, and so on, to deliver a set of pairwise relative effects while respecting the randomisation in the evidence.

The network meta-analysis process requires the combination of direct evidence and indirect evidence. Direct evidence between treatments A and B comes from combining the results of studies which compare these two treatments. Indirect evidence is drawn from combining other studies in the network. If two treatments, A and B, have a common comparator C, then A and B can be compared indirectly by combining the A versus C and B versus C studies as follows:

$$d_{AB} = d_{AC} - d_{BC}$$

with variance $Var(AB) = Var(AC) + Var(BC)$. If there are multiple studies that compare A versus C and B versus C, then the combined evidence can be used, $\mu^I_{AB} = \mu^D_{AC} - \mu^D_{BC}$, where $I$ denotes indirect evidence and $D$ denotes direct evidence. When both direct and indirect evidence are available, they can be combined into a mixed effect size, $\mu^M_{AB}$, using a weighted average.

### 2.2.2 Assumptions

When conducting a network meta-analysis, three assumptions are made. The first, carried over from standard meta-analysis, is the assumption of homogeneity for a fixed-effects model or purely random heterogeneity for a random-effects model. The second, transitivity, assumes that indirect comparison is a valid way of estimating the difference between two treatments that have not been directly compared. It is alternatively referred to as the similarity assumption. The third assumption, consistency, requires that the direct and indirect estimates are in agreement. Consistency can be thought of as an extension of transitivity over a closed loop of evidence.

### 2.2.3 The Consistency NMA Model

This section will present a consistency model for binomial data where the outcome of interest is the odds ratio. Suppose $M$ studies make two-arm comparisons with any of the $T$ treatments included in the network. We define $r_{ik}$ as the number

of events out of the total number of patients in each arm, $n_{ik}$, for arm $k$ of trial $i$, we assume that the data generation process follows a Binomial likelihood $r_{ik} \sim Binomial(p_{ik}, n_{ik})$ where $p_{ik}$ represents the probability of an event in arm $k$ of trial $i$, $k = 1, 2; i = 1, \ldots, M$. By restricting all RCTs to two-arm trials, there is only one effect estimated in each study, so the consideration of within-trial correlated estimates is unnecessary.

A set of $T - 1$ basic parameters for relative effects must be selected. Typically, these basic parameters are taken to be the relative effects between a specified baseline treatment and each other treatment in the network. Parameters estimating the treatment effects amongst the non-baseline treatments are called functional parameters, and need not be estimated as they can be represented as linear functions of the basic parameters. Inherent to this structure is the assumption of consistency.

A logit link function is used to map the probabilities to a continuous measure centered at zero. The models can be re-written with the treatment effect defined from a study-specific baseline treatment.

For clarity, a meta-analysis is presented first. For random effects,

$$logit(p_{i1}) = \mu_i$$

$$logit(p_{i2}) = \mu_i + \delta_{i12}$$

Each $\mu_i$ is a trial-specific baseline, representing the log-odds of the outcome in the baseline treatment while $\delta_{i12}$ are the trial-specific log-odds ratios of success on the treatment group 2 compared to group 1, with $\delta_{i12} \sim N(d_{12}, \tau^2)$. For fixed effects,

$$logit(p_{i2}) = \mu_i + d_{12}$$

,

which is equivalent to setting $\tau^2$ to zero.

Note that the subscripts indicating the treatments being compared are unnecessary in the case of only two arms; however, similar notation will be used for network meta-analysis where multiple treatments are considered. For this reason, the subscripts are included. For clarity, some additional notation is introduced here. Let $X$ and $Y$ represent variable treatments and $A, B, C, D$ will represent fixed treatments, with the convention that the first letter alphabetically in each trial serves as the baseline treatment. The subscript $b$ indicates a study-specific baseline.

Then, a fixed-effects model can be written as follows:

$$logit(p_{ik}) = \mu_{ib}, \text{ for } k = b$$

$$logit(p_{ik}) = \mu_{ib} + d_{bk}, \text{ for } k \text{ after } b$$

For a random-effects model, $d_{bk}$ is replaced with $\delta_{ibk}$ and yields:

$$logit(p_{ik}) = \mu_{ib}, \text{ for } k = b$$

$$logit(p_{ik}) = \mu_{ib} + \delta_{ibk}, \text{for } k \text{ after } b$$

where $\delta_{ibk} \sim N(d_{bk}, \tau^2_{bk})$ where the variance term must be estimated.

It is common to assume a constant between-study variance, $\tau^2$. The assumption

of equal variances implies that the correlation between any two treatment contrasts in a multi-arm trial is 0.5.

### 2.2.4 The Bayesian Framework

A Bayesian framework means that all parameters are considered random variables. These parameters are assumed to have a distribution, called a prior distribution, which is updated by the data available to yield a posterior distribution. The prior assigned to parameters must be specified with hyper-parameters. When the hyper-parameters are considered random variables with their own prior distributions, the result is a Bayesian hierarchical model.

The process of combining the prior distributions and the study-level data to generate posterior distributions is implemented with the use of a Markov Chain Monte Carlo (MCMC) method. MCMC methods make it possible to simulate the entire joint posterior distribution of the unknown parameters.

MCMC methods generate pseudo-random draws from probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one. As we proceed along the sequence, provided that certain regularity conditions are met, the distributions of the elements stabilize to a common distribution known as the stationary distribution. In MCMC, one constructs a Markov chain whose stationary distribution is a distribution of interest. The Markov chain begins at an arbitrary point and runs for a long time. After deleting several inital values (the burn-in), what remains is a sequence of dependent random variables all with the desired marginal

distribution (Schafer, 2010).

Care must be taken in checking convergence of these chains. Posterior distributions should be examined visually for spikes and unwanted peculiarities, and both the initial burn-in and posterior samples should be conservatively large. Over-dispersed starting values should be chosen for a number of independent chains so that convergence can be assessed. Thinning, keeping every $n$th value after the burn-in period, is a way of reducing the dependent nature of the chains.

### 2.2.5 The Bayesian Hierarchical Model

Specifying a Bayesian hierarchical model involves choosing prior distributions for several parameters. To model a network meta-analysis in a Bayesian framework, prior distributions must be provided to the basic parameters $d_{XY}$, the baseline treatment effects in each study $\mu_{ib}$, and the between-study variance $\tau^2$.

It is recommended that the following uninformative priors are used (Salanti et al., 2008a):

$$\mu_{ib} \sim N(0, 100000)$$

$$d_{XY} \sim N(0, 100000)$$

$$\tau \sim Uniform(0, 2)$$

The upper limit of 2 in the Uniform distribution represents a huge range of trial-specific treatment effects. Once specified, the study results are combined with these prior distributions to form joint posterior distributions for the parameters of interest.

Often, the choice to work in a Bayesian framework is made in order to bring prior

information to the modeling process, however this is not of interest in the case of NMA. Rather, the Bayesian framework is appropriate for NMA because it allows probabilistic conclusions to be made. In particular, the rank probability matrix would not be possible if a Bayesian framework were not employed.

### 2.2.6 Further Readings on Bayesian NMA

A discussion of how to conduct meta-analyses with multi-arm trials in a Bayesian framework is detailed by Lu and Ades (2006). In particular, within-study correlations between the parameters of interest must be taken into account.

Consistency of treatment effects must hold for the the consistency model described here. However, a common criticism of network meta-analysis is the difficulty assessing consistency. Methods for investigating consistency and alternative models are presented in Dias et al. (2010), Lu and Ades (2006), and Higgins et al. (2003).

### 2.2.7 Software for Network Meta-Analysis

As network meta-analysis becomes a well-known method in health research and other applications, software packages will be developed alongside the statistical models. Currently, one R package, GeMTC (generating multiple treatment comparisons) has been created to run network meta-analysis in a Bayesian framework (van Valkenhoef and Kuiper, 2013). This package allows for the simple implementation of a simulation study to investigate rank probabilities.

### 2.2.8 The Rank Probability Matrix

The Bayesian hierarchical model generates a set of estimates for the basic and functional parameters, $d_{XY}$. In addition to these effect parameters, the posterior probabilities that each treatment is best, second best, and so on are also of interest and can be calculated through the MCMC framework.

For each MCMC iteration, the treatments are ranked by their effect relative to an arbitrary baseline. A frequency table is constructed from these rankings and normalized by the number of iterations to provide rank probabilities.

These probabilities are presented in a matrix of size $T \times T$, where $T$ is the number of treatments in the network, displaying the probability that each treatment holds each of the 1st through $T$th possible rankings. Rows represent each of the treatments, while columns represent the ranks in descending order. A rank probability matrix with absolute certainty of the true ranks, with treatments ordered from most to least effective, will be the $I_T$ matrix. All rank probability matrices are doubly stochastic, which means all rows and columns sum to one with all entries falling between 0 and 1. For ease of use, the matrix can be reordered so that the treatments are ordered in rows from most to least effective. In a consistency model, the relative effects between any two treatments will support a common ordering of the treatments. This is the order used to order the rows of the rank probability matrix.

As this reordered matrix will be used frequently in future sections, it will be referred to as the ordered rank probability matrix. In the simulation setting, the order is fixed by the pre-determined treatment ranks. In the real data applications, the order is determined by ranking treatments in accordance with the estimates of

the pairwise differences between treatments yielded from the modelling process. An example of a $3 \times 3$ rank probability matrix that has been ordered is included in Table 2.1.

| | 1st | 2nd | 3rd |
|---|---|---|---|
| Treatment A | 0.95 | 0.04 | 0.01 |
| Treatment B | 0.03 | 0.75 | 0.22 |
| Treatment C | 0.02 | 0.21 | 0.77 |

Table 2.1: Example of an Ordered Rank Probability Matrix

## 2.3 Proposed Measures for Summarizing Rank Probability Matrices

One of the current shortfalls of network meta-analysis is the lack of an interpretable and simple measure to summarize the results (Salanti, 2012). With this concern in mind, five measures for summarizing rank probability matrices are considered. *Standardized determinant, standardized Frobenius norm, standardized trace, diagonal maximum* and *diagonal minimum* will be investigated through a simulation study. Each of these summary measures has been considered for simplicity of calculation and for the ease of interpretation. Each is calculated from the ordered rank probability matrix. Measures were selected that would yield maximal values from identity matrices - implying complete confidence in the ranks. Each measure has been standardized so that its optimal value is one.

Given an $n \times n$ matrix, $A$ with the entry in row $i$ column $j$ denoted $a_{ij}$, the summary measures are calculated as follows:

$$trace(A) = \sum_{i}^{n} a_{ii}$$

$$norm_F(A) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}$$

The determinant of a $3 \times 3$ matrix is calculated:

$$det(A) = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) - (a_{13}a_{22}a_{31} + a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32})$$

The formula for an $n \times n$ matrix is similar and can be found in any linear algebra textbook.

The remaining two summary measures, the diagonal maximum and the diagonal minimum, are single entries in the matrix and need not be calculated. Interpretations of each measure are included below.

**Standardized Absolute Determinant (det)** The determinant is calculated using all values from the matrix, and has a geometric interpretation representing a volume in the three-dimensional case. This measure will be larger for matrices that more closely resemble the identity matrix; maximum volume is obtained from the unit cube formed by the identity matrix. It must be standardized for comparison across different matrix dimensions, thus the absolute value and $T$th root is taken. Interchanging rows or columns of a matrix changes the sign of the determinant. By taking the absolute value, the order of the rows in the

rank probability matrix is inconsequential.

**Standardized Frobenius Norm (norm)** Matrix norms are used to quantify the size of a matrix and can be calculated in a variety of ways. Many norm calculations rely on row or column sums, which are irrelevant for doubly stochastic matrices. The Frobenius Norm is commonly used and increases as doubly-stochastic matrices approach the identity. It is standardized for comparison across different matrix dimensions by dividing the result by a factor of $\sqrt{T}$.

**Standardized Trace (trace)** The diagonal of the ordered rank probability matrix captures the total amount of correctly assigned probability. Thus, the trace is a natural choice for a summary measurement. It is standardized for comparison across different matrix dimensions by dividing the result by a factor of $T$.

**Diagonal Maximum (max)** The single largest value in the ordered rank probability matrix diagonal. This represents an upper bound on the amount of certainty there is in the matrix. As it uses only one entry, it is expected to be highly variable.

**Diagonal Minimum (min)** The smallest value in the ordered rank probability matrix diagonal. This represents a lower bound on the amount of certainty there is in the matrix. It can be thought of as an upper bound on the error of all probability ranks.

The goal of the summary measures is to assess the overall confidence in the rankings. Higher values for these summary measures reflect higher entries in the

rank probability matrix, which indicate more certainty of the proposed rankings. The trace, diagonal maximum and diagonal minimum measure values on the rank probability matrix diagonal, thus requiring the rank probability matrix to be ordered. The determinant and norm do not require this ordering.

Although the diagonal maximum and diagonal minimum are likely over-simplifying the information in the rank probability matrix, they have been included for inspection. If either of these measures behaves very similarly to one of the calculated measures, it might be unnecessary to perform any calculation at all and to use one of these values as a representation of overall confidence in the network.

The ideal summary measure will provide a tool to compare the overall confidence captured in different networks. As networks change in ways that should lead to less confidence in the ranks, (for example, increasing the number of treatments, reducing the number of studies, reducing the effect size), the summary measure should reflect these changes. Ideally, the changes in the summary measure will span much of the 0 to 1 range, so that differences between networks can be clearly seen. When calculated many times over the same network structure, the variance should be minimal. Through the simulation study and the real data analysis, features of the proposed summary measures should be revealed. Additional considerations that arise will also be discussed.

# Chapter 3

# Simulations

The goal of this simulation was to explore the two key questions posed in the introduction of this thesis. First, what summary measures might be used to describe the confidence in a set of ranks? That is, given an ordered rank probability matrix, what measurements can be taken to assess the overall confidence of the ranks given in that matrix? Five summary measures have been proposed and will be considered. Second, how are these summary measures affected by aspects of the network? Specifically, how will the summary measures change in response to altering the number of studies, the effect size and the geometry of the network? The design of this simulation and the results are presented in the following sections.

Before carrying out the simulation, several patterns were anticipated within and between these measures. For example, the diagonal maximum, trace and minimum will always be found in decreasing order, regardless of the original network. Other patterns make intuitive sense; increasing the effect size or number of studies should

result in stronger evidence to support the correct rankings, and produce higher summary measures. Other patterns are less intuitive, such as the changes in summary measures over different network geometries.

In order to confirm the patterns involving effect size and number of studies, and to explore the results under different network geometries, a total of 216 simulation scenarios were considered (3 effect sizes $\times$ 4 studies per comparison $\times$ 18 geometries).

## 3.1 Simulation Design

When designing this simulation, an effort was made to maintain very simple network structures. First, this simplicity allows attention to be focused on the summary measures and to ensure that they behave as expected in the most basic cases. Second, the use of the most simple networks made the implementation straight-forward through use of the GeMTC R package. Specifically, the decisions to use a dichotomous variable (thus effect size measured using odds ratio), to generate all studies without introducing any heterogeneity, and to model the network using a consistency model simplified the process. This simulation design, though unrealistic for comparison to real studies, is thought to be merely a starting point in the exploration of summary measures for rank probability matrices.

### 3.1.1 Parameters Considered

In order to explore the way in which these summary measures behave under different networks, several parameters were varied in the simulation study.

**Effect Size (ES):** Given that the least effective treatment would have a fixed probability of effectiveness of 0.1, three effect sizes (later called small, medium and large) were considered based on odds ratios of 1.2, 1.6 and 2. These effect sizes were fixed between successive treatments. Although odds ratios larger than 2 may be found in practice, preliminary simulations revealed that ordered rank probability matrices differ minimally, if at all, from identity matrices given odds ratios above 2.

**Network Geometry:** Four network geometries were considered in the simulation, one network of three studies and three networks of four studies.

**Triangle (TRI):** The only three-treatment network considered is a complete triangle. This shape was selected to represent the most simple and symmetric case.

**Complete Square (COMP):** The only fully symmetric case of a four-treatment network is a complete network. While rare in applications, the four-treatment symmetric case provides valuable insight into the role of symmetry.

**Star:** This four-treatment network only contains comparisons between each treatment and a single common comparator. All four treatments are placed in the centre of the star to create four permutations of this geometry.

**Dangle:** The final four-treatment network contains a complete triangle between three treatments and a fourth treatment compared to any one of

the triangle vertices. Rotating the treatments into different places in the network yields 12 permutations of this geometry.

**Number of Studies per Comparison (SpC):** The number of studies per comparison was fixed at 1, 2, 5 or 10. These values indicate the number of RCT's simulated for each pair of treatments connected in the given network geometry. Although imbalance in the number of studies per comparison is standard in real applications, only balanced cases have been considered here.

**Characteristics of Contributing Studies:** For simplicity, the contributing studies were not varied at all. Each contributing study was generated to be a two-arm study with 100 participants per arm.

A summary of the parameters is available in Table 3.1. A diagram of the different network geometries under consideration is provided in Figure 3.1. The black arrows represent basic parameters, and all other direct comparisons in the network are noted with grey lines.

### 3.1.2 Simulation Process

Simulation settings were chosen to allow one parameter to vary while others remained constant, providing an opportunity to investigate the role of each parameter in the contribution to rank probabilities and summary measures. For each fixed set of parameters, the simulation process required three steps: simulating the contributing studies, formatting the data, and applying the Bayesian hierarchical model to the network.

Table 3.1: Parameter Descriptions

| Parameters | Values |
|---|---|
| Effect Size measured in Odds Ratio | Large ($OR = 2$)<br>Medium ($OR = 1.6$)<br>Small ($OR = 1.2$) |
| Studies per Comparison | $10, 5, 2, 1$ |
| Network Geometry | Triangle<br>Complete Square<br>Star (4 permutations)<br>Dangle (12 permutations) |
| Study size | 100 per arm |
| Treatments per study | 2 |

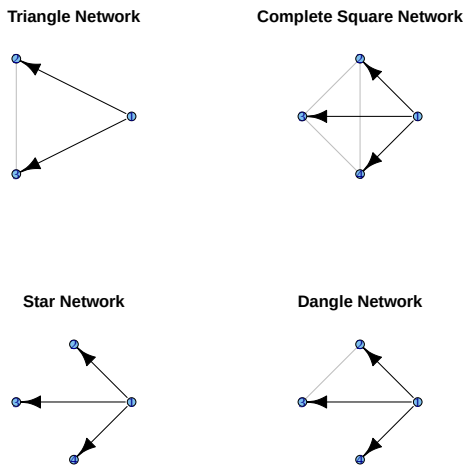


Figure 3.1: Simulated Network Geometries.

First, 60000 3- or 4-arm studies were simulated where the number of events out of 100 participants was randomly determined for each of the A through C or D treatments, based on a probability calculated inversely from the odds ratio of interest. Pairs of arms were taken from the available arms to match the direct comparisons in the geometry and the number of studies per comparison necessary for the simulation setting. For example, in the star network with one study per comparison, from the $60000 \times 4$ matrix of data generated, outcomes from treatments A and B would be selected from the first row, outcomes from treatments A and C would be selected from the second row, and outcomes A and D would be selected from the third row. Since the simulation was repeated 1000 times for each scenario, this process would be repeated beginning with the next unused row. Depending on the complexity of the geometry, anywhere from 3000 to 60000 of the randomly generated multi-arm studies were employed. One of the interpretations of the consistency assumption is that arms can be thought of as missing at random. Since the data is all generated randomly, it is in line with the assumptions of NMA to generate multi-arm studies and sample treatment arms from them in the simulation process.

The second step of the simulation process required reformatting the groups of generated studies for input into the GeMTC R package (van Valkenhoef and Kuiper, 2013). This involved placing the data in a long-form structure and ensuring that entries were classified correctly.

The third step was the implementation of the GeMTC package. It provides options for model type, the number of chains to be used in the MCMC process, variance scaling factor for the starting values, link and likelihood to be used, and

random or fixed effects model options. For this simulation, a consistency model with random effects was selected with binomial and logit used for the likelihood and link, respectively. The recommended default values, 2.5 and 4, were used for the variance scaling factor and number of chains, respectively. The model specifies which parameters are basic and the plot function displays these parameters. Samples are generated based on these specifications, and a Bayesian hierarchical model is built that takes into account all direct and indirect evidence to generate an estimate of the effect of all included treatments and a rank probability matrix. The tuning phase, burn-in, thinning and length of the MCMC chains can be specified. This simulation used the default values of 5000 tuning, 20000 burn-in and chain length of 20000, with a thinning factor of 10.

For each simulation setting, the 1000 rank probability matrices were stored in a $1000 \times 4 \times 4$ array. Summary measures from these arrays were taken in order to answer the questions of interest.

## 3.2 Results

Included in this section are tables and graphs that illustrate properties of the five summary measures when calculated from varying networks. The properties of these summary measures, the changes that take place by varying parameters, and the effect of introducing a new treatment are all considered.

To describe the simulation results clearly and succinctly, the following terms will be used repeatedly:

**Effect Size** The effect size refers to the change in treatment success expected if one treatment is substituted for another. In this simulation, the effect size refers to the three fixed odds ratios of 1.2, 1.6 and 2.0 (small, medium and large) that were used to calculate success probabilities and generate study results.

**Studies per Comparison** Within a network, the number of individual studies contributing to each direct comparison is fixed. This single value refers to the number of studies comparing all pair-wise direct comparisons, as this value is balanced in all simulation scenarios.

**Geometry** In this simulation, the influence of geometry is limited to the four network geometries considered.

### 3.2.1 Properties of Summary Measures

To display some of the trends in summary measures under various parameter changes, a table with selected results is provided in Table 3.2. The summary measures reported are based on the average values over the simulation size of 1000, and the accompanying standard deviation addresses the variability in these 1000 values.

There are several patterns to observe in this table. Both decreases in effect size and in studies per comparison lead to smaller values for all five summary measures. Modifying the network from a complete square to a star geometry has the same result. This change in geometry lowers the total number of studies in the network and introduces asymmetry in the network. It is likely that both of these factors contribute to the decreased summary measures, which will be discussed later in

more detail.

Smaller summary measures reflect lower values on the ordered rank probability matrix diagonal, and hence less confidence in the ranks. The decrease in these values is generally paired with an increase in standard deviation. It is useful to note that some summary measures are more variable than others. The determinant consistently yields more variability than the norm. The diagonal maximum varies less than the diagonal minimum, with the trace variability in between. The most variable measure is consistently either the diagonal minimum or the determinant, while the least variable measure is always either the norm or the diagonal maximum. Generally, the summary measures closer to 1 vary less, and summary measures that are further from 1 vary more.

While not all geometries are present, the complete square and star have been included here to demonstrate the loss of confidence when shifting from a symmetric geometry to a non-symmetric geometry. The variability of all values increases with the loss of symmetry. Inevitably, the number of studies required to have a symmetric network is less than the number of studies that exist in an asymmetric network when the studies per comparison are fixed. Further patterns will be explored in a number of plots below.

| Geom | ES | SpC | Det | DetSD | Norm | NormSD | Max | MaxSD | Min | MinSD | Tra | TraSD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Complete | 2 | 10 | 1.0000 | 0.0000 | 1.0000 | 0.0000 | 1.0000 | 0.0000 | 1.0000 | 0.0000 | 1.0000 | 0.0000 |
| | | 5 | 0.9999 | 0.0004 | 0.9999 | 0.0004 | 1.0000 | 0.0000 | 0.9999 | 0.0009 | 0.9999 | 0.0004 |
| | | 2 | 0.9858 | 0.0212 | 0.9871 | 0.0154 | 0.9972 | 0.0060 | 0.9759 | 0.0330 | 0.9865 | 0.0177 |
| | | 1 | 0.8548 | 0.0857 | 0.8870 | 0.0434 | 0.9477 | 0.0357 | 0.8029 | 0.0997 | 0.8727 | 0.0591 |
| | 1.6 | 10 | 0.9998 | 0.0012 | 0.9998 | 0.0012 | 1.0000 | 0.0000 | 0.9996 | 0.0024 | 0.9998 | 0.0012 |
| | | 5 | 0.9929 | 0.0222 | 0.9940 | 0.0117 | 0.9994 | 0.0027 | 0.9878 | 0.0284 | 0.9936 | 0.0145 |
| | | 2 | 0.8932 | 0.0962 | 0.9238 | 0.0447 | 0.9773 | 0.0356 | 0.8430 | 0.1221 | 0.9091 | 0.0698 |
| | | 1 | 0.6701 | 0.1508 | 0.7958 | 0.0506 | 0.8919 | 0.0733 | 0.6047 | 0.1381 | 0.7379 | 0.0920 |
| | 1.2 | 10 | 0.8273 | 0.1452 | 0.8962 | 0.0518 | 0.9691 | 0.0542 | 0.7607 | 0.1641 | 0.8628 | 0.0935 |
| | | 5 | 0.6588 | 0.1684 | 0.8109 | 0.0607 | 0.9171 | 0.0901 | 0.5530 | 0.1892 | 0.7194 | 0.1247 |
| | | 2 | 0.4511 | 0.1754 | 0.7016 | 0.0715 | 0.8021 | 0.1323 | 0.3727 | 0.1564 | 0.5417 | 0.1384 |
| | | 1 | 0.3333 | 0.1490 | 0.6417 | 0.0592 | 0.7135 | 0.1329 | 0.2881 | 0.1233 | 0.4401 | 0.1192 |
| Star | 2 | 10 | 0.9935 | 0.0158 | 0.9943 | 0.0118 | 0.9999 | 0.0005 | 0.9879 | 0.0270 | 0.9939 | 0.0135 |
| | | 5 | 0.9467 | 0.0743 | 0.9610 | 0.0307 | 0.9960 | 0.0084 | 0.9140 | 0.0915 | 0.9548 | 0.0463 |
| | | 2 | 0.6945 | 0.1343 | 0.8058 | 0.0469 | 0.9093 | 0.0607 | 0.6231 | 0.1319 | 0.7556 | 0.0820 |
| | | 1 | 0.3767 | 0.0852 | 0.6449 | 0.0099 | 0.7074 | 0.0600 | 0.3717 | 0.0690 | 0.5175 | 0.0430 |
| | 1.6 | 10 | 0.9425 | 0.0721 | 0.9585 | 0.0325 | 0.9961 | 0.0105 | 0.9056 | 0.1006 | 0.9508 | 0.0509 |
| | | 5 | 0.8097 | 0.1381 | 0.8845 | 0.0455 | 0.9698 | 0.0429 | 0.7332 | 0.1602 | 0.8481 | 0.0881 |
| | | 2 | 0.5185 | 0.1423 | 0.7335 | 0.0419 | 0.8434 | 0.0914 | 0.4412 | 0.1346 | 0.6162 | 0.0960 |
| | | 1 | 0.3180 | 0.1025 | 0.6321 | 0.0216 | 0.6841 | 0.0732 | 0.2962 | 0.0782 | 0.4530 | 0.0597 |
| | 1.2 | 10 | 0.5409 | 0.1565 | 0.7576 | 0.0498 | 0.8811 | 0.1004 | 0.4158 | 0.1635 | 0.6131 | 0.1213 |
| | | 5 | 0.4230 | 0.1626 | 0.6954 | 0.0576 | 0.7931 | 0.1221 | 0.2918 | 0.1380 | 0.4839 | 0.1210 |
| | | 2 | 0.3251 | 0.1460 | 0.6304 | 0.0535 | 0.6741 | 0.1267 | 0.1999 | 0.0899 | 0.3575 | 0.0972 |
| | | 1 | 0.2920 | 0.1221 | 0.6113 | 0.0427 | 0.6382 | 0.1106 | 0.1662 | 0.0805 | 0.3038 | 0.0900 |

Table 3.2: Table of Summary Measures from Simulation

The first two sets of graphs included below illustrate that all five summary measures behave as predicted with increases in effect size and studies per comparison. In Figure 3.2, all summary measures have been taken for each of the four geometries with the number of studies per comparison fixed at two. All summary measures increase with increases in effect size. It is somewhat evident here, and will be discussed later, that the triangle and complete square geometries yield higher summary measures than the star and dangle networks. There is a more significant increase in summary measures between the low and medium effect sizes than between the medium and high effect sizes.

A similar result is shown in Figure 3.3, where the number of studies per comparison is varied while the effect size is held constant at the medium value.

Although only one permutation for each of the star and dangle geometries has been included in Figures 3.2 and 3.3, it was verified that all permutations have the desired increases seen in these plots. For example, the four star permutations with changes to summary measures over increased studies per comparison are plotted in Figure 3.4. The rank of the star's central treatment is noted, S-1 indicates the best treatment is at the centre and so on. The summary measures all show an increasing trend, though there is some variation between the plots. Further investigation of the asymmetric geometries is included in the next section.

### 3.2.2 Comparison of Summary Measures

A clear visual of the changes in the summary measures over different geometries is depicted in Figure 3.5. Each network included had a medium effect size and two

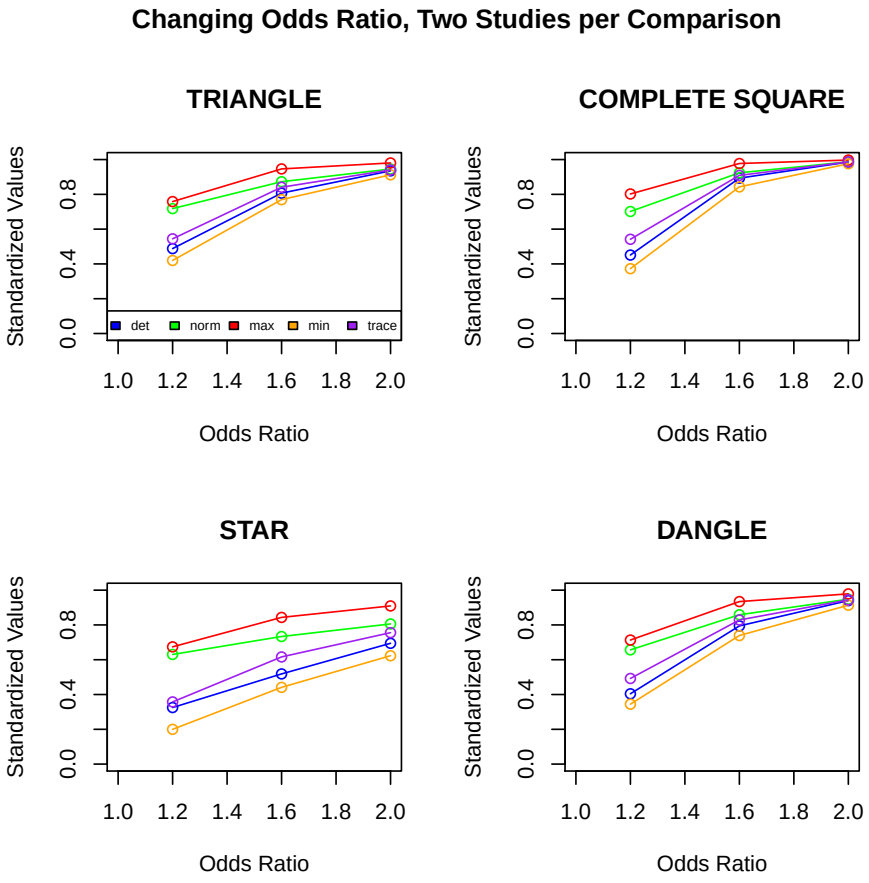


Figure 3.2: Summary Measures over Changes in Odds Ratios.

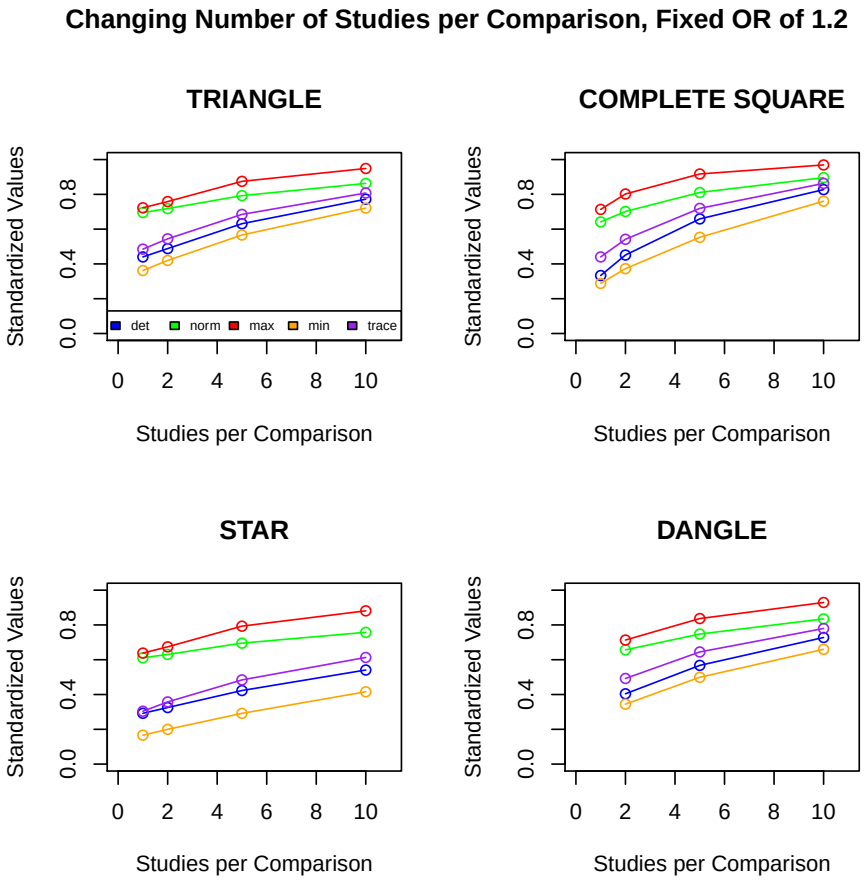


Figure 3.3: Summary Measures over Changes in Number of Studies per Comparison.

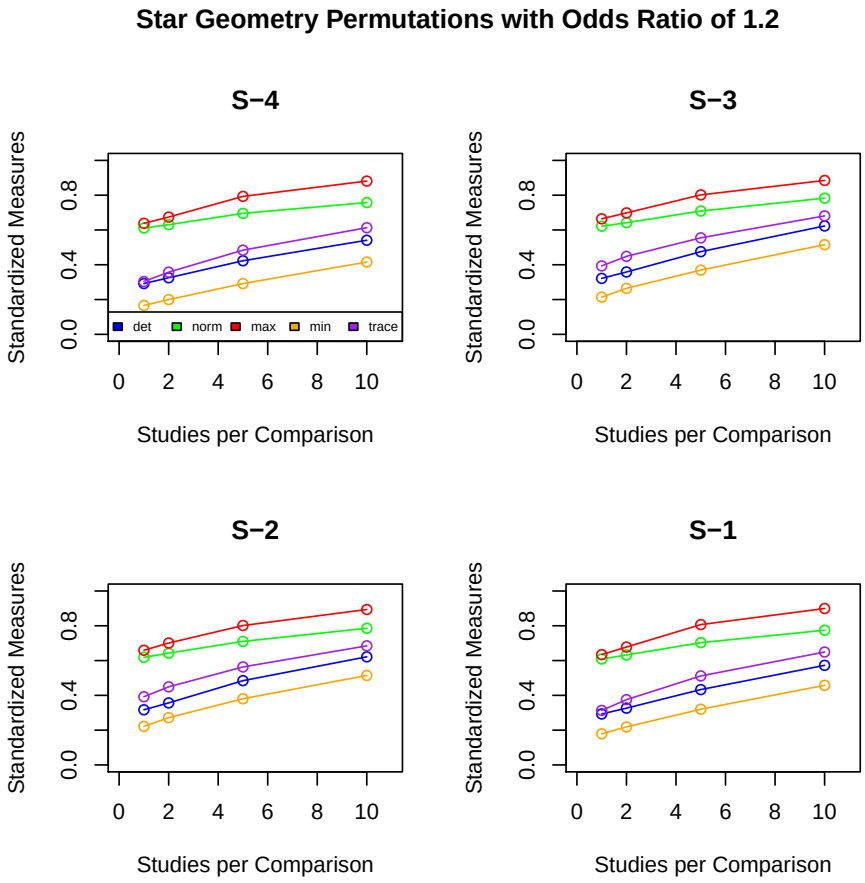


Figure 3.4: Changes to Summary Measures with Differing Permutations of Star Network.

studies per comparison.

To keep the labelling brief and intuitive, the star and dangle geometry permutations are labelled as follows: S-1 through S-4 represent star geometries with treatment ranks 1 through 4 at each center; dangle geometries are described by the rank of the treatment on the dangling arm and the rank of the treatment to which it is directly compared. For example, D2-3 indicates that the dangling treatment is the second best treatment and it has been compared directly with the third best treatment.

From this graph, it is clear to see that the triangle and complete square geometries yield higher summary measures. The star networks, which are the most asymmetric networks, have the lowest values. Within the asymmetric star and dangle geometries, the summary measures vary considerably. In the star permutations, having one of the extreme ranks, 1st or 4th, in the centre created the lowest summary measures. In the dangle permutations, having one of the central ranks, 2nd or 3rd, on the dangling arm created the lowest summary measures.

Regardless of geometry, the summary measures always take on values in the same order, from highest to lowest: diagonal max, Frobenius norm, trace, determinant and diagonal minimum.

A subset of the previous graph, looking only at the symmetric triangle and complete square geometries is shown in Figure 3.6. One might suppose that the network with only three treatments should yield more confident ranks. However, in this graph, both networks have been simulated using two studies per comparison. In total, the triangle network includes six studies while the complete square network contains twelve. It seems that these additional studies provide additional confidence

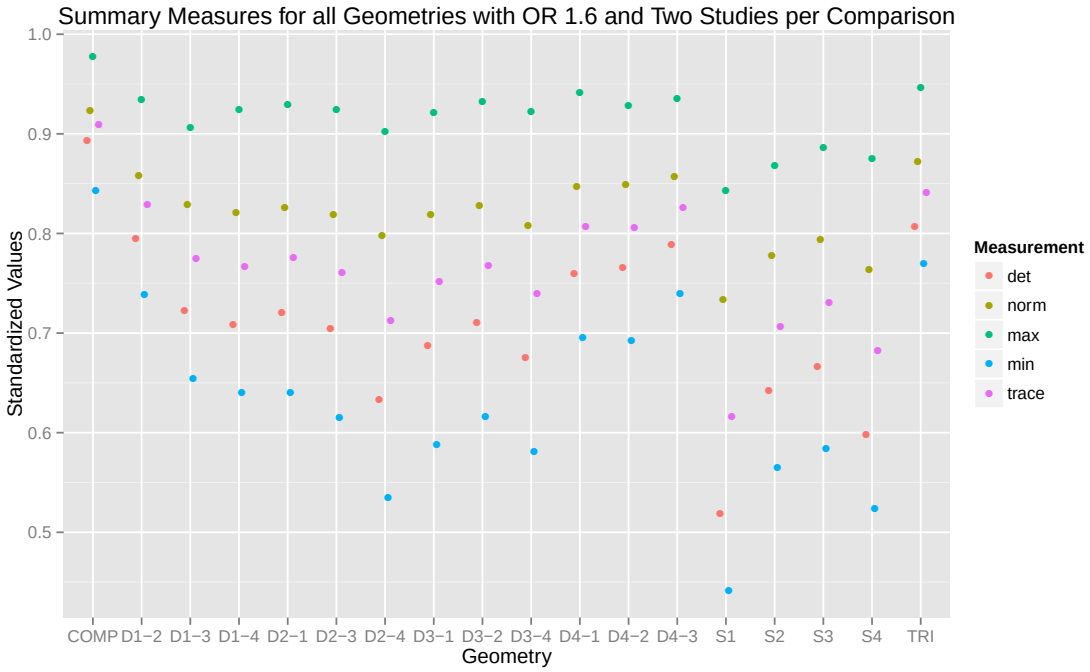


Figure 3.5: Summary Measures Over Four Geometries.

to the ranks, resulting in higher summary measures for the complete square network versus the triangle.

In order to verify that the triangle network is, in fact, easier to rank given equal amounts of information, the complete square network with two studies per comparison was compared to the triangle network with three studies per comparison. In this way, the treatments of both networks would be involved in exactly six studies. The values are very similar, but the triangle network yields slightly higher values, despite having fewer contributing studies overall. These differences isolate the effect of an added treatment to the network: given the same number of studies per treatment, fewer treatments yields more confidence in the ranking process and higher summary measures. The results are shown in Figure 3.7.

### 3.2.3 Introduction of a New Study and the Effects on Summary Measures

Although a network is analyzed at a specific point in time, it is important to remember that networks evolve over time. The dangle geometry was chosen for inclusion in this simulation because it seems to be a likely shape as a network grows. If three treatments exist and comparisons have been made between them, when a new treatment is introduced, it will likely be tested against only one of the pre-existing treatments creating a dangle geometry.

With this network growth in mind, consideration should be given to the case when only one study is available for the dangling comparison. When a new treatment is introduced, it is fair to assume that only one study will be available initially,

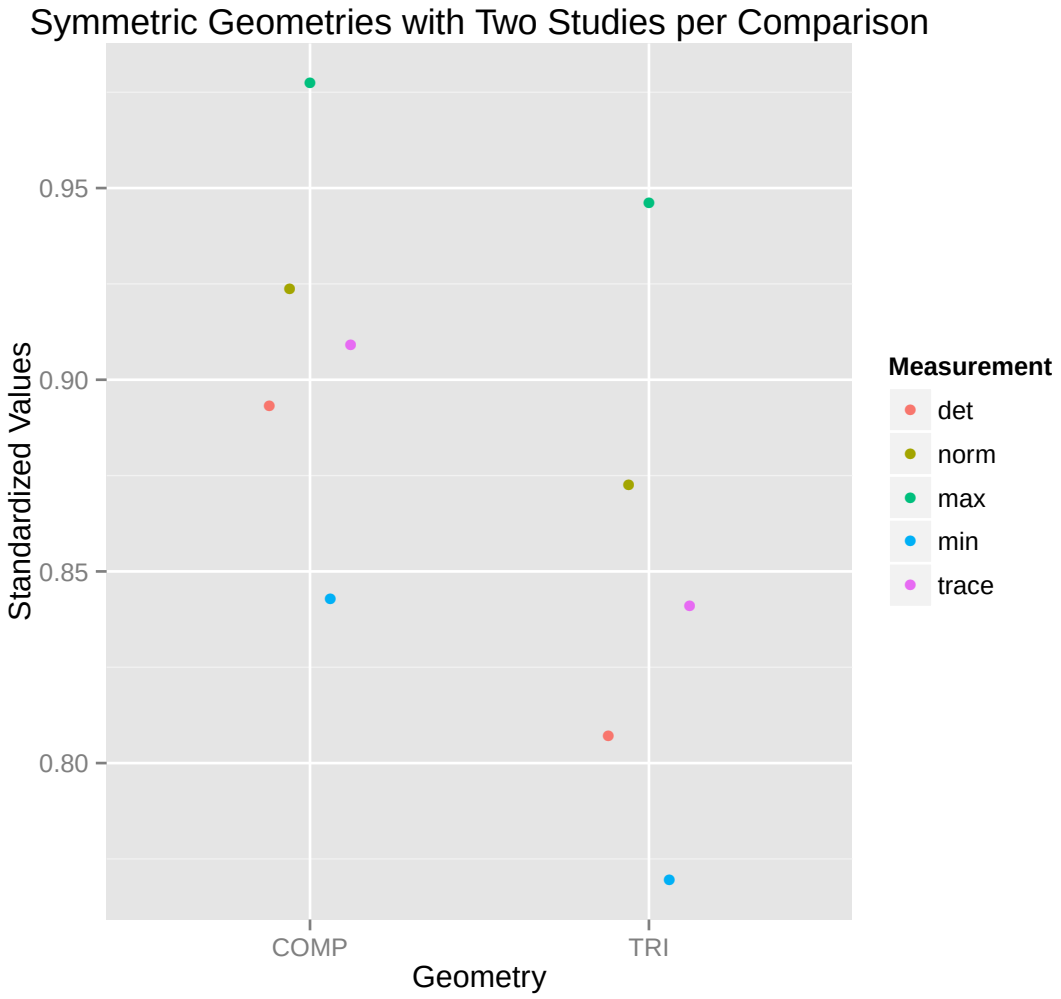


Figure 3.6: Comparison of Symmetric Networks with Three and Four Treatments.

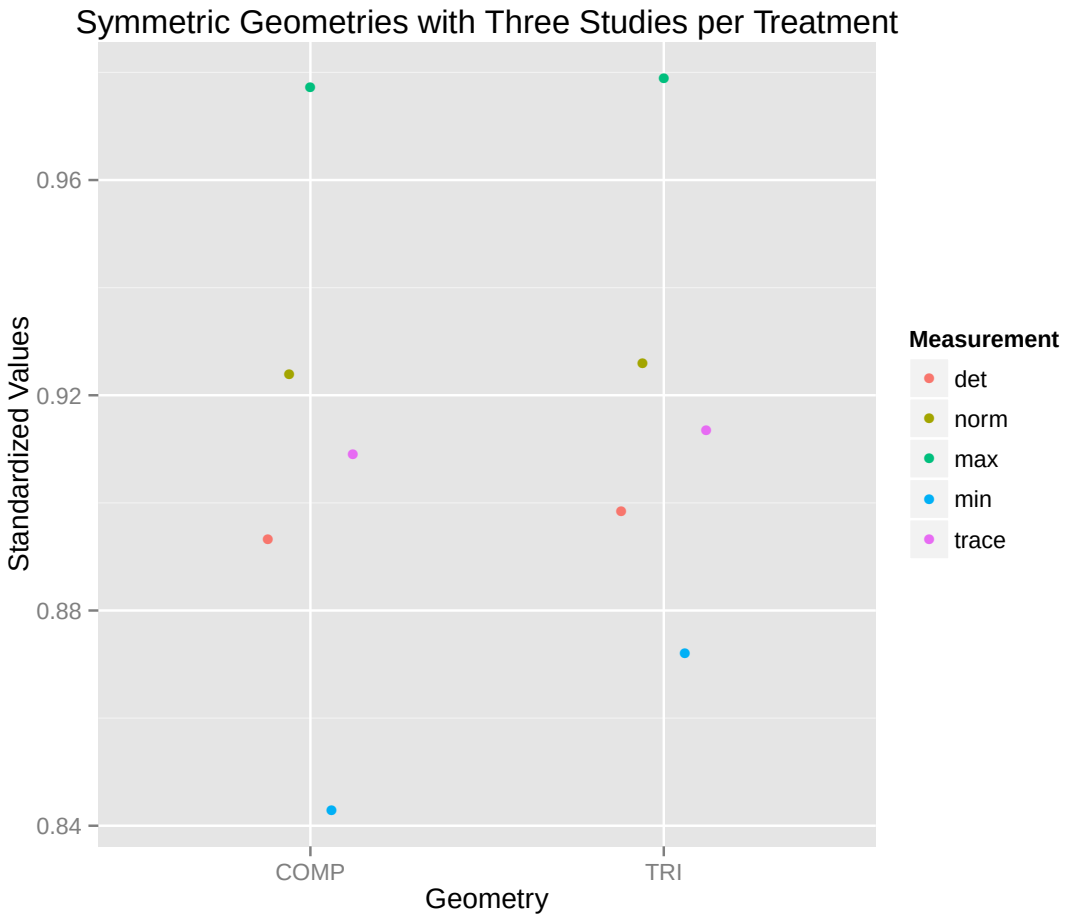


Figure 3.7: Matching Studies per Treatment in Symmetric Networks.

thus creating a triangle network with any number of studies per comparison, and a dangling arm with only one study per comparison.

To investigate the effect of adding a treatment to an existing triangle network, the summary measures of a triangle network with two studies per comparison have been compared to the summary measures of a dangle network with two studies per comparison in the closed loop and one study on the dangling arm. The results are displayed in Figure 3.8.

It is evident in the plot that adding a single treatment significantly lowers the summary measures, indicating a loss of confidence in the rankings. The rank of the newly introduced treatment plays a large role in how much the summary measures decrease - if the treatment on the dangling arm is ranked either 2nd or 3rd, this has the largest impact on lowering summary measures.

Assuming that a new treatment to the network will be studied beyond a single RCT, the improvement from having only one study on the dangling arm to having a matching number of studies to the other comparisons in the network is displayed in Figure 3.9. In this comparison, a medium effect size is taken, and the number of studies per comparison within the closed loop is five. The two plots illustrate the differences in summary measures when one or five studies is used to make the dangling comparison. Based on the plot, it is clear that the summary measures increase significantly with the additional information from the new studies. This pattern is reassuring that, although introducing a new study to a network might cause uncertainty in the ranks of the treatments temporarily, that with further research the confidence in the rankings can be regained.

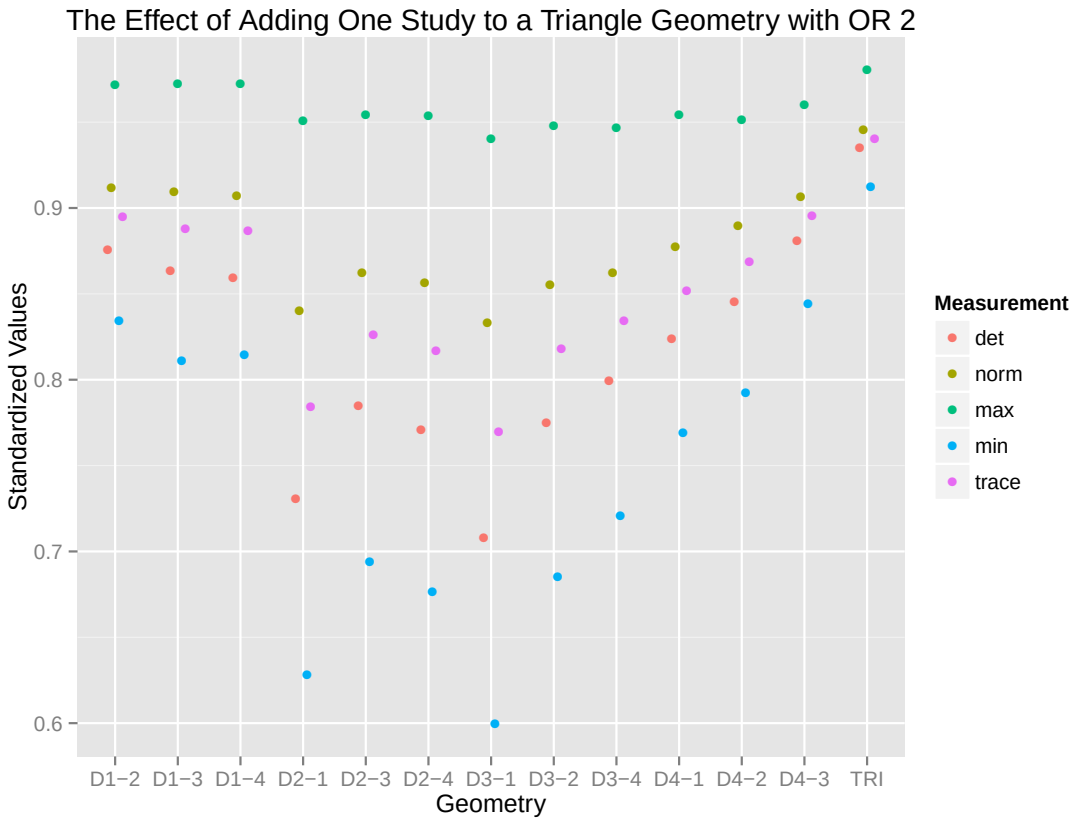


Figure 3.8: The Effect of Adding One New Treatment to a Triangle Network.

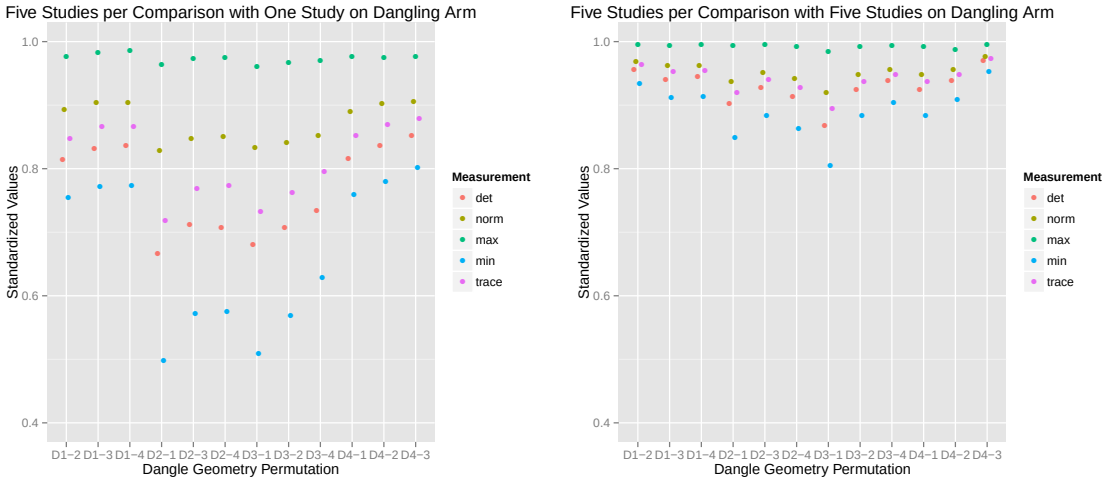


Figure 3.9: Dangle Geometry with One or Five Studies on the Dangling Arm

## 3.3 Comments on Summary Measures

Before exploring real data sets, a few comments about the summary measures are included.

First, the diagonal maximum and diagonal minimum values do not appear to be well chosen. For the maximum, the values are consistently high despite changes to the network, which can be misleading. If the goal is to get a sense of the overall confidence of a set of ranks, using a single rank probability is insufficient. If a network of four treatments has one that is clearly worst, the maximum will reflect the confidence in this single rank, while the interpreter may be primarily interested in the top two treatments. Similarly, the diagonal minimum does not provide enough information. It was also highly variable in the simulation study.

The norm, determinant and trace all provide a better overview of the ranks. The determinant is the most variable of these three measures, so either the trace

or the norm might be better options for summarizing the rank probability matrix. The applied data sets are used in the next section to gain more insight into which measures are most appropriate for summarizing the confidence in a set of ranks.

# Chapter 4

# Application to Real Data

In this chapter, each of the proposed summary measures is applied to previously published network meta-analyses. Each of these studies has been previously formatted into a file that is compatible with the GeMTC package for use in demonstrations of the software. For ease of labelling, the examples have been provided with reference names based on the study topic, provided in Table 4.1.

Table 4.1: Summary of Example Data Sets

| Ex | Reference Name | Data Source | Topic |
|---|---|---|---|
| 1 | Smoking | Lu and Ades (2006) | Smoking Cessation Methods |
| 2 | Thrombolysis | Lu and Ades (2006) | Thrombolysis after Acute MI |
| 3 | Anti-Depressants | Cipriani et al. (2009) | Anti-Depressant Efficacy |
| 4 | Systolic BP | Welton et al. (2009) | Interventions for CHD Patients |
| 5 | Cholesterol | Welton et al. (2009) | Interventions for CHD Patients |
| 6 | Parkinson | Dias et al. (2011c) | Drugs to Reduce Mean Off-Time |

Table 4.2: Summary of Example Networks

| Reference Name | Treatments | Studies | Participants | Direct Comparisons (Present/Possible) | Multi-Arm Studies |
|---|---|---|---|---|---|
| Smoking | 4 | 24 | 16737 | 6/6 | 2 |
| Thrombolytic | 8 | 26 | 145822 | 12/28 | 2 |
| Anti-Depressant | 12 | 111 | 24595 | 38/66 | 2 |
| Systolic BP | 2 | 9 | 2262 | 1/1 | 0 |
| Cholesterol | 2 | 14 | 3093 | 1/1 | 0 |
| Parkinson | 5 | 7 | 1613 | 6/10 | 1 |

## 4.1 Summary of the Real Data Examples

Each of the first three examples involves binary data and the odds ratio on the log scale is used to measure relative effects. The last three examples use continuous data with mean difference used to measure relative effects. Each data set will be described briefly and summarized in Table 4.2. In the next section, a consistency model is applied to the data and the proposed summary measures are calculated based on the rank probability matrix that is generated.

In order to use a consistency model for each of the real data sets, assessments from the original authors are relied upon to ensure appropriateness. When both fixed-effects and random-effects models are deemed appropriate, the random-effects has been applied here. This choice is based on the fact that contributing studies are usually different, and so the choice of random-effects is both cautious and realistic.

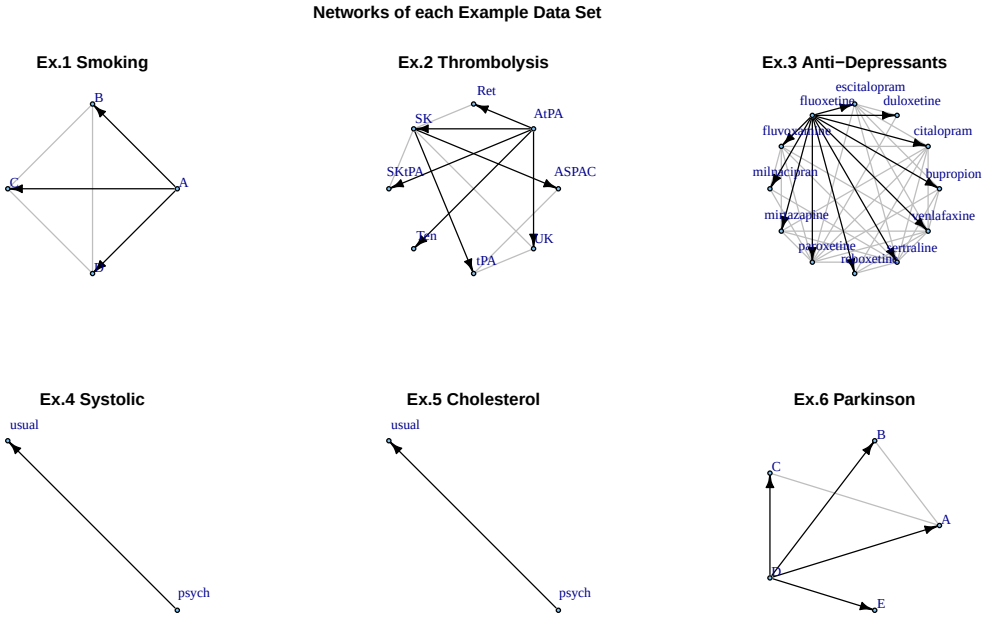


Figure 4.1: Network Geometries for Example Data.

### 4.1.1 Smoking

The first example is a smoking cessation study that compares the effectiveness of four methods. There are 24 trials included, two of which are three-arm while the remainder are two-arm. Each of the six possible pairwise comparisons is made directly in at least one two-arm trial. The four treatments of interest include no help (A), self-help (B), individual counseling (C), and group counseling (D). Visuals of this network and latter examples are available in Figure 4.1. The comparisons serving as basic parameters are highlighted in black, and all other direct comparisons in the network are noted with grey lines.

Lu and Ades (2006) apply both fixed-effects and random-effects models to the data. The authors find that the data is well fit under the random-effects model and poorly fit under the fixed-effects model. The choice is made to use a consistency model as no serious inconsistency is found. Applying the GeMTC default model this data set is appropriate: a consistency model with random effects. For the number and length of chains, as well as the burn-in, the default values are used.

Plots to assess convergence of the MCMC chains were produced for each example, though only the first is included here. In order to assess convergence, the R package coda is implemented (Plummer et al., 2006). Through this package, a Gelman and Rubin's convergence diagnostic and a Gelman-Rubin-Brooks convergence plot can be produced for inspection. Details on the theory of these diagnostic measures are available through the coda package and its references. From their description, "The potential scale reduction factor is calculated for each variable in x, together with upper and lower condence limits. Approximate convergence is diagnosed when the upper limit is close to 1." For each of the real data sets included, the upper limit remained between 1 and 1.02 for each variable. Figure 4.2 is a Gelman-Rubin-Brooks convergence plot for the smoking cessation example, which illustrates repeated calculations of the potential scale reduction factor, ensuring that values seen in the convergence diagnostic did not occur by chance. For each example, the convergence diagnostic and the convergence plot indicated that the chains converged.

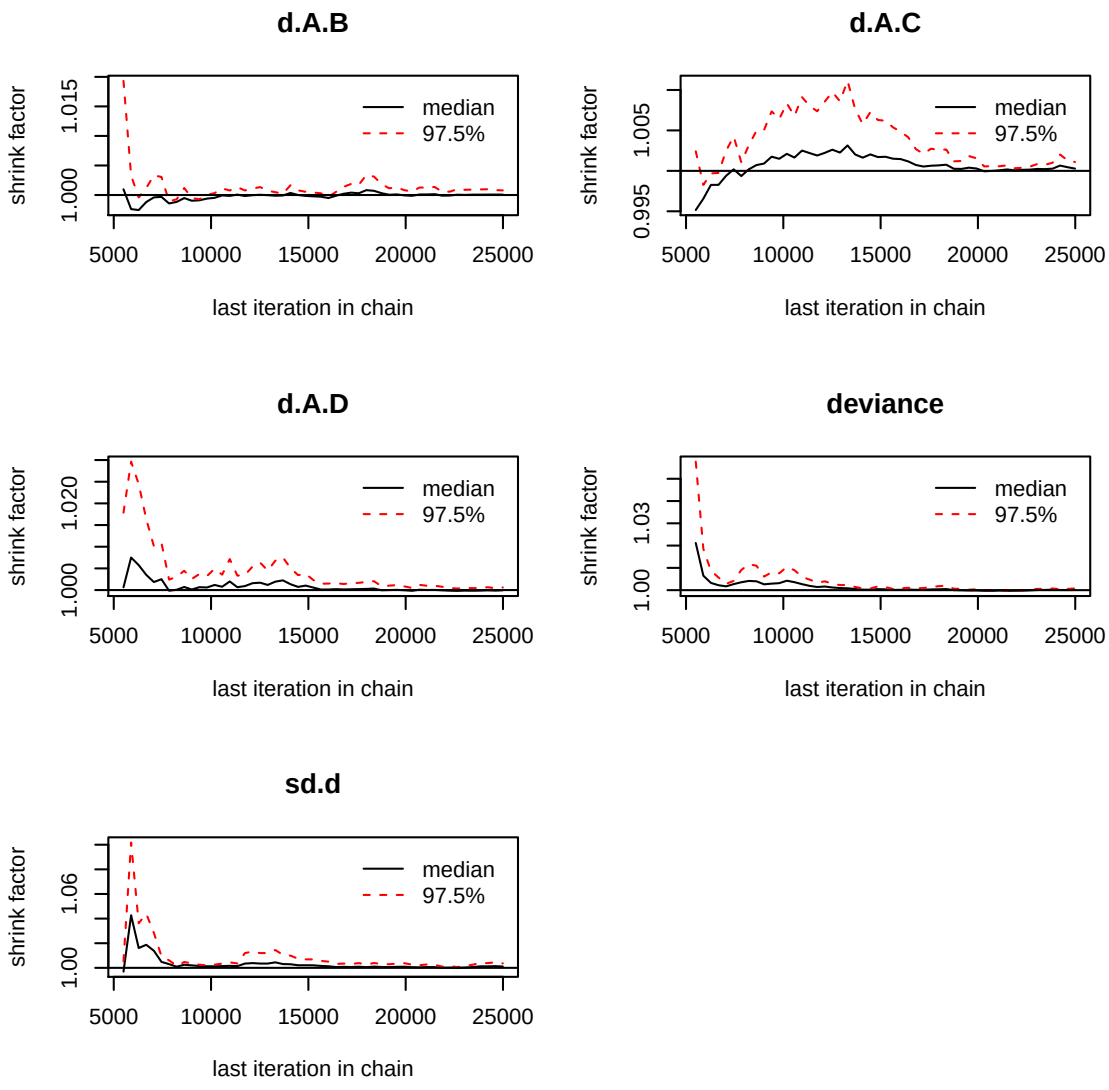


Figure 4.2: Gelman-Rubin-Brooks Convergence Plot for Smoking Cessation Example.

### 4.1.2 Thrombolysis

In the second example, 28 trials are combined to compare eight thrombolytic treatments after acute myocardial infarction. The treatments are streptokinase (SK), alteplase (tPA), accelerated alteplase (AtPA), reteplase (Ret), tenectedplase (Ten), streptokinase plus alteplase (SKtPA), urokinase (UK), and anistreptilase (ASPAC). Of the 28 potential pairs of comparisons, 13 pairwise comparisons are available from direct evidence.

The authors identify a high degree of inconsistency in the model. The inconsistency detected through multiple methods is identified to be caused by two trials in particular. (Discussion about methods for identifying inconsistency in this and other data sets is available in Dias et al. (2010)). Removing these two studies essentially removes the inconsistency issue. Thus, in order to apply a consistency model to this data, the violating studies were removed as suggested. As a result, 26 trials are included with 12 direct pairwise comparisons. In the consistency model, similar results are yielded between random-effects and fixed-effects models, so a random-effects model is selected here. Due to the large number of treatments, the MCMC chains and burn-ins were increased to length 50000.

### 4.1.3 Anti-Depressants

The third example compares the efficacy of 12 anti-depressants. One-hundred-eleven clinical trials are combined that include a total of 24595 participants. One-hundred-nine studies are two-arm trials and two are three-arm trials. There is direct evidence for 40 of 66 possible pairwise comparisons. The drugs considered for the treatment

of unipolar major depression in adults were buproprion, citalopram, duloxetine, escitalopram, fluoxetine, fluvoxamine, milnacipran, mirtazapine, paroxetine, reboxetine, sertaline and venlafaxine. The authors included studies that fit into a pre-defined range of follow-up time and had comparable drug dosage levels (classified as high, medium and low for each drug), and which did not have inadequate random allocation concealment and blinding. They defined response as the proportion of patients who had a reduction of at least 50% from the baseline score on the Hamilton depression rating scale or Montgomerysberg depression rating scale, or who scored much improved or very much improved on the clinical global impression at 8 weeks.

The assumption of consistency among RCTs was made. Upon exploration, statistically significant inconsistency was detected in three of 70 evidence loops, which was thought to align with what might reasonably happen by chance. Heterogeneity amongst the same pair-wise comparisons was moderate on average, which was expected due to the low number of studies defining each comparison. As a result, the consistency model with randome-effects was deemed appropriate. Due to the large number of treatments, the MCMC chains and burn-ins were increased to length 50000.

### 4.1.4 Systolic Blood Pressure and Cholesterol

The fourth and fifth examples consider the effect of psychological interventions on patients with coronary heart disease on several health outcomes. Since psychological interventions can involve different components and be difficult to classify, the

authors grouped treatments involving any of the following components as psychological interventions: educational, behavioural, cognitive, relaxation and support. Most interventions included more than one of these elements. These interventions were compared with usual care. A variety of outcome measures were considered both relating to heart health and psychological outcomes such as anxiety and depression. Three separate NMA's were conducted based on different outcome measurements, each using a slightly different set of RCTs. Two of those NMA's are being investigated in Examples 4 and 5. The fourth example involves nine studies where systolic blood pressure is the outcome of interest. The fifth example uses 14 studies with cholesterol as the outcome of interest.

The authors acknowledge the existence of heterogeneity in the varying definitions of usual care and in the various psychological treatments. For this reason, the random-effects model is clearly appropriate. Since there are only two treatments in this network, there is no chance of inconsistency, thus the consistency model is appropriate for both.

Although a network meta-analysis is an unnecessary framework to place on a network of only two treatments, including these studies provides greater diversity in the example networks considered with respect to number of treatments, number of studies, and total number of participants. Thus, these examples are included in the hope that they will help to reveal trends.

### 4.1.5 Parkinson

The sixth and final example compares five treatments for Parkinson's disease with evidence from seven trials. One of the seven trials is multi-arm. A total of 1613 participants are included in the contributing trials, and 6 of the possible 10 direct comparisons are available in the network. The treatments considered are placebo (A) and four active drugs (B-E) which are dopamine agonists used as adjunct therapy. More detail on the treatments is not provided in the paper. Mean off-time reduction in patients is the outcome of interest.

The random and fixed effects model both fit the data well, thus the random effects model is used here. An assessment of consistency is not included in the original paper, but is assumed in the modeling process so it is assumed appropriate for this example.

## 4.2 Results

For each of the examples included, the five summary measures of interest were taken from the ordered rank probability matrix. A summary of these values is presented in Table 4.3.

In the case of applications, unlike simulations, the true ranks of treatments are unknown. In order to use all of the proposed summary measures, the rank probability matrix must be rearranged into an ordered rank probability matrix without the assurance of a known order. This is the reality of applications. The order used is taken from the pair-wise relative effects generated in the modeling process. Since consistency is assumed, combining all of the pair-wise relative effects yields a list of

internally consistent ranks.

| | name | treatments | studies | det | norm | trace | max | min |
|---|---|---|---|---|---|---|---|---|
| 1 | Smoking | 4 | 24 | 0.667 | 0.755 | 0.714 | 0.889 | 0.597 |
| 2 | Thrombolysis | 8 | 26 | 0.108 | 0.451 | 0.262 | 0.505 | 0.169 |
| 3 | Anti-Depressants | 12 | 111 | 0.119 | 0.512 | 0.326 | 0.964 | 0.091 |
| 4 | Systolic BP | 2 | 9 | 0.776 | 0.825 | 0.801 | 0.801 | 0.801 |
| 5 | Cholesterol | 2 | 14 | 0.998 | 0.999 | 0.999 | 0.999 | 0.999 |
| 6 | Parkinson | 5 | 7 | 0.502 | 0.694 | 0.614 | 0.910 | 0.398 |

Table 4.3: Summary Measures for Example Networks

Several trends are identifiable from Table 4.3. Consistent with the simulations, for each example either the maximum or the norm is the highest summary measure and either the determinant or the minimum is the lowest summary measure. The two-treatment networks yield maximum, minimum and trace values that are identical, which is necessary due to the dimension.

The anti-depressant example, which has the highest number of treatments, yields the widest variety of summary measures. It is important to note that although the maximum value is 0.964, which was the probability that the 12th ranked treatment was in fact worst, the next highest entry in the diagonal of the ordered rank probability matrix was just above 0.5, with all others well below that. This is a clear example of why the diagonal maximum can be a misleading summary measure for the overall confidence in the ranks.

In order to explore potential relationships between network characteristics and the summary measures, several plots are included. The number of studies, number of treatments, studies per treatment, and participants per treatment are all considered as possible factors for the changes in summary measures between examples.

In Figure 4.3, the example networks are plotted by the number of studies included in the network. Looking at the plot one colour at a time, the trend in each summary measure is observable. Based on the plot, there is not a clear increase in summary measures with an increase in the number of contributing studies.

In Figure 4.4, the example networks are plotted by the number of treatments included in the network. Again, looking at the plot one colour at a time is beneficial. There is a clear positive relationship between the summary measures and the number of treatments in the network. This pattern is clearly seen in all summary measures except the maximum. The small amount of variation in the maximum over different networks is consistent with the findings of the simulation study.

In case the relationship between the summary measures and the networks is somewhat more complex, two additional plots are created that plot the example network's summary measures against the average number of studies per treatment and the average number of participants per treatment. These plots are displayed in Figure 4.5. Neither of these plots shows a relationship as clear as Figure 4.4.

Based on the simulations and the example data, it appears that the number of treatments in a network has a significant impact on the confidence in the rankings. Despite the fact that all summary measures are standardized for the number of treatments, there are still higher summary measures yielded by networks with fewer treatments. For comparison across networks of different sizes, the standardization process may require adjustment.

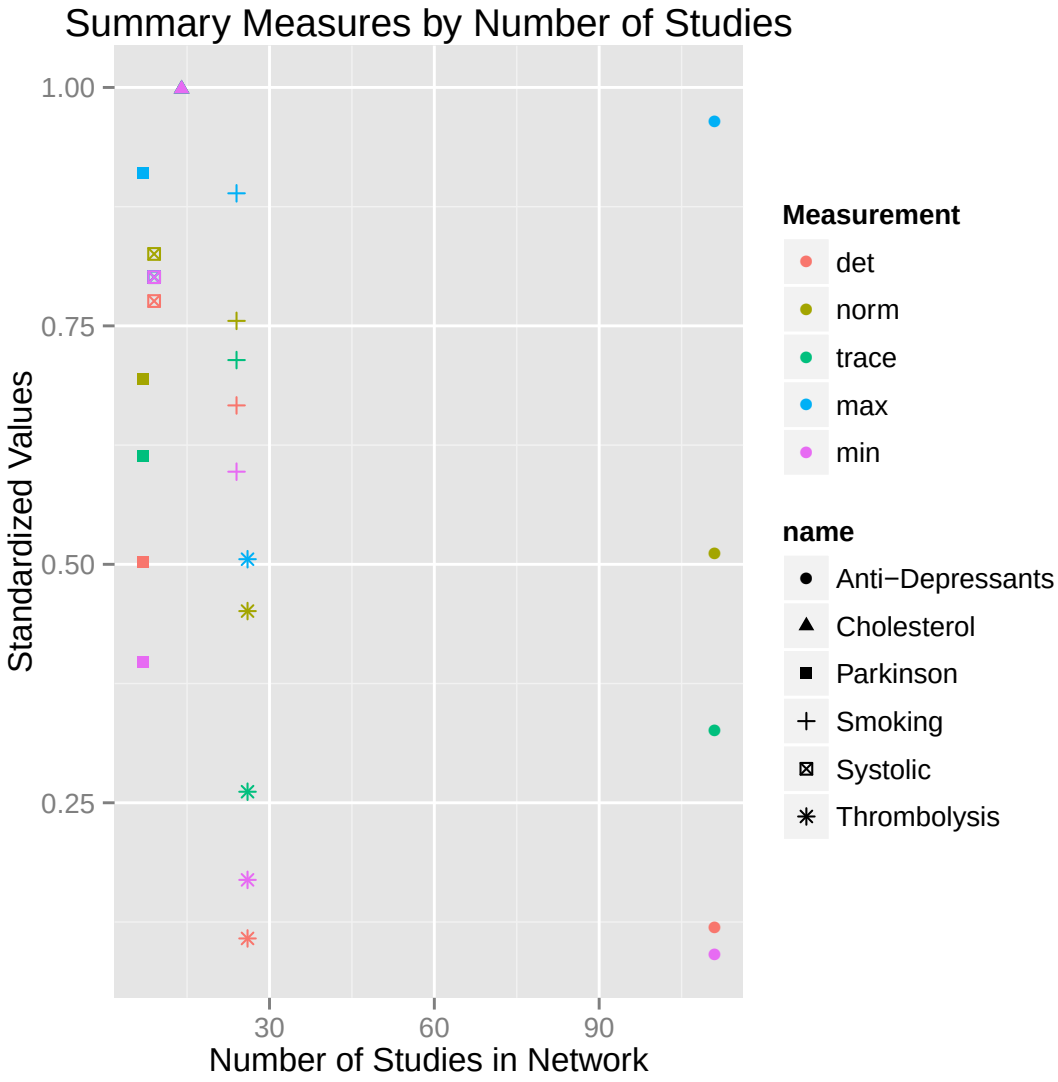


Figure 4.3: Summary Measures for Example Data plotted by Number of Studies in the Network.

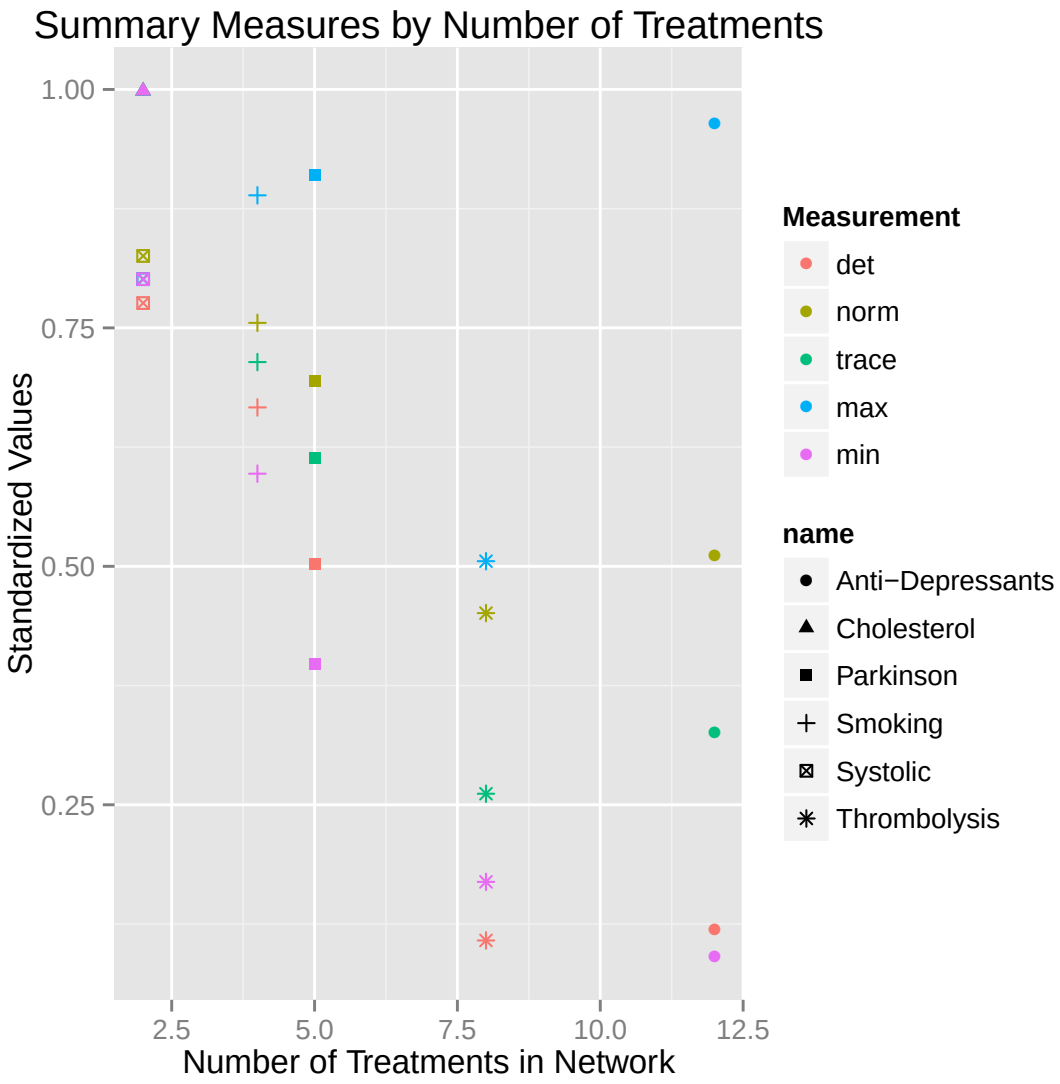


Figure 4.4: Summary Measures for Example Data Plotted by Number of Treatments in the Network.

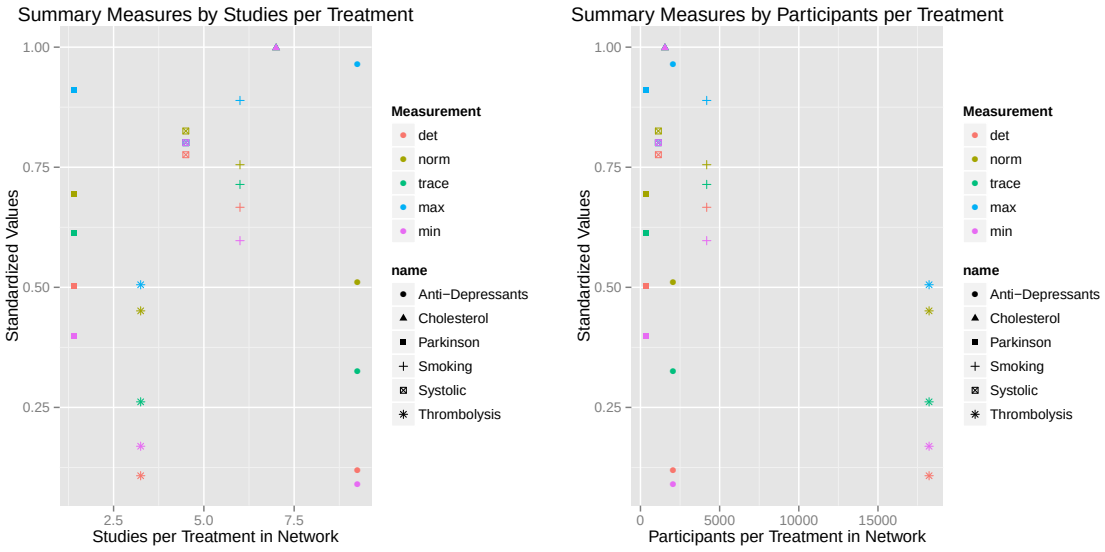


Figure 4.5: Summary Measures for Example Data Plotted by Participants per Treatment in the Network.

## 4.3 Implications for Summary Measures

The real data examples provide a much broader variety of networks than the simulation, which informs the understanding of various summary measures. It became clear in the anti-depressants example that the diagonal maximum is not reflective of the overall confidence in the rankings. Further, when a clear pattern emerged between the number of treatments in a network and the summary measures in Figure 4.4, the maximum did not show the same relationship as the other summary measures.

The real data examples also reveal that the determinant yields values around 0.1 for networks with 8 and 12 treatments. The network with 5 treatments had a determinant value near 0.5. There are several applications where 10 or more treatments might be included in a network meta-analysis. Based on these results,

the determinant might be a poor choice for a summary measure since the values appear to drop off significantly as the number of treatments increases.

An unanticipated issue identified in the analysis of the real data sets was the possibility of two treatments being found equally effective in the modeling process. This poses a problem in determining the treatment rankings for creating the ordered rank probability matrix. Since the determinant and norm are unaffected by changing the order of rows in the rank probability matrix, this issue only affects the reliability of the trace, diagonal maximum and diagonal minimum values. One solution to this issue might be to average the tied rows of the rank probability matrix and use the results in both row entries. However, this complicates the calculation of the summary measures, which were proposed based on ease of calculation and interpretation.

# Chapter 5

# Discussion and Future Directions

In the discussion section of this paper, findings of the simulation and real data analysis are summarized. The future directions section discusses limitations of the current work and provides recommendations for extending this research.

## 5.1 Discussion

This section provides suggestions as to which summary measures might effectively characterize the confidence in a set of treatment ranks yielded from a NMA. Based on the simulation study and the real data analysis, the following insights have been gained about each of the five summary measures considered:

**Standardized Absolute Determinant** The determinant does not require ordering the rank probability matrix, which allows issues with tied treatments to be avoided. Determinants were very low (near 0.1) for networks with 8 and

12 treatments, which suggests that little information may be gained from this measure in larger networks. The determinant has the potential to be weighted if certain ranks are considered more important than others, but this process would not be straight-forward. This measure was found to be one of the most variable in the simulation study.

**Standardized Frobenius Norm** The norm does not require ordering the rank probability matrix, which allows issues with tied treatments to be avoided. The values taken by the norm remained relatively high (all $> 0.45$) for all networks, despite higher numbers of treatments. The norm has the potential to be weighted if certain ranks are considered more important than others, but this process would not be straight-forward. This measure was found to be one of the least variable in the simulation study.

**Standardized Trace** The trace requires the rank probability matrix to be ordered, which might require tie breaking rules. The trace has the potential to be easily weighted if certain ranks are considered more important than others. The simulation revealed that its variance was higher than that of the norm.

**Diagonal Maximum** The maximum value does not vary with significant changes to networks, suggesting it does not capture enough information. It also requires the rank probability matrix to be ordered, which might require tie breaking rules.

**Diagonal Minimum** The minimum value does not allow for weighting of ranks deemed more important. It also requires the rank probability matrix to be

ordered, which might require tie breaking rules.

Based on the results of the simulation and real data sets, the Frobenius norm appears to be the best choice for a summary measure. It summarizes the confidence across all ranks and has lower variability than the trace or the determinant. The case of tied treatments in the network does not require special consideration.

All of the summary measures assume that the accuracy of each rank is equally important, which might not be the case. One possible issue for using the Frobenius norm is its inability to be weighted easily. However, if only one or two ranks are of interest to the investigator, then simply the columns of the rank probability matrix corresponding to the ranks of interest can be considered.

Several insights were gained from the simulation study beyond assessment of the summary measures. As predicted, all summary measures increased, reflecting confidence in the ranks, when the effect size or number of studies per comparison were increased. The symmetric networks, which have more studies per treatment than asymmetric networks, yielded higher summary measures. When the number of studies per treatment was fixed, the triangle network had higher summary measures than the complete square network. This was the first indication that the summary measures, although standardized, tend to decrease as the number of treatments in the network increases.

There was a significant loss of overall confidence when a new treatment was added to a triangle network through a single study. However, when studies involving the new treatment were added to the network, the confidence in the treatment ranks was largely regained.

The real data examples confirmed that the single-value summary measures, the diagonal maximum and the diagonal minimum, did not behave similarly to the three summary measures requiring calculation. The examples also provided insight into the relationship between the number of treatments in the network and the amount of confidence in the ranks.

## 5.2 Future Directions

The field of network meta-analysis is still very young. There are several ways for this research to be extended.

### 5.2.1 Extensions to the Simulation

The simulations conducted in this investigation represented over-simplified networks. The way in which studies were simulated did not introduce any heterogeneity or inconsistency. All studies were two-arm and simulated to be exactly the same size.

Increasing the number of treatments, varying the size of the individual studies, introducing inequality in the number of studies per comparison, violating assumptions and including multi-arm trials are just a few ways to extend this simulation study.

### 5.2.2 Theoretical Considerations

Absent from this thesis is an exploration into the theoretical properties of the proposed summary measures.

As the rank probability matrix is random, it has an accompanying distribution. So then do each of the proposed summary measures.

A theoretical approach is needed to fully understand the distributions of the rank probability matrix and the summary measures of interest. Once these distributions are known, the usefulness of these summary measures extends beyond point estimation to include confidence intervals and hypothesis testing.

Further, the distributions will help to set benchmark values for the summary measures that could identify the rank confidence as strong or weak, for example.

The hope is that this thesis will serve as a starting point for the possible assessments of a NMA rank probability matrix, and how it might be summarized with a simple and informative measure.

# Bibliography

Bucher, H. C., Guyatt, G. H., Griffith, L. E., and Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology*, 50(6):683–691.

Chalmers, I. (1993). The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences*, 703(1):156–165.

Chalmers, I., Hedges, L. V., and Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions*, 25(1):12–37.

Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J., Churchill, R., Watanabe, N., Nakagawa, A., Omori, I. M., McGuire, H., et al. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet*, 373(9665):746–758.

Dias, S., Sutton, A. J., Welton, N. J., and Ades, A. (2011a). Nice dsu technical support document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Last updated April 2012; Available from http://www.nicedsu.org.uk.

Dias, S., Welton, N., Caldwell, D., and Ades, A. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in medicine*, 29(7-8):932–944.

Dias, S., Welton, N. J., Sutton, A. J., and Ades, A. (2011b). Nice dsu technical support document 1: Introduction to evidence synthesis for decision making. Last updated April 2012; Available from http://www.nicedsu.org.uk.

Dias, S., Welton, N. J., Sutton, A. J., and Ades, A. (2011c). Nice dsu technical support document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. Last updated April 2012; Available from http://www.nicedsu.org.uk.

Dias, S., Welton, N. J., Sutton, A. J., Caldwell, D., Lu, G., and Ades, A. (2011d). Nice dsu technical support document 4: Inconsistency in networks of evidence based on randomised controlled trials. Last updated April 2012; Available from http://www.nicedsu.org.uk.

Fleiss, J. (1993). Review papers: The statistical basis of meta-analysis. *Statistical methods in medical research*, 2(2):121–145.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8.

Higgins, J. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, 327(7414):557.

Hoaglin, D. C., Hawkins, N., Jansen, J. P., Scott, D. A., Itzler, R., Cappelleri, J. C., Boersma, C., Thompson, D., Larholt, K. M., Diaz, M., et al. (2011). 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 in Health*, 14(4):429–437.

Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., Lee, K., Boersma, C., Annemans, L., and Cappelleri, J. C. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ispor task force on indirect treatment comparisons good research practices: part 1. *Value in Health*, 14(4):417–428.

Jonas, D., Wilkins, T., Bangdiwala, S., Bann, C., Morgan, L., Thaler, K., Amick, H., and Gartlehner, G. (2013). Findings of bayesian mixed treatment comparison meta-analyses: Comparison and exploration using real-world trial data and simulation. (prepared by rti-unc evidence-based practice center under contract no. 290-2007-10056-i.).

Li, T., Puhan, M., Vedula, S., Singh, S., Dickersin, K., et al. (2011). Network meta-analysis-highly attractive but more methodological research is needed. *BMC medicine*, 9(1):79.

Lu, G. and Ades, A. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in medicine*, 23(20):3105–3124.

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

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

Mills, E. J., Ghement, I., O'Regan, C., and Thorlund, K. (2011). Estimating the power of indirect comparisons: a simulation study. *PLoS One*, 6(1):e16237.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

Normand, S.-L. T. (1999). Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in medicine*, 18(3):321–359.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, 3(2):80–97.

Salanti, G., Higgins, J. P., Ades, A., and Ioannidis, J. P. (2008a). Evaluation of networks of randomized trials. *Statistical methods in medical research*, 17(3):279–301.

Salanti, G., Kavvoura, F. K., and Ioannidis, J. P. (2008b). Exploring the geometry of treatment networks. *Annals of internal medicine*, 148(7):544–553.

Salanti, G. and Schmid, C. (2012). Special issue on network meta-analysis.

Schafer, J. L. (2010). *Analysis of incomplete multivariate data*. CRC press.

Song, F., Clark, A., Bachmann, M. O., and Maas, J. (2012). Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC medical research methodology*, 12(1):138.

Spine, M. A. (2010). Facts, fallacies, and politics of comparative effectiveness research: Part i. basic considerations. *Pain Physician*, 13:E23–E54.

Thorlund, K. and Mills, E. (2012). Stability of additive treatment effects in multiple treatment comparison meta-analysis: a simulation study. *Clinical epidemiology*, 4:75.

van Houwelingen, H. C., Arends, L. R., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in medicine*, 21(4):589–624.

van Valkenhoef, G. and Kuiper, J. (2013). *gemtc: GeMTC network meta-analysis*. R package version 0.5-2.

Welton, N. J., Caldwell, D., Adamopoulos, E., and Vedhara, K. (2009). Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *American journal of epidemiology*, 169(9):1158–1165.