

國立臺灣大學公共衛生學院
流行病學與預防醫學研究所
博士論文



Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University
Doctoral Dissertation

網絡統合分析排名之可信度
Assessing the Reliability of Treatment Rankings
in Network Meta-Analyses

吳昶麋
Yun-Chun Wu

指導教授：杜裕康 博士
Advisor: Yu-Kang Tu, D. D. S., Ph. D.

中華民國111年6月
June 2022



國立臺灣大學博士學位論文
口試委員會審定書

論文中文題目：網絡統合分析排名之可信度

論文英文題目：Assessing the Reliability of

Treatment Rankings in Network Meta-Analyses

本論文係吳昀麋君（學號 D03849009）在國立臺灣大學
流行病學與預防醫學研究所完成之博士學位論文，於民國
111 年 1 月 19 日承下列考試委員審查通過及口試及格，特此
證明。

口試委員：

杜福康

（簽名）

（指導教授）

李文宗

東植忠

陳偉華

黃維誌

致謝



非常感謝我的博士班論文指導老師—杜裕康老師，謝謝老師從論文题目的構思、分析、撰寫、期刊選擇、到投稿，都給予我非常寶貴的建議與指導，不厭其煩地與我進行討論和不斷地修改潤飾文章，這論文才得以完成，非常感謝老師的指導。謝謝論文口試的委員們—蕭朱杏老師、李文宗老師、陳錦華老師、東雅惠老師、余聰老師、James Hodges 老師，謝謝您們給予我很多重要的建議和修改方向。

在博士求學生涯中，我想特別的感謝林先和老師，謝謝您總是給我很多的支持與鼓勵，讓我能夠堅持到最後。我也非常感謝從大學就是我的導師的蕭朱杏老師，能夠遇到您當我的導師，真是何其幸運，謝謝您總是讓我感受到無比的溫暖。我也很謝謝在西雅圖的 Mohsen Naghavi 老師，謝謝您啟發我很多人生智慧。

非常感謝我的家人一路的支持，謝謝爸爸陪著我到西雅圖，給我到國外闖蕩一年的勇氣。謝謝媽媽總是煮最好吃的晚餐等我回家，讓我有暖暖的力量。謝謝我的大姊，跟我一起在博士班努力，彼此打氣！謝謝我的二姊，陪伴我參與博士生涯中大大小小的重要時刻，安撫我每一次的緊張。謝謝我的三姊，總是用最幽默的方式，讓我知道不要把自己的人生過得太難！謝謝我的姊夫們和在我博士班過程中陸續出生長大的姪子姪女們，你們讓我的生活多采多姿！

謝謝每一位陪伴我走完博士路的朋友，謝謝你們對我的好，我將謹記在心頭。謝謝博士班同學們，芊芊、尼可、巧兒、悅哥、渲渲、姿婷、意婷、舜淳、書如、王睿，有你們一起度過的博士生涯，是我最難忘而美好的回憶。謝謝我在西雅圖



的朋友們，Toño、Angela、Liane、Thomas、Jonathan、Chun、Yixian，你們讓我覺得西雅圖就像是我第二個家。謝謝 Kyle Foreman，開啟我的獨旅人生，讓我意識到閉關寫作是多麼的重要。謝謝研究室一起打拼的夥伴們，曾醫師、銘杰、琪婕、伊婷，謝謝陪我運動的夥伴們，俞伶、小伍、柏威、柏辰、易瑄、馬柔、苡暉、Anne，謝謝蕙竹、怡暄、Ken、Tim、博文，每一次和你們的對話總是會讓我學習到很多。謝謝 Andrei Akhmetzhanov，帶我去跑越野跑，開拓我的視野，讓我可以更寬闊的心看這個世界。

八年的時間，謝謝我自己，走完這趟旅程，堅持到這裡。

昀麋 2022/6/26

中文摘要




近年來國際上的臨床指引廣泛採用網絡統合分析的結果去提供疾病治療的建議。透過網絡統合分析，整合多個研究的直接證據與間接證據，去估算多種治療方法或者介入措施之間的差異，藉此可以提供實證醫學一個有利的工具來填補當前的知識缺口。然而，由於網絡統合分析中所包含的多為有效的治療方式，因此它們之間的效果的差異往往很小或是沒有達到統計上顯著，故其結果並不容易解讀。因此，就有研究者提出使用排名的方法來去簡化治療之間比較結果的解讀。

透過排名，可以讓網絡統合分析結果的資訊簡單化，也是將實證數據轉化為臨床實務的一種方式。排名讓複雜的網絡統合分析結果容易解讀，然而，不論資料多寡，只要可以進行網絡統合分析，就能取得其治療的排名。但是，在排名上有差別的兩個治療，並不表示他們之間的差異就很顯著。因此，使用排名卻不報告排名的可信度，往往會導致誇大解讀不同介入或治療之間的差異。

目前排名可信度的評估方法包括不確定性評估和穩健性評估。當前排名不確定性的評估方法，會受到網絡所包含的治療數目所影響，因此被批評此指標是資訊缺乏的。而穩健性評估與不確定性評估之間的關聯為何，目前並未有定論。因此，本論文旨在建立評估網絡統合分析排名可信度的方法，以強化對網絡統合分析的解讀和應用。本論文所探討的問題如下所述：

1. 發展網絡統合分析中治療排名不確定性的替代指標。
2. 探討排名的穩健性和不確定性關聯性。

針對以上研究問題，首先，本論文提出應用標準化熵這個度量，來將每種治療的排名機率分佈轉換為一個數值指標，以促進對治療排名不確定性的精確解讀。

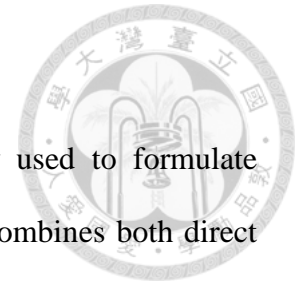


標準化熵是一個介於 0 到 1 之間的指標，越大表示不確定性越高。與傳統的指標相比，此指標不會受到包含在網絡當中的治療數目多寡影響，因此，它可用於比較網絡統合分析中不同治療，或是和不同網絡統合分析之間治療排名的不確定性。而本論文利用網絡統合分析資料庫，使用標準化熵評估 157 篇已發表的網絡統合分析，其中排名不確定性高的網絡統合分析占約三分之二。此外，本論文利用已發表的網絡統合分析，來探討排名不確定性和穩健性的關係。從結果看到，與預期相符的是，排名不確定性很低時，相對的排名穩健性也很高。然而，排名穩健性高並非總是對應於低不確定性，具有高穩健治療的排名也可能同時具有高不確定性。因此，當穩健性高的時候，並不表示此排名未來不容易改變，只能說在此網絡所包含的試驗中，沒有單一一個試驗是會對排名有很大影響的。

在報告排名時，利用標準化熵來呈現排名不確定性可以讓我們避免對排名的過度解度。目前已發表的網絡統合分析，排名的不確定性極高，表示其排名可能會在未來有新的試驗加入時改變。而透過一次排除一個試驗來看排名穩健性的方法，只能審視目前包含的試驗是否對排名會有很大的影響，並非與排名不確定性有絕對的相關性。

關鍵詞：實證決策、網絡統合分析、排名、不確定性、標準化熵、穩健性

Abstract



In recent years, network meta-analysis (NMA) has been widely used to formulate recommendations in the guidelines for managing diseases. NMA combines both direct and indirect evidence to compare multiple treatments and has been shown to be a useful tool for bridging the knowledge gap in evidence-based medicine. Since the differences among active treatments in the efficacy or harm are likely to be small, researchers develop methods to rank treatment for aiding the interpretation of treatment comparisons.

Ranking makes information from NMA simpler and is also a way to translate evidence into clinical practice. However, although ranking facilitates the interpretation of complex results from NMAs, its reliability has caused much controversy. As ranking can always be obtained from NMA, the difference in ranking does not mean that difference between treatments is statistically significant. Therefore, using rankings without reporting the reliability of the rankings may either make interpretation difficult or exaggerate the small differences between treatments.

Currently, ranking uncertainty and ranking robustness are two methods for evaluating the reliability of ranking. However, the current method used to evaluate the uncertainty of ranking would be affected by the number of treatments included within the network. Therefore, it is criticized as uninformative. The association between ranking uncertainty and ranking robustness has not been fully explored. Thus, the objective of this dissertation is to develop methods to facilitate the interpretation and application of NMA rankings. This dissertation aims to address the following objectives:

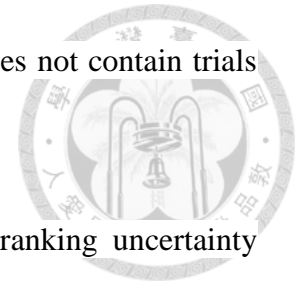
1. Develop an alternative method to measure the uncertainty of treatment ranking from NMA.
2. Explore the association between the uncertainty of ranking and robustness of ranking



For the first objective, I proposed Normalized Entropy, which transforms the distribution of ranking probabilities into a single quantitative measure to facilitate a refined interpretation of uncertainty of treatment ranking. I showed that as Normalized Entropy ranges from 0 to 1 and is independent of the number of treatments, it can be used to compare the uncertainty of treatment rankings within an NMA and between different NMAs. Normalized Entropy is an alternative tool for measuring the uncertainty of treatment ranking by improving the translation of results from NMAs to clinical practice and avoiding naïve interpretation of treatment ranking. I also evaluated the uncertainty of ranking for 157 published NMAs. Among them, two-thirds of NMAs have high or very high ranking uncertainty.

In the results of the second objective of the dissertation, the association between uncertainty and robustness of ranking was explored. The results showed that low uncertainty corresponds to high robustness. When the uncertainty of ranking is very low, treatment ranking is unlikely to be altered by deleting a trial from the complete data. However, good robustness of ranking does not always correspond to low uncertainty. NMA with robust treatment ranking may have high uncertainty of treatment ranking. Therefore, if the network does not contain a trial that significantly impacts the ranking, even if the uncertainty is high, the ranking robustness can still be high. The high robustness of ranking does not mean that the ranking will not be easily changed when

new trials are added in the future, but it means that the network does not contain trials that have a significant impact on the treatment ranking.



When reporting rankings, using Normalized Entropy to present ranking uncertainty prevents us from naïve interpretation of treatment ranking. Among the current published NMAs, most of them have high uncertainty of ranking, and their rankings may have a higher possibility to be changed when new trials are added into the network in the future. The robustness of ranking, which is evaluated by the leave-one-trial-out approach to identify trials included in the network that substantially influence the treatment ranking, is not entirely related to the uncertainty of ranking.

Keywords: evidence-based decision-making; network meta-analysis; ranking; uncertainty; Normalized Entropy; robustness

Contents



論文口試審定書	i
致謝	ii
中文摘要	iv
Abstract	vi
Contents	ix
List of Tables	xii
List of Figures	xiii
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: LITERATURE REVIEW	4
2.1 Development of Network Meta-Analysis	4
2.2 Models and Assumptions for Network Meta-Analysis	6
2.2.1 Model for Network Meta-analysis	7
2.2.2 Assumptions of Network Meta-Analysis	8
2.3 Ranking	10
2.3.1 Ranking Algorithms	11
2.3.2 Uncertainty of Treatment Ranking	14
2.3.3 Robustness of Ranking	20
2.3.4 Factors Affecting Treatment Ranking	21
2.3.5 An Illustrative Example	22
2.4 Decision Making under Uncertainty	31
2.4.1 Entropy	31
2.4.2 Applications of Entropy	33
2.4.3 Criteria of Normalized Entropy	34

CHAPTER 3: AIMS AND OBJECTIVES	36
CHAPTER 4: MATERIALS AND METHODS	38
4.1. Current and Proposed methods.....	38
4.1.1. 95% CI of SUCRA.....	38
4.1.2. Normalized Entropy	39
4.1.3. Euclidean Distance	43
4.1.4. Variance and Standard Deviation	44
4.2. Simulations	47
4.2.1. Comparing Normalized Entropy and 95% CI of SUCRA.....	47
4.2.2. Comparing Normalized Entropy and P(Best).....	49
4.3. Reanalysis of NMAs.....	50
4.3.1. NMA Database	50
4.3.2. Four Examples for Comparing Current and Purposed Methods	51
4.3.3. Two Examples for Graph Illustration.....	52
4.4. Robustness of Ranking.....	53
4.4.1. Cohen’s kappa coefficients.....	53
4.4.2. Treatment-level and NMA-level assessment.....	55
4.4.3. Association between the Uncertainty and Robustness of Ranking	55
CHAPTER 5: Results.....	58
5.1 Proposed Methods.....	58
5.1.1 Comparing Normalized Entropy, Rankogram, and the Width of 95% CI of SUCRA.....	58
5.1.2 Comparing Normalized Entropy, Normalized Variance, and Normalized Standard Deviation.....	64
5.2 Simulations	67



5.2.1 Comparing Normalized Entropy and 95% CI of SUCRA.....	67
5.2.2 Comparing Normalized Entropy and P(Best).....	70
5.3 Ranking Uncertainty of Published NMA	73
5.3.1 The Distribution of Ranking Uncertainty for Published NMAs	73
5.3.2 Two Illustrative Examples	79
5.4 Association between Uncertainty and Robustness of Treatment Ranking..	85
5.4.1 NMA-level Association between Uncertainty and Robustness.....	95
5.4.2 Treatment-level Association between Uncertainty and Robustness....	100
5.4.3 Regression Analysis	101
CHAPTER 6: Discussion and Conclusions	102
6.1 Using Normalized Entropy to Measure Uncertainty of Rankings	102
6.2 Strengths and Limitation of Normalized Entropy	103
6.3 Is Providing Uncertainty Intervals in Treatment Ranking Helpful?	104
6.4 How is Normalized Entropy Related to Variance?	105
6.5 High Robustness Does Not Always Imply Low Uncertainty of Treatment	
Rankings	106
6.6 Evaluation at NMA-level, Treatment-level, and Trial-Level	107
6.7 Limitations of the Study of Robustness and Uncertainty of Ranking	108
6.8 Presentation of Uncertainty with Ranking.....	108
6.9 Conclusions.....	109
6.10Future work.....	109
REFERENCE	110

List of Tables

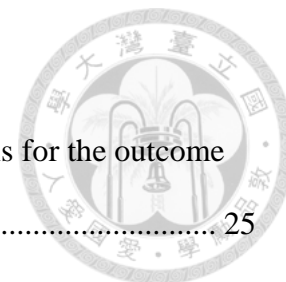


Table 1. Results of network meta-analysis and pairwise meta-analysis for the outcome AHI in the illustrative example	25
Table 2. Results of network meta-analysis and pairwise meta-analysis for the outcome ESS in the illustrative example.....	26
Table 3. Ranking probabilities for the outcome AHI.....	27
Table 4. Formula and the value range for entropy and variance	45
Table 5. Variables used for regression analysis of 60 NMAs and 348 treatments.....	57
Table 6. Ranking probabilities, the width of 95% CI of SUCRA, and Normalized Entropy of the four network meta-analysis studies	61
Table 7. Ranking probabilities, Entropy/Normalized Entropy, and Variance/Normalized Entropy of the four network meta-analysis studies	66
Table 8. Summary of the 157 NMAs.....	75
Table 9. Summary of the 60 NMAs.....	87
Table 10. Basic characteristics of 60 NMAs	88
Table 11. Uncertainty of ranking evaluated by average normalized entropy and robustness of ranking evaluated by LOTO and Cohen’s kappa for 60 NMAs. 93	
Table 12. Comparison of five levels between the uncertainty of treatment ranking quantified by the average normalized entropy and robustness of treatment ranking quantified by the (A) Average (B) Minimum (C) Maximum value of quadratic Cohen’s kappa within the network	98
Table 13. Results of each model to explore the association between the robustness of treatment ranking and uncertainty of treatment ranking for 60 NMAs and 348 treatments	101

List of Figures



Figure 1. Summary of Findings (SoF) table	18
Figure 2. Network map of the illustrative example	24
Figure 3. Cumulative rankograms and ranking for the illustrative example	29
Figure 4. Clustered ranking plot for the illustrative example.....	30
Figure 5. Probabilities and the corresponding entropy with base two logarithms for binary outcome	32
Figure 6. Maximum and minimum of Entropy value in different number of treatments	42
Figure 7. The probabilities distribution for the maximum value of (A) entropy and (B) variance.....	46
Figure 8. Rankograms for the (A) Example 1 and (B) Example 2.....	62
Figure 9. Simulation results for the relationship between Normalized Entropy and the width of 95% CI of SUCRA for top 3 ranked treatments within network with 3 to 10 treatments	69
Figure 10. Range of Normalized Entropy corresponding to the probability of being the best equal to (A) 0.5, (B) 0.7 or (C) 0.9 for network with 3 to 25 treatments included	72
Figure 11. Flowchart of empirical dataset selection	74
Figure 12. The distribution of ranking uncertainty levels for 157 published NMAs	76
Figure 13. The association of ranking uncertainty in Normalized and the range of 95% CI of SUCRA for each treatment of 156 published NMAs by number of treatments included in the NMA	78
Figure 14. Rankings for two outcomes with uncertainty in 5 levels	82
Figure 15 (A) SUCRA-based rank with uncertainty of 6 treatments (B) 4 treatments (C) 2 treatments for example	83

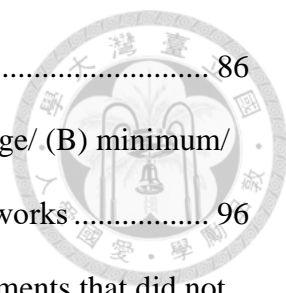


Figure 16. Flowchart of the study selection process 86

Figure 17. Scatter plots of average normalized entropy and (A) average/ (B) minimum/
 (C) maximum quadratic weighted Cohen’s kappa for 60 networks 96

Figure 18. Scatter plot of normalized entropy and percentage of treatments that did not
 change rank for 348 treatments within 60 NMAs 100

Figure 19. A hypothetical example for the relationship between precision of estimates
 and uncertainty of ranking 105

CHAPTER 1: INTRODUCTION

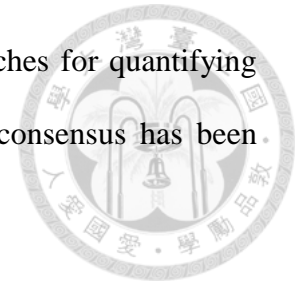


In recent years, evidence from network meta-analysis (NMA) has been widely used to support recommendations in guidelines for disease management¹⁻³. This methodology incorporates both direct and indirect evidence to simultaneously estimate relative effects treatments within the network, even for treatment comparisons without head-to-head trials⁴⁻⁶. NMA provides evidence on the selection of the best treatment strategy, but the interpretation of its results may sometimes be challenging⁷⁻⁹, especially when the differences between active treatments are likely to be small and statistically non-significant.

Therefore, ranking treatments are proposed to facilitate the interpretation of comparative effectiveness and to support clinical decision making¹⁰, which is often determined by the mean rank or the surface under the cumulative ranking curve (SUCRA), a summary index for ranking probabilities. As ranking can always be obtained from NMA, it is important to know how the uncertainty of the ranking is and how likely the ranking is to be altered by new evidence.

However, the lack of reporting uncertainty of treatment ranking for NMA has been a great concern¹¹. According to a review of 121 published NMA studies, 52 studies reported treatment ranking, but only 9 report the uncertainty of treatment ranking¹². Therefore, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-NMA guideline and Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) working group suggested that the uncertainty of ranking for each treatment should be reported, either by rankogram (i.e., the distribution of the ranking probabilities), the credible/confidence intervals (CIs) of the mean rank, or the

interquartile range (IQR) of the median rank^{13,14}. Different approaches for quantifying the uncertainty of ranking have not yet been compared, and no consensus has been reached on how to quantify uncertainty.



An empirical study, reviewing 58 published NMA studies, used the 95% CI of SUCRA and the mean rank to quantify the uncertainty of ranking and found a substantial degree of uncertainty in treatment ranking¹⁵, raising doubts about the level of evidence from NMA¹⁶. Their results showed that although the ranking of treatments is a useful presentation, it may give rise to naïve interpretation of NMA results, such as small differences between treatments being exaggerated without taking the uncertainty of the ranking into consideration¹⁷.

In addition, the uncertainty of ranking^{18,19} and robustness of ranking²⁰ are two concepts related to the reliability of ranking. The uncertainty of ranking can be visualized by the distribution of ranking probabilities of a treatment. The more concentrated the ranking probabilities, the lower the uncertainty of ranking is. On the other hand, the robustness of ranking measures how sensitive the ranking is to subtle alterations of a dataset. The approach proposed to evaluate the robustness of ranking empirically is to remove one trial and then determine how the ranking would change²⁰. When the agreement between the two rankings derived from the complete dataset and the modified dataset with one trial removed is high, the treatment ranking of an NMA is considered robust.

Both the uncertainty and robustness of ranking have been applied to evaluating the reliability of treatment ranking^{18,20}, but whether these two approaches would yield similar conclusions on the reliability of ranking has not yet been fully explored. One study examined the association between uncertainty and robustness of ranking²⁰.

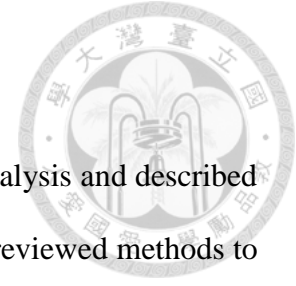
However, that study only analyzed two NMAs, which is too small to generalize.

In this dissertation, I reviewed different approaches for measuring uncertainty of ranking and proposed an alternative method, Normalized Entropy, to measure the uncertainty of treatment ranking in NMA. The real examples and simulations were used to demonstrate the advantages of using Normalized Entropy over the current approach, including rankogram, the 95% CIs of SUCRA, and mean rank, and to provide general rules for evaluating the uncertainty of ranking.

An empirical study was also conducted on the relationship between the uncertainty and the robustness of treatment ranking by using a database of NMAs. These two concepts are often presented in NMA results to show the reliability of ranking; however, they are rarely both reported and compared in an NMA. Therefore, in this dissertation, whether the high robustness of treatment ranking is associated with low uncertainty or whether they are two independent concepts were investigated.

This dissertation is organized into six chapters. Chapter 1 is the introduction and overview. Chapter 2 is the literature review, including a brief description of the development of NMA, core models and assumptions of NMA, ranking algorithms used in NMA, and current approaches to quantifying the uncertainty and robustness of ranking. Literature about the information theory, which is applied to measure the uncertainty of treatment ranking in this dissertation, is also reviewed. Based on the review, the current challenges and the aim of this dissertation are described in Chapter 3. The proposed method, normalized entropy, and the methods used to compare the performance of different methods are presented in Chapter 4. In Chapter 5, results are shown and described. Finally, the discussion and conclusion are in Chapter 6.

CHAPTER 2: LITERATURE REVIEW



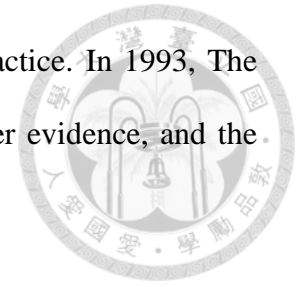
In this chapter, I first reviewed the development of network meta-analysis and described the core models and assumptions for network meta-analysis. I then reviewed methods to rank treatments within a network meta-analysis and current approaches to quantifying the uncertainty and robustness of ranking. Finally, I explained the rationale of the information theory, which provides the basis of the proposed method for evaluating the uncertainty of ranking.

2.1 Development of Network Meta-Analysis

Karl Pearson is the first to come up with the idea of meta-analysis²¹. He summarized the association between inoculation and the incidence/mortality of typhoid fever. He used data from six studies, one from India and the others from Africa. He estimated the relative risk (RR) of each study and then pooled the results by using the unweighted average. The result is considered the first research with the concept of meta-analysis, albeit this approach is different from the current standard method. Later in 1976, Glass coined the term meta-analysis for the statistical combination of results from two or more separate studies²². At that time, the meta-analysis research was mostly in the field of education or psychology.

In the 1980s, Archie Cochrane advocated the evidence-based medicine movement. He said that every patient should be cared for with the most effective intervention, which is based on the well-designed evidence, such as randomized controlled trials (RCTs)²³. Since RCTs are experiments on humans, the sample sizes of trials are usually not too large. Using meta-analysis to pool results from different studies can therefore increase the statistical power and broaden the evidence. Since then, systematic review and

meta-analysis have become an important foundation in clinical practice. In 1993, The Cochrane Collaboration was established for the goal to make better evidence, and the organization was named after Archie Cochrane.



Pairwise meta-analysis was proposed to integrate several studies for the same direct comparison of two treatments. However, most clinical trials compared new treatments with the placebo or standard care, and the head-to-head trials are limited or absent. Therefore, network meta-analysis (NMA), or called multiple treatment comparison (MTC), has been proposed to compare more than two treatments included within the network for the same condition. This method not only combines direct and indirect comparison to strengthen evidence but also can estimate the relative effects for those comparisons without direct evidence^{4,24-30}.

Computing the difference between two treatments by using indirect comparisons was first proposed by Bucher et al. in 1997³¹. His method preserved the power of randomization, but he didn't pool the direct and indirect evidence together because he believed that summary estimates should still base on direct comparisons whenever it is available. On the contrary, Higgins and White proposed that using indirect comparisons, which they called external evidence of direct comparison, can lead to more precise estimations³². In 2002, Lumley presented a linear mixed model to combine direct and indirect evidence and coined the term network meta-analysis²⁴. He mentioned that if the indirect evidence is consistent with direct evidence, then the results should be combined, and therefore, the uncertainty would become less. However, his method is not available for trials with three or more arms. Lu and Ades proposed a hierarchical Bayesian method that can include multi-arm trials in 2004⁴, and White et al. also proposed a corresponding frequentist method in 2012 and wrote a STATA package in 2015^{33,34}.

These two approaches are most widely used for network meta-analysis today.



2.2 Models and Assumptions for Network Meta-Analysis

Both Bayesian and frequentist statistical models have been developed for network meta-analysis. The only difference between the two frameworks is that the Bayesian approach could set prior distribution for estimating parameters. In 2014, there was a systematic review for published network meta-analysis. They found that among the 121 network meta-analyses published before 2012, 75% of them used a Bayesian approach³⁵. The computer code for the Bayesian network meta-analysis by using the free software WinBUGS has been available on the website, so it has become popular at the beginning. However, WinBUGS is not a user-friendly tool for people who are not familiar with Bayesian analysis, and it often stuck without showing the reasons.

After Ian White implemented the network package into Stata software for conducting NMA in 2015^{6,36}, the frequentist method has gained greater popularity. Although Bayesian method has the strength to add the prior knowledge into the analysis, most people used non-informative prior when conducting Bayesian network meta-analysis, and this usually results in a negligible difference between the results from the Bayesian framework or frequentist framework.

The netmeta package in R can also conduct the frequentist network meta-analysis³⁷. The difference in their approaches to network meta-analysis between STATA and R is that the options for heterogeneity variance estimation under the random effect model are different. The method of moments estimator (MM) is used in R, while both the maximum likelihood estimator (ML) and the restricted maximum likelihood estimator (REML) are available in the Stata software.



In this section, I will introduce the model and the general assumptions for network meta-analysis, irrespective of whether the Bayesian or frequentist approach is being undertaken.

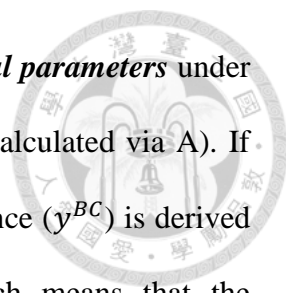
2.2.1 Model for Network Meta-analysis

The model described here uses the “contrast-based” model, which means that the effects are estimated by using treatment contrast. The model includes parameters for estimating the heterogeneity (variation of treatment effects between studies) and inconsistency (variation of treatment effects between different kinds of treatment combinations, which is usually called “design”). Consider that we have a network with T treatments, including A, B, C..., etc. Let $d=1, 2, 3... D$, which are the designs (the set of treatments compared within the study, such as AB, ABC, BCD...) within the network. The effect measure of a treatment contrast is $y_{d,i}^{AJ}$, which may be a mean difference or log odds ratio of treatment contrast of treatment J to treatment A, for i^{th} study in the d^{th} design. The general network meta-analysis model^{6,34} is

$$y_{d,i}^{AJ} = \delta^{AJ} + \beta_{d,i}^{AJ} + \omega_d^{AJ} + \varepsilon_{d,i}^{AJ}, \quad J = B, C, \dots, T$$

A is chosen as the reference treatment in this formula. The meaning of each dependent variable is described below. δ^{AJ} represents the summary effect of treatment contrasts between J and A; $\beta_{d,i}^{AJ}$ represents the heterogeneity in the J-A contrast between studies within designs; ω_d^{AJ} represents the inconsistency in the J-A contrast between designs; $\varepsilon_{d,i}^{AJ}$ is a within-study error term.

The model will estimate the $T - 1$ *basic parameters* first, such as the summary effect of each treatment contrast: $y^{AB}, y^{AC}, y^{AD}, \dots, y^{AT}$ etc. After the basic parameter is



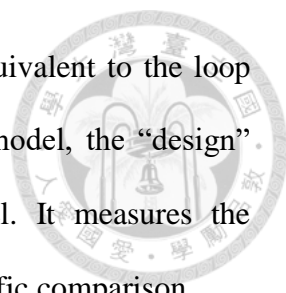
estimated, they use these basic parameters to calculate the *functional parameters* under the consistency model. That is, $y^{BC} = y^{AC} - y^{AB}$ (B versus C is calculated via A). If there is direct evidence from y^{AB} and y^{AC} , then the indirect evidence (y^{BC}) is derived from y^{AB} and y^{AC} based on the transitivity assumption, which means that the common comparator A is transitive the effect to which it is linked (y^{AB} and y^{AC}).

2.2.2 Assumptions of Network Meta-Analysis

The key assumptions of NMA are homogeneity within each pairwise comparison in the network, transitivity among different comparisons, and consistency of direct and indirect evidence. So far, most tools have been developed to detect inconsistency of direct and indirect evidence. To examine whether the consistency assumptions are violated or not, several approaches, such as the design-by-treatment interaction models^{38,39}, loop inconsistency models⁴⁰, and node-splitting models^{41,42}, have been proposed to evaluate the inconsistency between the direct and indirect evidence within a network meta-analysis. I reviewed the basic concept of these methods.

If there is evidence from y^{BC} , y^{AC} and y^{AB} , loop inconsistency means that the effect size from direct evidence (y^{BC}) is substantially different from the effect size estimated by indirect evidence ($y^{AC} - y^{AB}$) within the same loop formed in a network. It also means that loop inconsistency only exists in the loop within a network. There is no loop inconsistency when the loop is formed by a multi-arm trial because the multi-arm trial is internally consistent. However, while there are multi-arm trials within a network, inconsistency may exist between different designs.

Therefore, the design-by-treatment interaction model is proposed by Higgins and White to measure not only loop inconsistency but also design inconsistency^{33,34}. When there is



no multi-arm trial, the design-by-treatment interaction model is equivalent to the loop inconsistency model. Among the design-by-treatment interaction model, the “design” variable is added as a study-level effect modifier in the model. It measures the inconsistency between different designs within the network for specific comparison.

Node-splitting model, proposed by Dias et al., on a treatment contrast splits direct evidence from all the remaining indirect evidence and compares the estimates of the treatment contrast given by the direct and indirect evidence⁴¹. White renamed this method as “side-splitting” because the comparison of two treatments within a network is a “side” of a network map. He also proposed a symmetrical parameterization method, which improves the accuracy and solve the problem of probable different estimates arising as data in different order⁶. Although this approach is different from the design-by-treatment interaction model, the side-splitting model has been proved as a special case of the design-by-treatment interaction model⁴².

Currently, most network meta-analysis studies evaluate the inconsistency through the above three methods. If the consistency assumption is violated, researchers need to find out the possible causes of inconsistency within the NMA. It is important to evaluate whether the NMA produces valid results. However, a lack of significant inconsistency does not necessarily prove no inconsistency, as the statistical power may be low when the uncertainty in the estimates of treatment effects is large⁴³⁻⁴⁵. Therefore, except that these assumptions cannot be violated, the uncertainty of the overall NMA is required to be evaluated. In addition, while these approaches to the evaluation of inconsistency may identify the location of the inconsistency within a network, they may not, however, identify trials that give rise to the inconsistency.

2.3 Ranking

While NMA can estimate relative effects of all pairs of treatments involved in the network, interpretation for decision-making is essential. Except for the placebo, treatments involved in the network are mostly active treatments. However, the differences among active treatments in the efficacy or harm are likely to be small, which made the difference of effect hard to achieve statistical significance though it is still meaningful to point out the difference. Researchers, therefore, develop methods to rank treatments based on the NMA results, which could provide the answer to the frequently asked question in clinics: “which treatment is the best, or the second, or the third, or so on, among all the options?”.

Ranking is a useful tool for providing a priority list which adopted in several different domains, such as website searching page, recommendation system, world university ranking, health performance ranking, voting system, ... etc. In these domains, the ranking could save searching time, create more profits for business, and strengthen resource management. Similarly, ranking treatments based on their efficacy and safety could provide a treatment hierarchy for doctors to choose treatments for patients when there are multiple options for the same condition.

In this section, I am going to review the algorithms currently used to summarize and determine the ranking of treatments within an NMA. The current methods for evaluating the uncertainty of ranking will also be reviewed. Then, issues about using ranking for decision, and current challenges on interpreting and presenting the treatment ranking will both be discussed. Factors affecting treatment ranking will also be discussed. Some guidelines for presenting ranking and its uncertainty are then described.

Finally, an example is used to demonstrate how an NMA use ranking to report their results.



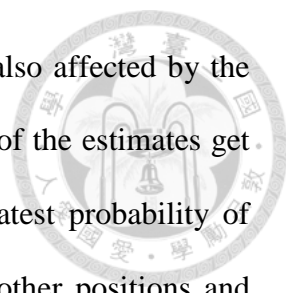
2.3.1 Ranking Algorithms

Generally, taking point estimates as constants to rank is the simplest ranking algorithm. However, the precision of point estimates would be neglected when the rank is simply according to the point estimates. Here, for NMA, researchers used ranking probabilities to take the precision of estimates into consideration. I am going to review the following methods for summarizing ranking, including ranking probabilities, mean rank, SUCRA¹⁰, and P-score⁴⁶.

2.3.1.1 Ranking probabilities

In 1998, Marshall and Spiegelhalter proposed a simulation method for obtaining ranking probabilities and intervals for ranks⁴⁷. This method randomly selects a draw from each estimate and then ranks within each draw. After carrying out a 10,000 times simulation, they identified the 5000th as median rank and 250th and 9750th as 95% confidence interval for rank.

While NMA could be conducted by Bayesian or frequentist approach, the simulation method for mean and confidence interval of ranking could proceed under both approaches. Within the Bayesian approach, the probabilities of being the best and other positions can be obtained from the posterior distributions derived from Markov Chain Monte Carlo simulations⁴⁷. Within the frequentist framework, ranking probabilities can be obtained from simulations, using the variance-covariance matrix of the estimates of treatment differences as the parameters.



In addition to differences in effect sizes, ranking probabilities are also affected by the precision of the estimates of these differences. When the variances of the estimates get smaller, the ranking probabilities are more likely to have the greatest probability of being at just one position and small probabilities of being at the other positions and therefore become more informative.

Some NMA studies use the probability of being the best ($P(\text{best})$) to rank treatments. However, it has been warned that “ranking of treatments based solely on the probabilities for each treatment of being the best should be avoided”³⁶, because the variation of the estimates would be neglected if the probability of being the best treatment is used to determine the ranking. In addition, a treatment may have a very high probability of being the second-best treatment while the probability of being the best is very low.

The difference between using probabilities of being the best and taking the probabilities of other positions into consideration is similar to the difference between plurality voting and preference list voting. Plurality voting is that every voter can only be allowed to vote for one option. Preference list voting is that every voter can submit a preference list that lists the options in a ranked order⁴⁸. To construct ranking probabilities, it requires a preference list of treatments in each simulation and summarizes the probabilities based on 1000 draws. In the voting system, both systems have their pros and cons, but in the NMA, taking the probabilities of other positions into consideration is obviously a more robust way to determine to rank.

2.3.1.2 Mean rank

Mean rank, as the name implies, is equal to the ranking probabilities of each rank order

to multiply the rank order itself for each treatment. The lower the mean rank means, the better the treatment. The formula for mean rank is written as:

$$\text{mean rank}(i) = \sum_{k=1}^n k \times p(j = k),$$

the ranking probabilities $p(j = k)$ for each treatment j in the $k = 1, 2, \dots, n$ position.

2.3.1.3 SUCRA

In order to summarize the ranking probabilities of each treatment, the cumulative ranking probabilities from the first rank for each treatment are used. The graph displayed the curve of cumulative ranking probabilities, and the surface under this curve is defined as SUCRA.

To calculate SUCRA, the ranking probabilities $P(j = k)$ for each treatment j in the $k = 1, 2, \dots, n$ position, which is summed to 1, within a network with n treatments, are used. The cumulative ranking probabilities, $\text{cum}(p(j = k))$, for each treatment j ranked as the b^{th} best or better, is used to calculate SUCRA, which is a summary value for the surface under the cumulative ranking probabilities curve. The formula for SUCRA is written as:

$$\text{SUCRA}(i) = \frac{\sum_{k=1}^{n-1} \text{cum}(p(i=k))}{n-1}.$$

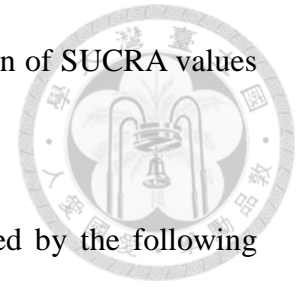
The treatment with a larger SUCRA value has a higher rank.

Therefore, SUCRA would be 1 when a treatment is 100% certain to be the best, and SUCRA would be 0 when a treatment is 100% certain to be the worst. The treatment with a larger SUCRA value has a higher rank.

Note that before calculating the ranking probabilities, whether a large effect size is



better or worse needs to be determined. Otherwise, the interpretation of SUCRA values would be contrary to the expatiations.



The relationship between mean rank and SUCRA can be presented by the following formula: mean rank = $n - (n - 1) \times \text{SUCRA}$, where n is the total number of rank. Since mean rank and SUCRA have the equivalent formula, the ranking given by the mean rank or SUCRA would be totally the same.

2.3.1.4 P-score

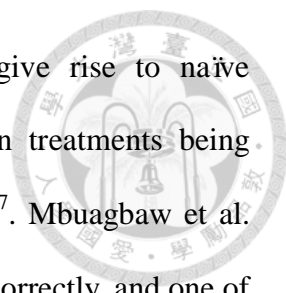
In 2015, Rucker et al. proposed P-score, which is also equal to SUCRA, but it can retrieve from the frequentist framework without conducting the simulation (the resampling method)⁴⁶. They calculated the one-side p-value of rejecting the null hypothesis for every pair of treatments, and then took the average of the p-value of each paired comparison for the specific treatment to get P-score. The formula for P-score is as below.

$$\bar{P}_i = \frac{1}{n-1} \sum_{j, j \neq i}^n P_{ij}$$

P_{ij} is the certainty that the effects of treatment i is greater than that of any other treatment j .

Therefore, mean rank, SUCRA, and P-score are ranking treatments in the same way. Again, the ranking derived by the P-score would be the same as it is determined by mean rank or SUCRA.

2.3.2 Uncertainty of Treatment Ranking



Although ranking treatments is a useful presentation, it may give rise to naïve interpretation of NMA results, such as small differences between treatments being exaggerated without taking the uncertainty of ranking into account¹⁷. Mbuagbaw et al. pointed out five reasons that ranking may mislead if not interpreted correctly, and one of them is that while the ranking is based on the low certainty evidence, the ranking is untrustworthy. Actually, while Salanti et al. proposed SUCRA rank treatments, they suggested calculating the uncertainty of ranking¹⁰, but few NMAs followed their guidance.

In this section, I will review current methods for presenting the uncertainty of ranking, including rankogram and 95% CI of SUCRA.

2.3.2.1 Rankogram

Ranking probabilities of each treatment have been used to rank treatments of a network meta-analysis. The ranking probabilities $P(j = k)$ for each treatment j in the position $k = 1, 2, \dots, n$ are summed to 1 within a network with n treatments. The graphical presentation of ranking probabilities of each treatment is known as rankogram. When the ranking probabilities are distributed more evenly, then the uncertainty is high; on the other hand, if the ranking probabilities are more concentrated in the specific rank, the uncertainty of ranking is low.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-NMA guideline¹³ suggested that the uncertainty of ranking for each treatment should be reported by using a rankogram. However, it is challenging to know the extent of uncertainty by looking at the ranking probabilities (or rankogram) when the number of treatments is large⁴⁹.

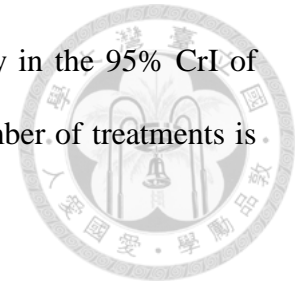
2.3.2.2 95% CI of SUCRA

The other approach used to evaluate the uncertainty of ranking is 95% credible interval (CrI) /confidence interval (CI) of mean rank or SUCRA^{15,50}. Under the Bayesian approach, the 95% CrI of mean rank and SUCRA can be obtained; under the frequentist approach, the 2.5 and 97.5 quantile from 1000 times draws of SUCRA and mean rank is used to present the 95% CI. The wider the range of 95% CrI/CI of SUCRA and mean rank represents higher uncertainty.

The 95% CI of SUCRA is obtained by simulation. For example, in a network with three treatments, labeled as A, B, and C, the estimates of their relative effects are obtained with the 95% CI for B versus A and C versus A from an NMA. Suppose that the absolute effect size of A is 0, and 1000 draws were randomly selected to compute the effect sizes of B and C from the distributions of the mean differences between them and A. In each draw, the ranking of A, B, and C can then be obtained, and SUCRA for each treatment can be computed as well. Based on 1000 draws, the 2.5 and 97.5 percentiles of SUCRA are considered the 95% CI of SUCRA. The width of 95% CI of SUCRA has been interpreted as an index for the uncertainty of treatment ranking; the greater the width, the greater the uncertainty is.

Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Working Group also recommends presenting 95% credible interval of mean rank to report in the Summary of Findings (SoF) tables in the latest version published in 2019^{9,51-55} (Figure 1). The difficulty with using 95% CrI of mean rank to represent the uncertainty of treatment ranking is that there are no consistent criteria to define whether the uncertainty is high or low. The criteria vary with the number of treatments. For

example, when the total number of treatments is 3, the uncertainty in the 95% CrI of mean rank from 1 to 2 can be considered high. But if the total number of treatments is 10, the range from 1 to 2 may be considered low uncertainty.



An empirical study, reviewing 58 published NMA studies, used 95% CI of SUCRA and the mean rank to quantify the uncertainty of ranking and found a substantial degree of uncertainty in treatment ranking⁵⁶, raising doubts on the level of evidence from NMA¹⁶. Recommendations based on the evidence with high uncertainty should be acted upon differently from those with low uncertainty⁵⁷. The importance of reporting the uncertainty of ranking has been recognized⁵⁶; however, no consensus has yet been reached on how to quantify the uncertainty.

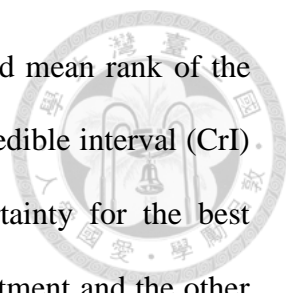
Figure 1. Summary of Findings (SoF) table



Bayesian NMA-SoF table

BENEFITS							
Estimates of effects, credible intervals, and certainty of the evidence for chemoprevention of colorectal cancer in individuals with previous colorectal neoplasia							
<p>Patient or population: Individuals with previous colorectal neoplasia</p> <p>Interventions: Low and high dose aspirin, nonaspirin non-steroidal anti-inflammatory drugs (NSAIDs), calcium, vitamin D, folic acid</p> <p>Comparator (reference): Placebo</p> <p>Outcome: Prevention of advanced neoplasia; range of follow up between three to five years</p> <p>Setting: Outpatient</p>							
Total studies: 21 RCT Total Participants: 12088	Relative effect** (95% CrI)	Anticipated absolute effect*** (95% CrI)			Certainty of evidence	Ranking**** (95% CrI)	Interpretation of Findings
		Without intervention	With intervention	Difference			
● Aspirin + calcium + vitamin D (1 RCT; 427 participants)	OR 0.71 (0.18 to 2.49) Network estimate	74 per 1000 [†]	53 per 1000	21 fewer per 1000 (61 fewer to 110 more)	⊕⊕○○ Low Due to Imprecision ^{†,‡}	3 (1 to 10)	-
● Calcium + vitamin D (1 RCT; 1028 participants)	OR 0.91 (0.52 to 1.63) Network estimate	74 per 1000 [†]	67 per 1000	7 fewer per 1000 (36 fewer to 47 more)	⊕⊕○○ Low Due to Imprecision ^{†,‡}	6 (1 to 10)	-
● Aspirin + folate (2 RCT; 916 participants)	OR 0.73 (0.43 to 1.19) Network estimate	74 per 1000 [†]	54 per 1000	20 fewer per 1000 (42 fewer to 14 more)	⊕⊕○○ Low Due to Imprecision ^{†,‡}	4 (2 to 8)	-
● Aspirin, high dose (3 RCT; 917 participants)	OR 0.81 (0.50 to 1.28) Network estimate	74 per 1000 [†]	60 per 1000	14 fewer per 1000 (37 fewer to 21 more)	⊕⊕○○ Low Due to Imprecision ^{†,‡}	5 (2 to 9)	-

*Source of this figure: Yepes-Nuñez, Juan José, et al. "Development of the summary of findings table for network meta-analysis." Journal of Clinical Epidemiology 115 (2019): 1-13.



In 2016, an empirical study explored the uncertainty of SUCRA and mean rank of the top three best treatments within 58 NMAs by presenting the 95% credible interval (CrI) of SUCRA and mean rank¹⁵. To summarize the extent of uncertainty for the best treatments, the overlapped of 95% CrI of SUCRA between best treatment and the other treatments is used; to evaluate the extent of the uncertainty of the whole NMA, they used the proportion of ranking probability of being the best exceed 50%, 75%, and 85%. They showed that more half the best-ranked interventions have no difference between the second, third, or fourth intervention. In addition, 27.6% of 58 NMAs do not have P(best) larger than 50%. Therefore, treatment ranking derived from NMA studies have a substantial degree of imprecision. Some researchers, therefore, questioned the appropriateness of taking NMA as the highest level of evidence.

In this empirical study, 21 out of 58 NMAs had at least one top-three best intervention with 95% CrI of SUCRA ranging from 0% to 100%, and hence the approach they used to present uncertainty of ranking has been criticized as uninformative⁴⁹. Therefore, it is hard to say whether the uncertainty of ranking among NMAs is really high or just because the method they used is easier to derive wide intervals.

In addition, despite the emphasis on the importance of reporting uncertainty of treatment rankings, few studies conduct this evaluation. According to an NMA's methodological systematic review³⁵, they reviewed previous 121 published NMAs and found that among 52 articles reported ranking of treatments, only 9 studies (17%) reported the uncertainty of treatment ranking by the credible interval of mean rank or SUCRA or presenting all ranking probabilities for each treatment in each ranking position. While the NMA and its ranking continue to provide evidence for clinical practice, the worrisome may increase as not knowing the certainty of ranking from the

NMA.



2.3.3 Robustness of Ranking

In addition to the uncertainty of ranking, several studies used the robustness of ranking to measure how sensitive the treatment ranking is to subtle alterations in the dataset. The methods they used to measure the robustness of ranking are slightly different from each other. Mills et al. in 2013 measured the difference in estimated effects and probability of being the best by leaving one treatment out of the whole dataset at a time. Zhang et al. in 2016 measured the impact on the coefficients by leaving one trial out of the whole dataset at a time. Daly et al. in 2019 also used the leave-one-trial-out method, but they measured the agreement by Cohen's kappa between rankings from the remaining dataset and the complete dataset.

In 2013, Mills et al. used 18 network meta-analyses to explore the effects of excluding treatments from network meta-analyses. They used the Brier score, the average squared difference between the outcome measure with and without one or more treatments, to identify which treatment has the most impact on the results. They also calculated the change in treatment ranking and ranking probabilities. They found that the exclusions could make a large impact on the results.

In 2015, Zhang et al. developed average relative distance (ARD)⁵⁸, which is also analogous to cook's distance, by calculating the difference of regression coefficients to detect outliers in an NMA. ARD method evaluates the average influence of a trial on an NMA by removing a trial from the analysis and then calculating the change in the coefficients relative to the coefficients estimated by the full data. The formula is given as⁵⁸:

$$ARD_i = \frac{1}{n} \sum_{k=1}^n \left| \frac{\hat{\eta}_k - \hat{\eta}_{k(i)}}{\hat{\eta}_k} \right|,$$

where $\hat{\eta}_k$ is coefficient estimate for treatment k from the full data, and $\hat{\eta}_{k(i)}$ is the coefficient estimate for treatment k from the data without trial i .



In 2019, Daly et al. used Cohen's kappa to evaluate the agreement of SUCRA-based ranking between subset and complete dataset. They compared their method to the 95% CI width of rank but hardly have a conclusion since they only used five studies in the analysis.

Although the above three studies used different outcome assessments, they all used the leave-one-trial/treatment-out (LOTO) approach. The limitation of this LOTO approach is that while they remove one trial/treatment out of the NMA, it may lead to a disconnected network. Therefore, the NMA cannot be conducted.

While these studies explored which treatment or which trial may affect treatment ranking, threshold analysis^{59,60} use another point of view to identify how much evidence has to change before the recommendation changes and what recommendation would be. These kinds of sensitivity analysis would help the decision maker to know the reliability of these results.

2.3.4 Factors Affecting Treatment Ranking

Several studies emphasized the importance of exploring the factors that affect treatment ranking. They called for studies to explore possible factors^{49,61,62}, such as network geometry, the number of treatments and studies included, size and risk-of-bias of included studies, baseline risk variability, between-study heterogeneity, inconsistency between treatment comparisons, small-study effects, frequency of events in

dichotomous outcome data, type of outcome data, and choice of the treatment effect measure.



2.3.5 An Illustrative Example

For showing how ranking translates the NMA results for decision-making, I used the published NMA that I analyzed as an example⁶³. The NMA compared the short-term efficacy of minimally invasive treatments for adults with obstructive sleep apnea (OSA). Seventeen interventions for outcome apnea-hypopnea index (AHI) and fourteen interventions for outcome Epworth sleepiness scale (ESS) were included in the network. The network maps for these two outcomes are presented in Figure 2.

I conducted both NMA and pairwise meta-analyses for these two outcomes. The point estimates and confidence intervals for 136 pairwise comparisons for AHI and 91 pairwise comparisons for ESS from pairwise meta-analysis and NMA are presented in Table 1 and Table 2. The pairwise meta-analysis results are shown in the upper right triangle, while the results from NMA showed in the lower left triangle in Table 1 and Table 2. Those comparisons without direct evidence can only obtain the estimates of their comparative effects from the NMA.

By observing the NMA results in Table 1 and Table 2, I find that some comparisons can conclude their relative effects using statistical significance, but some cannot. For the outcome AHI, only one-third (45/136) comparisons reached statistically significant results; for the outcome ESS, about one-third (31/91) comparisons reached statistically significant results. Nevertheless, the other two-third comparisons without statistically significant results may still have clinical implications. Therefore, instead of interpreting NMA results by whether the hypothesis is true or false, providing probabilities of being

the best, the second... etc., and summarized by SUCRA or mean rank (Table 3) would be helpful for decision making.





Figure 2. Network map of the illustrative example

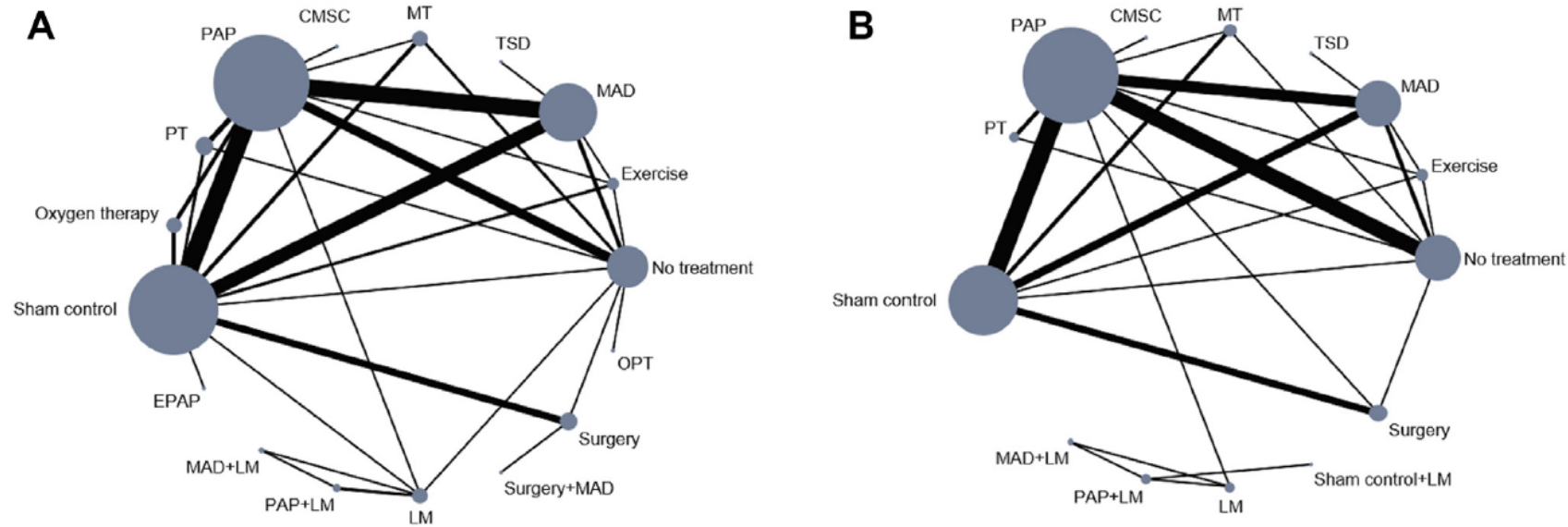


Figure 2 Networks of treatment comparisons for the reduction of (A) primary outcome apnea-hypopnea index (AHI) and (B) secondary outcome Epworth sleepiness scale (ESS) in adults with obstructive sleep apnea (OSA). The size of the nodes corresponds to the number of trials that studied the treatments. The directly compared interventions are linked with a line, the thickness of which corresponds to the number of trials that assessed the comparison. Abbreviations: CMSC, cervico-mandibular support collar; EPAP, nasal expiratory positive airway pressure; LM, lifestyle modification; MAD, mandibular advancement device; MT, myofunctional therapy; OPT, oral negative pressure therapy; PAP, positive airway pressure; PT, positional therapy; TSD, tongue stabilizing device.

*Source of this figure: Gao, You-Ning, et al. "Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: a systematic review and network meta-analysis of randomized controlled trials." *Journal of the Formosan Medical Association* 118.4 (2019): 750-765.



Table 1. Results of network meta-analysis and pairwise meta-analysis for the outcome AHI in the illustrative example

(A) Apnea-hypapnea Index														
PAP+LM	10.70 (6.05 to 15.35)													15.86 (3.92 to 27.81)
7.60	MAD+LM													11.50 (5.96 to 17.04)
3.83	-3.78	Surgery+MAD												0.95 (-2.30 to 4.20)
-15.44	-23.04	-19.26	PAP	6.89 (5.32 to 8.47)										56.00 (31.99 to 60.01)
-5.45	-13.06	-9.28	9.99	MAD										17.00 (6.45 to 27.55)
4.77	-2.83	0.95	20.21	10.22	Surgery									16.00 (4.90 to 27.10)
15.58	7.98	11.75	31.02	21.03	10.81	LM								6.70 (1.58 to 11.82)
1.56	-6.04	-2.26	17.00	7.02	-3.21	-14.02	CMSC							19.10 (8.19 to 30.01)
2.66	-4.95	-1.17	18.09	8.11	-2.12	-12.92	1.09	MT						49.20 (35.96 to 62.43)
5.43	-2.18	1.60	20.86	10.88	0.65	-10.15	3.86	2.77	EPAP					32.87 (26.86 to 38.87)
-4.45	-12.05	-8.27	10.99	1.00	-9.22	-20.03	-6.01	-7.10	-9.87	TSD				25.55 (16.70 to 34.41)
-4.83	-12.44	-8.66	10.60	0.62	-9.61	-20.41	-6.40	-7.49	-10.26	-0.39	PT			12.35 (8.95 to 15.76)
2.27	-5.34	-1.56	17.71	7.72	-2.51	-13.31	0.70	-0.39	-3.16	6.72	7.10	Exercise		2.77 (-3.69 to 9.24)
-6.26	-13.87	-10.09	9.18	-0.81	-11.04	-21.84	-7.83	-8.92	-11.69	-1.81	-1.43	-8.53	OPT	6.80 (2.83 to 10.77)
21.16	13.55	17.33	36.60	26.61	16.39	5.58	19.59	18.50	15.73	25.61	25.99	18.89	27.42	Oxygen therapy
12.23	4.62	8.40	27.66	17.68	7.45	-3.35	10.66	9.57	6.80	16.67	17.06	9.96	18.49	-8.93
7.84	0.23	4.01	23.28	13.29	3.06	-7.74	6.27	5.18	2.41	12.29	12.67	5.57	14.10	-13.32
														-4.39
														-8.74
														to -0.04

*Source of this figure: Gao, You-Ning, et al. "Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: a systematic review and network meta-analysis of randomized controlled trials." Journal of the Formosan Medical Association 118.4 (2019): 750-765.

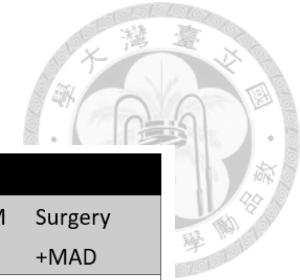
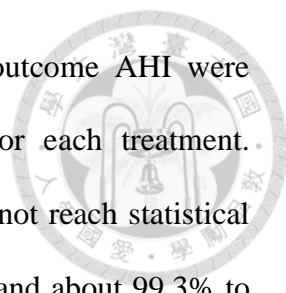


Table 3. Ranking probabilities for the outcome AHI

Ranking	Treatment																
	No treatment	Sham control	PAP	MAD	Surgery	LM	CMSC	MT	EPAP	TSD	PT	Exercise	OPT	Oxygen therapy	PAP+LM	MAD+LM	Surgery +MAD
Best	0	0	87.2	0	0	0	1.2	0	0	4.9	0	0	5.9	0	0.7	0	0.1
2nd	0	0	12.1	9.2	0	0	7.1	0	0.8	21.5	10.7	0.2	28.9	0	6.9	0.6	2
3rd	0	0	0.7	27.6	0	0	6.6	0.3	1.4	14.2	19.7	0.9	16.8	0	7.5	1.2	3
4th	0	0	0	31.8	0.1	0	5.7	1	2.1	9.5	24.2	2.1	11.2	0	7.3	1.6	3.5
5th	0	0	0	20.7	0.5	0	7	3.3	3	10.9	22.4	4.6	11.4	0	8.8	2.2	5.2
6th	0	0	0	8.1	1.7	0	9.5	8.7	5.6	10.2	13.8	10.2	8.4	0.1	12.3	3.7	7.6
7th	0.1	0	0	2.2	5.4	0	8.8	13.6	7	7.8	6.2	15	6.4	0.1	12.6	5.5	9.3
8th	0.8	0	0	0.3	10.4	0.1	8.1	17.8	8.6	5.7	2.1	16.1	4.3	0.2	10.2	6.3	9.1
9th	2.7	0	0	0	15.8	0.2	7.3	17.5	8.7	4.4	0.7	15.2	2.7	0.2	8.3	6.5	9.7
10th	6.9	0	0	0	20.3	0.3	6.6	15	9.2	3.5	0.2	12.8	1.5	0.4	7	7.1	9.2
11th	15.1	0.1	0	0	19.7	0.9	6.7	10.8	9.9	2.5	0.1	9.7	1.1	0.6	6.1	7.9	8.9
12th	24.9	1.3	0	0	14.6	1.6	6.4	7.1	10.3	2.1	0	7	0.7	1	5.1	8.7	9.3
13th	27.1	8.3	0	0	8.3	3.7	5.8	3.7	11.1	1.4	0	4.3	0.4	1.8	4.3	11.8	8.1
14th	18.1	28.4	0	0	2.7	8.3	5.2	1	10	0.8	0	1.4	0.2	3.2	2.2	12	6.5
15th	3.7	41.1	0	0	0.5	17.4	3.6	0.2	6.3	0.4	0	0.3	0.1	6.7	0.7	14.7	4.4
16th	0.5	19.1	0	0	0.1	46.9	3.4	0	4.8	0.3	0	0.1	0.1	13.7	0.1	7.8	3.1
Worst	0	1.7	0	0	0	20.7	1	0	1.2	0	0	0	0	72	0	2.4	0.9
MEAN RANK	12.4	14.7	1.1	4	10.2	15.6	8.3	8.9	10.6	5.1	4.4	8.7	4.2	16.4	7.2	11.6	9.5
SUCRA	0.3	0.1	1	0.8	0.4	0.1	0.5	0.5	0.4	0.7	0.8	0.5	0.8	0	0.6	0.3	0.5

*Source of this figure: Gao, You-Ning, et al. "Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: a systematic review and network meta-analysis of randomized controlled trials." *Journal of the Formosan Medical Association* 118.4 (2019): 750-765.



In Table 3, the ranking probabilities of each treatment for the outcome AHI were presented, and both mean rank and SUCRA were calculated for each treatment. Although the relative effects of PAP compared to TSD or OPT do not reach statistical significance, PAP has 87.2% possibility to be the best intervention and about 99.3% to be either the best or the second-best intervention. Therefore, PAP would be the first treatment to recommend for improving AHI.

Based on the mean rank and SUCRA, ranking of treatments for the outcome AHI is: PAP > MAD > OPT > PT > TSD > PAP+LM > CMSC > Exercise > MT > Surgery + MAD > Surgery > EPAP > MAD+LM > No treatment > Sham control > LM > Oxygen therapy. However, not every treatment has a high probability of being in each position. For example, the second-ranked treatment MAD has more even distributed probabilities. The reason for this is that MAD and third-ranked (OPT), fourth-ranked (PT), fifth-ranked (TSD) treatments actually have less than 1 in mean difference for AHI. Therefore, using ranking to say that MAD is better than OPD, PT, and TSD may exaggerate small differences¹³ between these interventions.

Some guidelines suggested presenting ranking probabilities (Table 3) or the graph such as cumulative rankograms (Figure 3), which can display several treatments together in one graph to avoid misinterpretation. However, it is complicated to check for each treatment, especially when the number of treatments is large, just like this example. Therefore, if there is a simple index for showing the uncertainty of ranking, it would be easier to point out which ranking is reliable. Moreover, the uncertainty of ranking may provide the weight for multiple outcomes, like AHI and ESS in this example, instead of equal weight for both outcomes (Figure 4).



Figure 3. Cumulative rankograms and ranking for the illustrative example

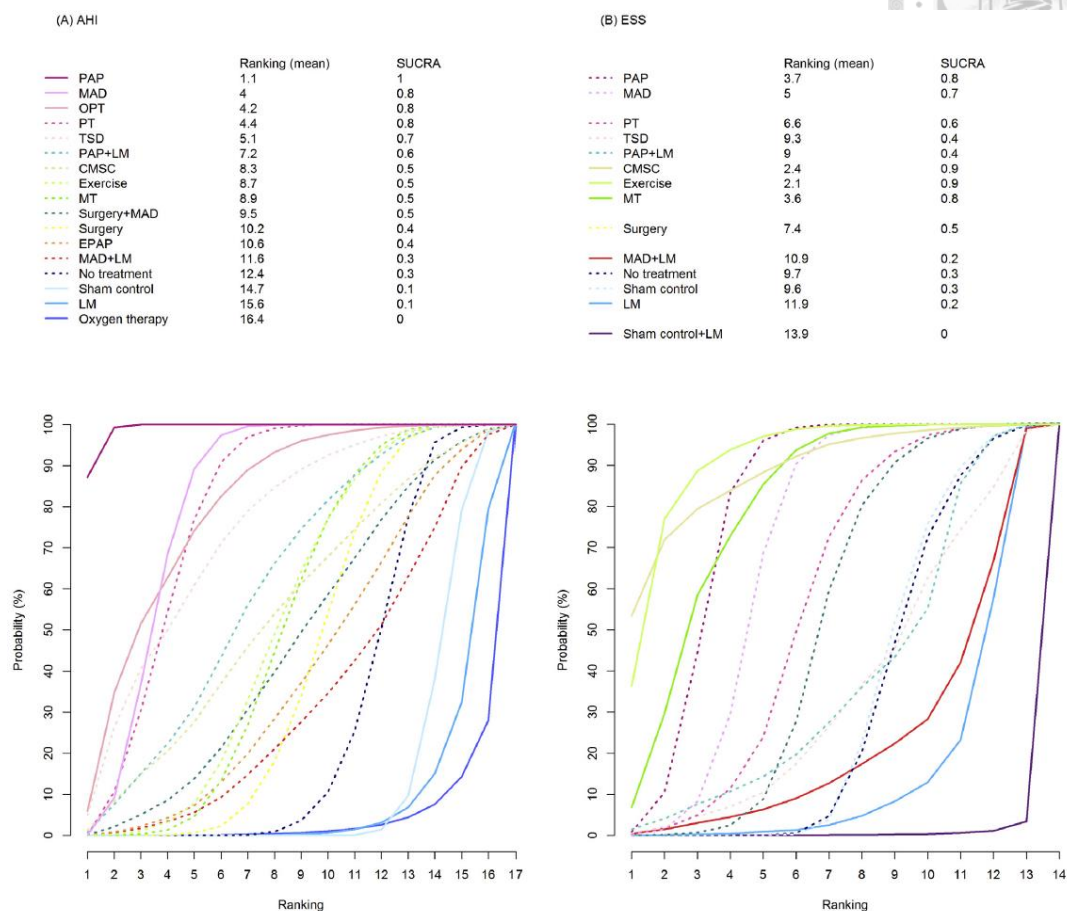


Figure 3 Cumulative rankograms: plots of the surface under the cumulative ranking curves (SUCRAs) for the reduction of (A) apnea-hypopnea index (AHI) and (B) Epworth sleepiness scale (ESS) of all interventions in the adult obstructive sleep apnea (OSA) networks. Ranking indicates the probability to be the best treatment, the second best, and so on, among the different interventions under evaluation. A larger SUCRA score indicates a more effective intervention. Abbreviations: CMSC, cervico-mandibular support collar; EPAP, nasal expiratory positive airway pressure; LM, lifestyle modification; MAD, mandibular advancement device; MT, myofunctional therapy; OPT, oral negative pressure therapy; PAP, positive airway pressure; PT, positional therapy; TSD, tongue stabilizing device.

*Source of this figure: Gao, You-Ning, et al. "Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: a systematic review and network meta-analysis of randomized controlled trials." *Journal of the Formosan Medical Association* 118.4 (2019): 750-765.



Figure 4. Clustered ranking plot for the illustrative example

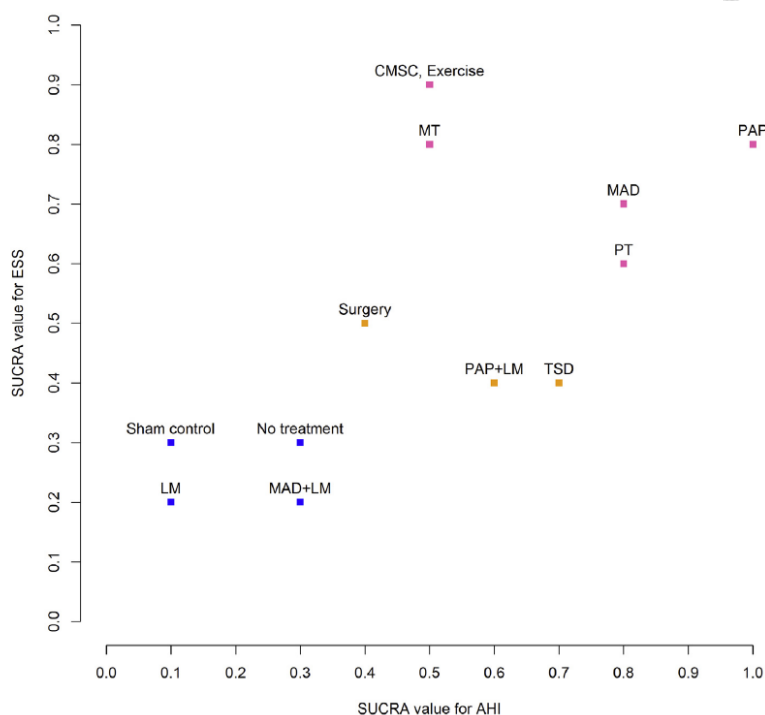
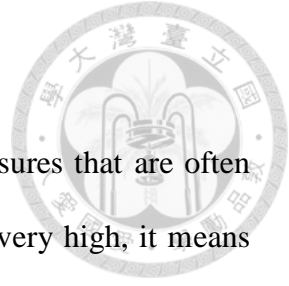


Figure 4 Clustered ranking plots of the adult obstructive sleep apnea (OSA) network based on the surface under the cumulative ranking curve (SUCRA) values for the reduction of the primary outcome apnea-hypopnea index (AHI) and the secondary outcome Epworth sleepiness scale (ESS). Each color represents a group of treatments that belong to the same cluster. Treatments lying in the upper right corner are more effective and acceptable than the other treatments. Here, the pink dots indicate the more effective interventions, while the blue dots indicate the less effective interventions. Abbreviations: CMSC, cervico-mandibular support collar; LM, lifestyle modification; MAD, mandibular advancement device; MT, myofunctional therapy; PAP, positive airway pressure; PT, positional therapy; TSD, tongue stabilizing device.

*Source of this figure: Gao, You-Ning, et al. "Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: a systematic review and network meta-analysis of randomized controlled trials." *Journal of the Formosan Medical Association* 118.4 (2019): 750-765.



2.4 Decision Making under Uncertainty

To quantify the uncertainty, probabilities and variance are two measures that are often used to describe this concept. While the probability is very low or very high, it means that it is very certain to be true or to be false. Therefore, the uncertainty is low. In the mid-20th century, Claude Shannon proposed the concept of ‘entropy,’ which calculate uncertainty from probabilities, and established the foundation of information theory. In epidemiology and biostatistics, the uncertainty is usually expressed by the variance of the measurement. For example, Nikolakopoulou et al. proposed a method for measuring pairwise meta-analysis information to monitor the precision of pairwise comparisons within living network meta-analyses^{64,65}. They defined precision as the inverse of variance from the pair comparison within network meta-analysis, and they called the precision as “amount of information”. In this definition, the smaller variance means smaller uncertainty and higher information. However, the uncertainty of ranking is not equal to the variance of estimates.

The following sections will explain the calculation of entropy-based uncertainty from probabilities. Some applications of entropy in different domains will also be reviewed. Lastly, how entropy is appraised in different fields and the criteria currently used for this index will also be discussed.

2.4.1 Entropy

Entropy was initially used in the thermodynamic system to describe the disorder or randomness in a heat or energy system in the 19th century. In 1948, Claude Shannon extended this concept to measure the impurity of the elements in a set in the communication field⁶⁶. They named this as Shannon’s entropy, which measures the

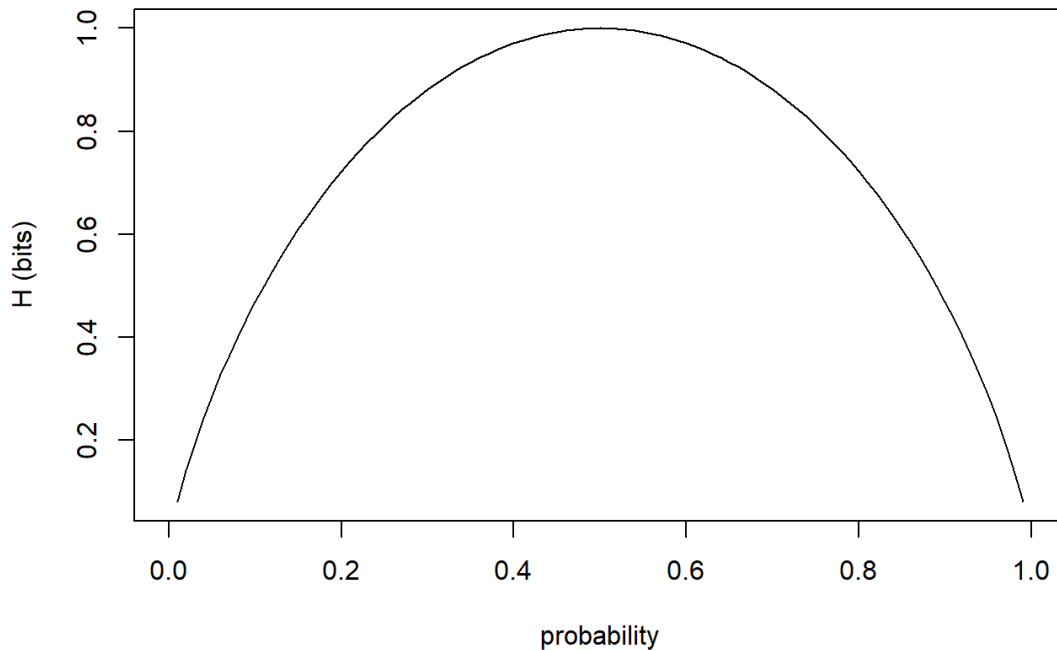
uncertainty for a set of outcomes with different probabilities. Suppose there are l possible outcomes, and Shannon's entropy formula is then given as:



$$H(t) = - \sum_{i=1}^l P(t = i) \times \log_b(P(t = i))$$

where $P(t = i)$ is the probability of the outcome equal to i . b is the base of the logarithm, and the unit of entropy would depend on the choice of b . As the base-2 logarithms is used, the entropy is measured in bits. The relationship between probability and entropy with base 2 logarithm for the binary outcome is presented in Figure 5. The greatest entropy is attained when the probability is evenly distributed. The larger entropy value means higher uncertainty and vice versa.

Figure 5. Probabilities and the corresponding entropy with base 2 logarithm for binary outcome



The normalized entropy⁶⁷ is used to give measures independent of the number of outcome l . The formula of normalized entropy is as follow,

$$H(t)^{Normalized} = \frac{H(t)}{H(t)^{maximum} - H(t)^{minimum}}$$

$$= -\frac{1}{\log_b(l)} \sum_{i=1}^l P(t=i) \times \log_b(P(t=i)).$$



Normalized entropy is that entropy divided by the range of maximum entropy and minimum entropy, which are $\log_b(l)$ and 0, respectively. Therefore, normalized entropy is between 0 and 1. If the number of outcomes is binary, and the base is 2, then normalized entropy is equal to entropy.

Several studies in different fields have ever used normalized entropy as a measure to compare the uncertainty of two systems^{68,69}.

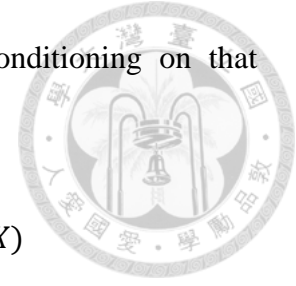
2.4.2 Applications of Entropy

Entropy has been used in many fields, including physics, economics, biology...etc. Here I took three examples, including decision tree analysis in machine learning, model selection in the latent growth curve, and classification in logistic regression to show how these fields used entropy as an index to measure uncertainty.

In machine learning, decision tree analysis is a classification method to help the process of decision making. The decision tree analysis is to divide samples in a dataset into smaller subsets with higher homogeneity of attributes. The entropy is proposed to measure the homogeneity of attributes in a subset. A lower entropy means higher homogeneity in a subset. Therefore, they compare the entropy reduction, which is information gain, to decide which attribute is the best for classification. The greater the

reduction in entropy, the greater information is gained from conditioning on that attribute X. The formula is as follows:

$$\text{Information gain} = IG(Y, X) = H(Y) - H(Y|X)$$



The attribute with the most information gain would be the preferred method to category subsets.

Similar to decision analysis in machine learning, the latent growth curve model in structural equation modeling is also used to categorize heterogeneous populations into subgroups. They group people with similar growth trajectory property together. For determining how many subgroups would be optimal to be classified in the latent growth curve model, entropy is used to show the precision of classification^{70,71}. The formula is as below.

$$E_k = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik} \ln \hat{p}_{ik})}{n \ln k}$$

where \hat{p}_{ik} is the probability that an individual i would endorse that class k . They called the measure entropy but used 1 minus normalized entropy. Therefore, the higher the value indicates better classification precision⁷².

For logistic regression, entropy is still used as a measure to evaluate how good classification is, which is still similar to the concept of using latent growth curve and decision tree analysis. However, I cannot know which correct group for each observation in those two fields, but I can know that in logistic regression. Therefore, it made it possible for comparison between classification rate and entropy.

2.4.3 Criteria of Normalized Entropy

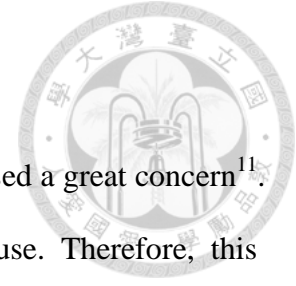
To determine how many clusters is necessary to be classified is always not an easy question. Currently, there is an existing criterion for latent growth curve analysis. Generally, they use 0.8 as a threshold for 1 minus normalized entropy⁷³, which indicates that good separation of classes when normalized entropy is lower than 0.2.

Since logistic regression can know the correct groups for each observation, entropy can be compared to classification rate or other evaluation approaches, such as the area under the ROC (receiver operating curve). Their results show that AUC is over 0.9 while normalized entropy is around and less than 0.55.

The other study suggested to categorize the 1 minus Normalized Entropy into four groups: perfect (entropy = 1), high (entropy = 0.8), medium (entropy = 0.6), and low (entropy = 0.4)⁷².

Therefore, in conclusion, from the literature review, normalized entropy below 0.4 is quite well, while it is not so well when the normalized entropy is over 0.6.

CHAPTER 3: AIMS AND OBJECTIVES

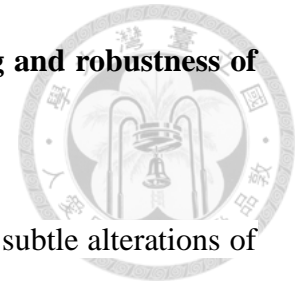


Whether the treatment ranking for NMA is trustworthy has been raised a great concern¹¹. However, the current approaches still have limitations in their use. Therefore, this dissertation aims to develop a methodology to evaluate the performance of treatment ranking from NMAs. Two objectives of this dissertation are briefly described as follows:

1. Developing an alternative method to measure the uncertainty of treatment ranking from NMA

Rankogram has been used to show ranking uncertainty. However, it is not always straightforward to compare the differences in the distribution of probabilities by inspecting rankograms. The 95% CI/CrI of the mean rank and SUCRA¹², and the interquartile (IQR) of the median rank have also been used to show ranking uncertainty. However, using these methods to compare the uncertainty of ranking between treatments within the same network or across networks may be inaccurate because the range of values given by these methods is related to the numbers of treatments within an NMA. If the uncertainty of treatment ranking can be accurately quantified, the research resource can be given priority to those treatments with highly uncertain efficacy. Thus, in the first objective of the dissertation, I proposed an alternative method, Normalized Entropy, to transform the distribution of ranking probabilities into a single quantitative measure for the uncertainty of treatment ranking in NMAs. I used simulation and empirical examples to demonstrate the strengths of Normalized Entropy as an alternative indicator for the uncertainty of ranking in NMAs.

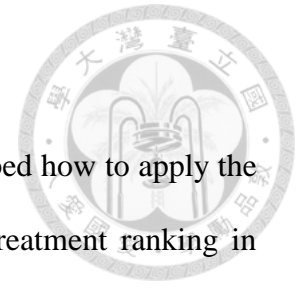
2. Explore the association between the uncertainty of ranking and robustness of ranking



The robustness of ranking measures how sensitive the ranking is to subtle alterations of a dataset. One approach proposed to empirically evaluate the robustness of ranking is to remove one trial and then to evaluate how the ranking would change.²⁰ When the agreement between the two rankings derived from the complete dataset and the modified dataset with one trial removed is high, the treatment ranking of an NMA is considered robust. Using uncertainty and robustness of ranking to evaluate the reliability of treatment ranking have been applied to published NMAs,^{18,20} but whether these two approaches would yield similar conclusions on the reliability of ranking has not yet been fully explored. In the second objective of the dissertation, I aimed to empirically investigate the relationship between the uncertainty and the robustness of treatment ranking by using a database of NMA. I would like to examine whether the high robustness of treatment ranking is associated with low uncertainty or whether they are two independent concepts.

These two research questions were written as two papers and submitted to journals for publication, which are: (1) Using Normalized Entropy to Measure Uncertainty of Rankings for NMAs, and (2) High Robustness Does Not Always Imply Low Uncertainty of Treatment Rankings: an empirical study of 60 Network Meta-analyses.

CHAPTER 4: MATERIALS AND METHODS



Materials and methods are described in five sections: First, I described how to apply the current and proposed methods for estimating the uncertainty of treatment ranking in NMA. Second, the simulations used to show the differences of using Normalized Entropy and 95% CI of SUCRA or the probability of being best (P(Best)) to quantify ranking uncertainty are presented. Third, the published NMA database, and the four selected examples, used to compare the uncertainty of ranking using Normalize Entropy and current methods are described. The other two examples also illustrate the uncertainty of ranking using Normalized Entropy. Fourth, the robustness of the ranking is compared with the uncertainty of the ranking. Lastly, the software and R code for calculating Normalized Entropy from ranking probabilities are presented in the fifth section.

4.1. Current and Proposed methods

Instead of measuring the uncertainty of ranking through the resampling process, such as 95% CI of SUCRA, I intend to estimate it directly from ranking probabilities. Normalized entropy and Euclidean distance are two indicators I have applied to evaluate the ranking uncertainty. Although I finally proposed to use Normalized Entropy as the uncertainty indicator of NMA ranking, I also described how to apply Euclidean distance to quantify uncertainty of ranking here for reference. Moreover, while variance and entropy have long been widely used in different fields to quantify the uncertainty, I discussed their differences in defining the uncertainty of ranking probabilities for NMAs.

4.1.1. 95% CI of SUCRA

The current approach to evaluating the uncertainty of ranking is to compute the 95% CI of mean rank or SUCRA¹⁵. Since mean rank and SUCRA are mathematically equivalent: $mean\ rank = n - (n - 1) \times SUCRA^{19}$, in the remaining of this dissertation I only compared the 95% CI of SUCRA to the proposed method.

To calculate SUCRA, I used the formula for SUCRA as mentioned in the Chapter 2.3.1.3:

$$SUCRA(i) = \frac{\sum_{k=1}^{n-1} cum(p(j=k))}{n-1}.$$

To derive the 95% CI of the SUCRA value, a simulation approach to deriving 95% CI of rank was used²⁰. I used the following example to explain the simulation process to obtain the 95% CI of SUCRA in an NMA.

In a network with three treatments, labelled as A, B, and C, we obtained the estimates of their relative effects with 95% CI for B versus A, and C versus A from a NMA. Suppose that the absolute effect size of A is 0, and we randomly select 1000 draws to compute the effect sizes of B and C from the distributions of the mean differences between them and A. In each draw, the ranking of A, B, and C can then be obtained, and we can also compute SUCRA for each treatment. Based on 1000 draws, the 2.5 and 97.5 percentiles of SUCRA is 95% CI of SUCRA⁷⁴. The width of 95% CI of SUCRA has been interpreted as an index for the uncertainty of treatment ranking; the greater the width, the greater the uncertainty is.

4.1.2. Normalized Entropy

Shannon's entropy was first proposed in the communication field for measuring the impurity of the elements in a set⁶⁶. Here, I propose to apply Shannon's entropy to measure the uncertainty of treatment ranking by using the ranking probabilities for each

treatment. For evaluating the degree of uncertainty over the differences in the ranking probabilities derived from the NMA, I applied Shannon's entropy formula:

$$H(x) = -\sum_{k=1}^n p(x = k) \log_b p(x = k), \quad (1)$$

where $p(x = k)$ is the probability of being ranked k for a specific treatment x in the network, b is the base of logarithm, and the unit of entropy would depend on the choice of b . When the base-2 logarithms is used, the entropy is measured in bit (short for binary digit), which is the smallest unit of data in a computer. For two treatments within an NMA, the relationship between probability and entropy is presented in Figure 5 (Chapter 2.4.1).

When the probability of being ranked first and second is either 100% or 0%, indicating that the ranking is absolutely certain, the corresponding entropy is 0, which represents the lowest level of uncertainty. When the probability is 50%, the corresponding entropy is 1 (the highest), indicating that the outcome is uncertain. While the minimum value of entropy is zero regardless of the number of treatments involved, the maximum value of entropy increases with the number of treatments. I presented the maximum and minimum value of entropy for different numbers of treatments included in the NMA in Figure 6.

Since the maximum entropy of treatment varies with the total number of treatments included in the network, we used Normalized Entropy, which rescaled entropy by dividing the range of maximum and minimum entropy for n treatments in a network⁶⁷. Therefore, Normalized Entropy ranges from 0 to 1, and is independent of the number of treatments. The formula of Normalized Entropy is given as follows:

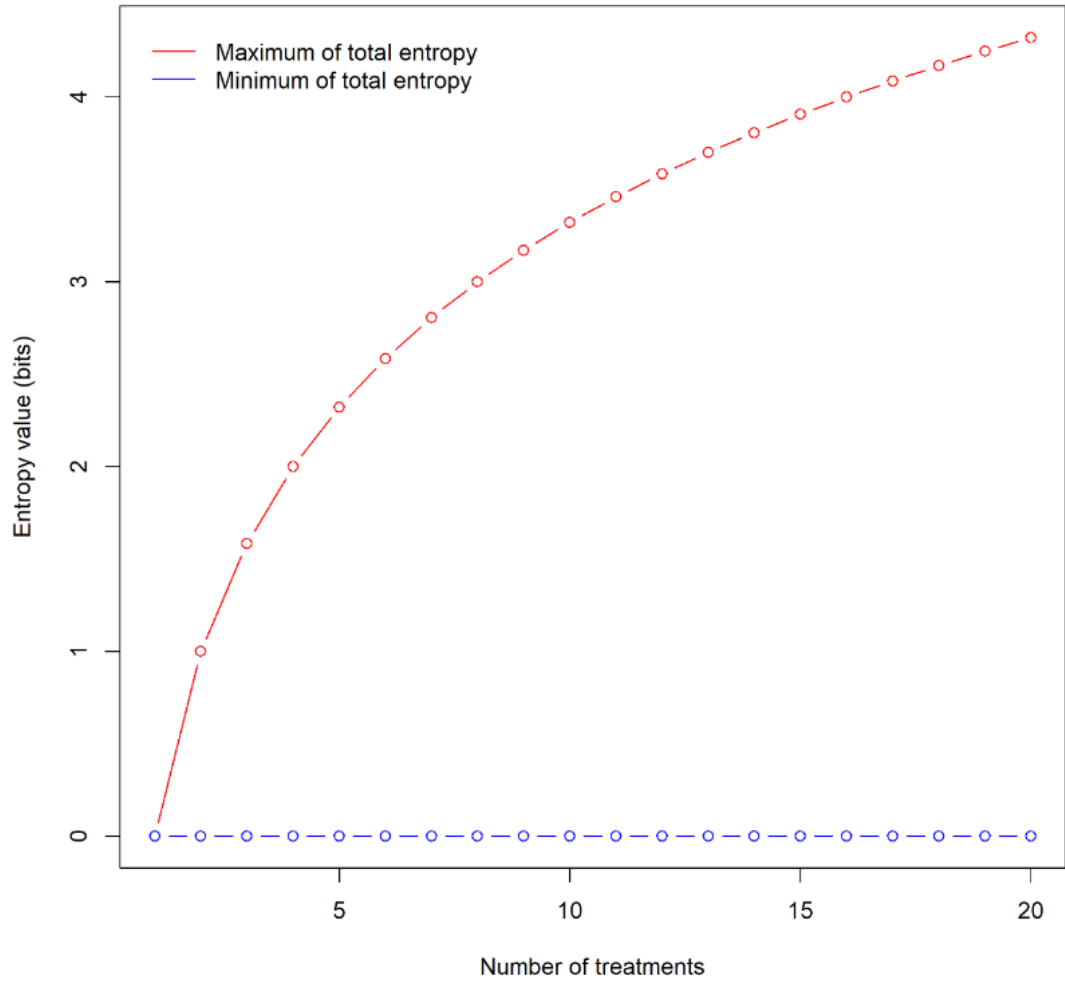
$$\begin{aligned}
& H(x)^{Normalized} \\
&= \frac{H(x)}{H(x)^{maximum} - H(x)^{minimum}} \\
&= \frac{-\sum_{k=1}^n p(x = k) \log_2 p(x = k)}{n \times \left(-\frac{1}{n} \log_2 \frac{1}{n}\right) - 0} \\
&= -\frac{1}{\log_2(n)} \sum_{i=1}^n p(x = k) \times \log_2 p(x = k).
\end{aligned}$$

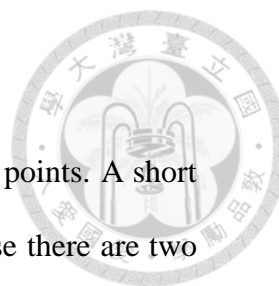


For an NMA, the most precise scenario (i.e., absolute certainty in the ranking of treatments) is that each treatment has 100% probability of being in one ranking position and 0% probability for the other positions. Under this scenario, the entropy is zero bit, and Normalized Entropy is zero. In the least precise scenario, the probability of being at each position for each treatment is equal; for instance, the probability of each position is 25% for a treatment in a NMA with four treatments. As a result, the total entropy value under this scenario is 2 bits. Then, it is divided by 2, which is the difference in the Entropy values between the most precise and least precise scenarios when there are 4 treatments, resulting in 1 (2 divided by 2) being the Normalized Entropy for each treatment.

To compare Normalized Entropy to the current approaches, we conducted two simulations to compare 95% CI of SUCRA to the Normalized Entropy and P(Best) to the Normalized Entropy. The process of two simulations were described in the following two sections.

Figure 6. Maximum and minimum of Entropy value in different number of treatments





4.1.3. Euclidean Distance

Euclidean distance measures the straight line distance between two points. A short distance of two points means that they are close in location. Suppose there are two points, $p_1 = [p_{1,1} p_{1,2} \dots p_{1,n-1} p_{1,n}]$ and $p_2 = [p_{2,1} p_{2,2} \dots p_{2,n-1} p_{2,n}]$, in an n -dimensional Euclidean space, the distance between them is given by the formula:

$$d(p_1, p_2) = \| p_1 - p_2 \| = \sqrt{\sum_{i=1}^n (p_{1,i} - p_{2,i})^2},$$

where n is the dimension of the space.

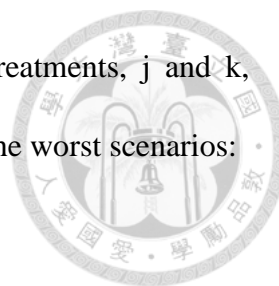
A short distance between the ranking probabilities of two treatments indicates a great similarity. In an NMA, the greater the distance between two treatments, the more divergent their ranking probabilities.

The formula for the distance between the ranking probabilities of any two treatments j and k is as below,

$$\text{Distance}_{j,k} = d(p_j, p_k) = \sqrt{\sum_{i=1}^n (p_{ji} - p_{ki})^2},$$

where n is the total number of ranking positions, i.e. the number of treatments involved in a NMA.

For any two treatments in an NMA, the best scenario is that each treatment has its own 100% probability of being in one position of the ranking and 0% probability for the other positions. Under this scenario, the distance between any two treatments is $\sqrt{2}$. In the worst scenario, the probability of being at each position for each treatment is identical, and the distance between two treatments under the worst scenario is zero. I



defined the Relative Distance as the distance between any two treatments, j and k , divided by $\sqrt{2}$, i.e., the difference in distance between the best and the worst scenarios:

$$\text{Relative Distance}_{j,k} = \frac{d(p_j, p_k)}{\sqrt{2}}.$$

Among an NMA with n treatments, each treatment has $n - 1$ relative distances to the other treatments (the relative distance of the treatment and itself would be zero).

The average relative distance of a treatment j is given as:

$$\text{Average Relative Distance}_j = \frac{\sum_{k=1}^n (\text{Relative Distance})_{j,k}}{n-1},$$

where n is the total number of treatments.

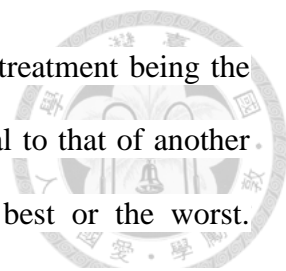
An NMA involving n treatments has n average relative distances, and the mean of n average relative distances can be used as a measure for the precision of a NMA:

$$\text{Precision in Euclidean distance} = \frac{\sum_{j=1}^n \left(\frac{\sum_{k=1}^{n-1} (\text{Relative Distance})_{j,k}}{n-1} \right)_j}{n},$$

where j and k are any two treatments and n is the total number of treatments.

4.1.4. Variance and Standard Deviation

To quantify uncertainties, it might be straightforward to come up with the variance from the statistical perspective. While calculating the uncertainty of ranking probabilities, variance considers the central location of ranking probabilities, but entropy does not (See the formula listed in Table 4). Therefore, entropy does not distinguish the difference between different distributions, such as unimodal and bimodal, with the same probability set, i.e., entropy does not take the order of ranking positions. For example,



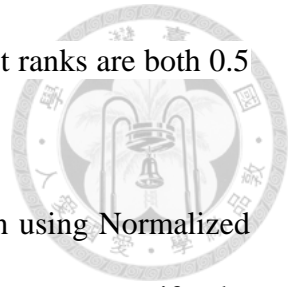
in a network with four treatments, if the ranking probabilities of a treatment being the best and the second-best are both 0.5, its entropy would be identical to that of another treatment with ranking probabilities of 0.5 for being either the best or the worst. However, their variances of ranking probabilities would be different. The variance would show that with the same probabilities set, the uncertainty of the probabilities with bimodal distribution (the second scenario in the example) is higher than unimodal distribution (the first scenario in the example). Also, as the standard deviation is the square root of the variance, the standard deviation is an index that is often used to represent a similar meaning with variance. Compared to variance, standard deviation could distinguish between different levels of uncertainty under the unimodal distribution, which is essential because the bimodal distribution of ranking probabilities is rarely seen in NMAs.

Table 4. Formula and the value range for entropy and variance

	Entropy	Variance
Formula	$H(x) = -\sum_{k=1}^n p(x = k) \log_2 p(x = k),$ <p>$p(x = k)$ is the probability of being ranked k for a specific treatment x in the network</p>	$\text{Var}(X) = E(X - \mu)^2,$ $X = P(x = i),$ $\mu = \sum_{i=1}^n P(x = i) * i$
Maximum	$\log_2 n$	$(n - 1)^2 / 4$
Minimum	0	0

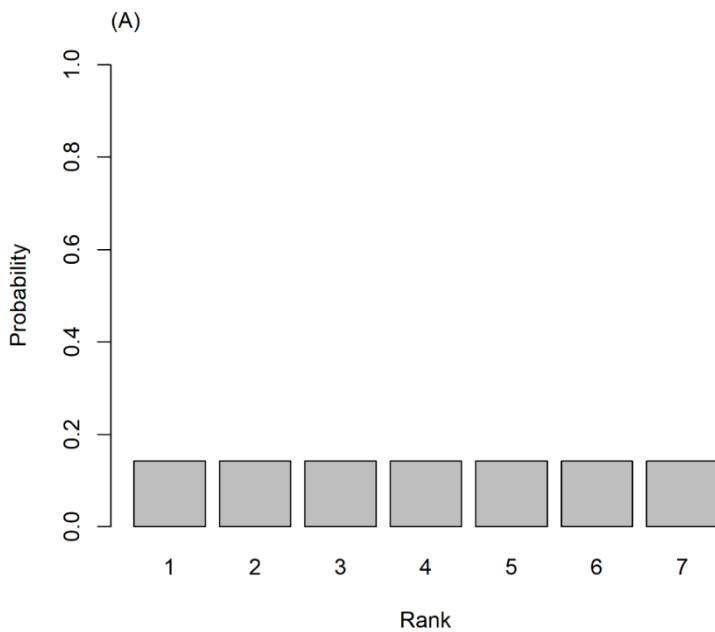
The minimum value of entropy or variance is zero when a treatment's ranking probability is 100% at one position. The maximum value of Entropy occurs when the ranking probabilities are identical for each rank (Figure 7 (A)). For Variance, the

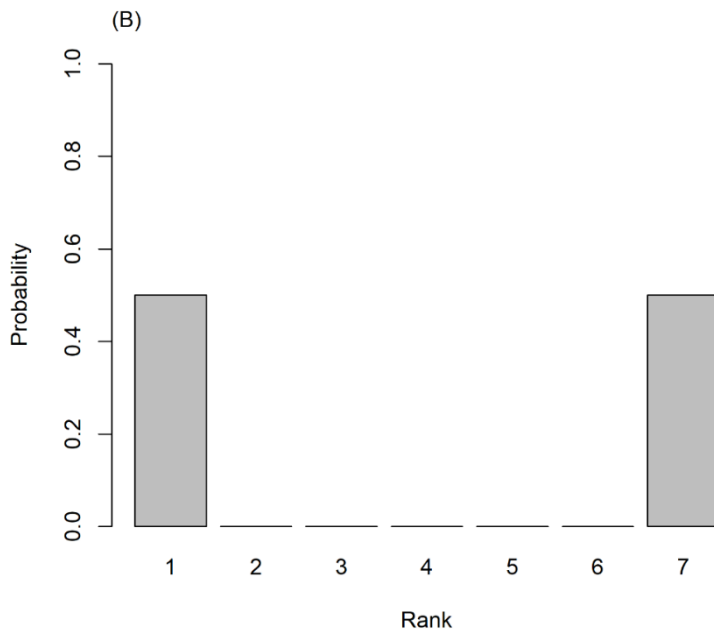
maximum value occurs when the probabilities of the first and the last ranks are both 0.5 (Figure 7 (B)).



I will use four NMA studies to demonstrate the differences between using Normalized Entropy, Normalized Variance, and Normalized Standard Deviation to quantify the ranking uncertainty of NMA.

Figure 7. The probabilities distribution for the maximum value of (A) entropy and (B) variance and standard deviation





4.2. Simulations

In the following two sections, I undertook simulations to compare Normalized Entropy to 95% CI of SUCRA and P(Best).

4.2.1. Comparing Normalized Entropy and 95% CI of SUCRA

The simulation aims to compare the performance of Normalized Entropy and 95% CI of SUCRA on measuring uncertainty of treatment ranking under scenarios with different numbers of treatments and different amounts of information within a network. I considered fully connected networks with K treatments, which are labeled as 1, 2, 3, ..., K . I simulated the contrast estimates $\hat{\theta}_{K1}$ for treatment K against treatment 1 by assuming the following multivariate normal (MVN) model:

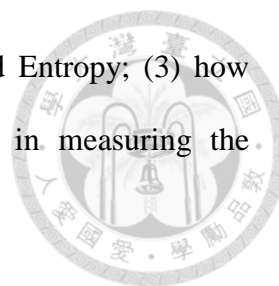
$$\begin{pmatrix} \hat{\theta}_{21} \\ \hat{\theta}_{31} \\ \hat{\theta}_{41} \\ \vdots \\ \hat{\theta}_{K1} \end{pmatrix} \sim MVN \left((\Phi^{-1}(0.975) + \Phi^{-1}(0.8)) \right. \\ \left. \times \begin{pmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \end{pmatrix}, \frac{1}{f} \begin{pmatrix} 1 & 0.5 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 0.5 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & 0.5 & \cdots & 1 \end{pmatrix} \right),$$



where f is the fraction of maximum information. I defined the maximum information ($f = 1$) as the amount of information that attains 80% power to detect a difference between each adjacent pair of treatments under a two-sided significance level $\alpha = 5\%$. For example, if only two treatments are included, when the contrast estimates $\hat{\theta}_{21}$ is drawn from $N(\Phi^{-1}(0.975) + \Phi^{-1}(0.8) = 2.80, 1)$, there is a 80% chance that I can detect the difference between two treatments. When I tune f to less than 1, it leads to a larger error variance, and, consequently, the data contains less information to detect the difference in treatment effects. This definition of maximum information and information fraction was also adopted in continuously updated NMA to monitor the trend in the results of NMA²³.

I simulated 1000 sets of contrast estimates for each combination of $K \in \{3,4,5, \dots, 10\}$ and $f \in \{0.1\%, 0.2\%, 0.3\%, \dots, 100\%\}$ to compute the 95% CI of SUCRA and Normalized Entropy. Results for the top three treatments within each network were presented in scatter plots to show the relationship between Normalized Entropy and 95% CI of SUCRA.

I then investigate (1) how the total number of treatments within a network influence the performance of these two indices; (2) how the width of the 95% CIs of SUCRA,



ranging from 0% to 100%, is related to the values of Normalized Entropy; (3) how sensitive the 95% CI of SUCRA and Normalized Entropy are in measuring the uncertainty of treatment ranking.

4.2.2. Comparing Normalized Entropy and P(Best)

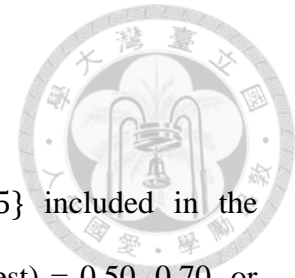
The simulation aims to investigate the relationship between the level of uncertainty measured by Normalized Entropy and the probability of being the best treatment. A previous study¹⁵ used arbitrary thresholds for the probability of being best (P(Best)), such as 0.50, 0.75, and 0.85, to define the levels of uncertainty for NMA. Despite being criticized for not considering ranking probabilities in other positions and thereby overlooking variations of the ranking probability distribution, P(Best) still serves as an intuitive index for ranking uncertainty. In this simulation, I would like to investigate the possible range of Normalized Entropy under P(Best) of 0.50, 0.70 and 0.90 to provide a relative scale of Normalized Entropy to the conventional P(Best).

Given a fixed $P(\text{Best}) = p$, the highest Normalized Entropy is attained when the uncertainty of ranking is the highest, i.e. when the ranking probabilities are evenly distributed, each with a probability of $\frac{1-p}{n-1}$. The Normalized Entropy under this scenario can be calculated as:

$$NE_{highest} = \frac{-p \times \ln(p) + \left[-\left(\frac{1-p}{n-1}\right) \times \ln\left(\frac{1-p}{n-1}\right) \right] \times (n-1)}{-\frac{1}{n} \times \ln\left(\frac{1}{n}\right)}$$

On the other hand, the lowest Normalized Entropy is attained when there is the most precise ranking probability distribution, i.e. when the ranking probability for the adjacent treatment is $1 - p$, and the rest are 0. The Normalized Entropy under this scenario can be written as,

$$NE_{lowest} = \frac{-p \times \ln(p) + [-(1-p) \times \ln(1-p)]}{-\frac{1}{n} \times \ln\left(\frac{1}{n}\right)}$$



I simulated datasets for number of treatments $n \in \{3,4,5, \dots, 25\}$ included in the network for the range of Normalized Entropy, conditioned on $P(\text{best}) = 0.50, 0.70,$ or 0.90 . For example, for a NMA with five treatments and $P(\text{best}) = 0.50$, the highest normalized entropy is attained for the ranking probabilities: 0.5, 0.125, 0.125, 0.125, 0.125; the lowest normalized entropy is attained for the ranking probabilities: 0.5, 0.5, 0, 0, 0.

4.3. Reanalysis of NMAs

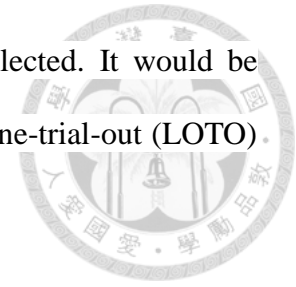
I used datasets of previously published NMAs before 2015, from a database maintained by Petropoulou et al. at the Institute of Social and Preventive Medicine (ISPM), University of Bern⁷⁵. The database can be downloaded by using the R package *nmadb*.⁷⁶ I used *network* package⁶ written for the statistical software STATA (version 14, Stata Corp, 4905 Lakeway Drive, College Station, Texas, USA) from R to undertake the frequentist NMA and to obtain ranking probabilities for all treatments. I then used R to quantify the uncertainty of ranking. The figure generation was also undertaken using statistical software R.

4.3.1. NMA Database

I used NMAs flagged as verified and those with odds ratio and mean difference as outcome measures in the arm-based data format. I firstly divided studies by their outcome measures (odds ratio or mean difference) and whether the outcome is beneficial or harmful.

To compare the robustness and uncertainty of ranking, NMAs with ten or fewer

treatments and more than two trials in each comparison were selected. It would be excluded if a network became disconnected when using the leave-one-trial-out (LOTO) approach.



4.3.2. Four Examples for Comparing Current and Purposed Methods

I selected four examples from the R package *nmadb*, which included a database of raw data for previously published NMA studies⁷⁶. I also used rankograms for example 1 and example 2 to show that it is easier to evaluate the uncertainty of treatment ranking by using a single quantitative measure than by inspecting the distribution of ranking.

The first example I selected is a NMA with low uncertainty in ranking. This example is an NMA of agents to prevent of postoperative recurrence in Crohn's disease, which included a total of four interventions and 1,507 participants⁷⁷. The second example I selected is an NMA with high uncertainty in ranking for most treatments within the network. This example is an NMA of erythropoiesis-stimulating agents for CKD patients with anemia, which included seven interventions and 12,103 participants⁷⁸. The third and fourth examples are selected to show the possible wrong evaluation that the uninformative nature of 95% CI of SUCRA can cause. The third example is an NMA of selective digestive or oropharyngeal decontamination and topical oropharyngeal chlorhexidine to prevent death in general intensive care, including 29 trials and 12,800 participants⁷⁹. The fourth example is an NMA of efficacy and safety of low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients, including 14 trials and 35,325 participants⁸⁰. Using these examples, I demonstrated the similarity and discrepancy between Normalized Entropy and the width of 95% CI of SUCRA in evaluating ranking uncertainty, and sought explanations by examining the

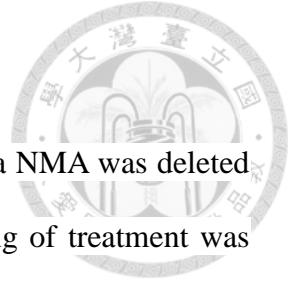
distribution of ranking probabilities.



4.3.3. Two Examples for Graph Illustration

I illustrated normalized entropy calculation and presentation in the two examples. In Example 5 (Treatment of Latent Tuberculosis Infection), the study-level data is not provided, but the ranking probabilities are reported. Normalized entropy calculation only requires ranking probabilities of each treatment. Therefore, I used the web-based tool called WebPlotDigitizer (the website: <https://automeris.io/WebPlotDigitizer/>) to extract ranking probabilities from the rankogram. I then quantified the uncertainty of ranking within an NMA using normalized entropy. In Example 6 (Fluid resuscitation in Sepsis), I used an example from the *nmadb* package and used the *netmeta* package in R to perform NMA and obtain the probability of ranking for each treatment through *rankogram* function in the *netmeta* package. The R function for calculating Normalized Entropy is as below. The input for this function is ranking probabilities (a), and the number of treatments (k). The rows of ranking probabilities matrix a are rankings (1st, 2nd, 3rd, ...), and each column is one treatment included in the NMA.

```
entropy_arm_function <- function(a, k) {  
  e = -a * log(a, 2)  
  e[!is.finite(e)] <- 0  
  e_colsum = colSums(e)  
  e_max = k * (-1/k) * log2(1/k)  
  e_prop = e_colsum/e_max  
  return(e_prop)  
}
```



4.4. Robustness of Ranking

To evaluate the robustness of treatment ranking²⁰, each trial within a NMA was deleted from the network in turn during each re-analysis and a new ranking of treatment was computed. To assess the robustness of ranking for a specific treatment, the percentage of included trials was calculated, the alternate deletion of which did not change that treatment's ranking position (treatment-level assessment). For NMA-level assessment, the agreement of the rankings between the complete dataset and the dataset with one trial being deleted was assessed by using quadratic weighted Cohen's kappa coefficients. I computed Cohen's kappa coefficients for assessing the agreement of ranking for each dataset with one trial removed, and I took the average of those kappa coefficients to represent the robustness of treatment ranking for the whole NMA.

4.4.1. Cohen's kappa coefficients

The observed agreement of two rankings can be intuitively measured by calculating the probability of being the same rank in the complete and the modified datasets with one trial removed, which is called the observed probability of agreement between two rankings, p_o . However, there is always probability of being the same rank by chance, so Cohen's kappa takes the chanced agreement out of calculations by subtracting the probability of being the same rank by chance, namely the expected probability of agreement between two rankings, p_e , from the observed probability. The difference between p_o and p_e is then divided by the perfect agreement of two rankings. The formula of Cohen's kappa coefficient (κ) is shown as below,

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

I used the weighted Cohen's kappa to take the magnitude of alterations in the ranking positions of a treatment between two lists of treatment rankings into consideration by incorporating weights into calculations of the observed and expected probabilities. The formula is given as below,

$$p_o = \sum_i \sum_j w_{ij} p_{ij}$$

$$p_e = \sum_i \sum_j w_{ij} p_{i+} p_{j+}$$

i and j are the ranking positions of the original dataset and the modified dataset with one trial removed, respectively. p_{ij} is the joint probability when the ranking is i in the original dataset and is j in the modified dataset, and p_{i+} or p_{j+} indicate the marginal probability when the ranking is i in the original dataset or j in the modified dataset. w_{ij} is the quadratic weight, which is calculated by $1 - \left(\frac{i-j}{k-1}\right)^2$, and k is number of treatments included. I used the quadratic weighted Cohen's kappa coefficient (κ), which gave more weight for the smaller difference in rank. Quadratic weighted kappa coefficient (κ) ranged from -1 to 1.

Take an NMA with three treatments as an example. If the rankings of the complete dataset and modified dataset with one trial removed are 1, 2, 3 and 1, 3, 2, respectively. Only the rank of the first treatment is unaffected, so the observed probability of agreement between two rankings is 1/3. The expected probability is calculated by two marginal probabilities of being the same rank multiplied by each other, which is 1/9 for each treatment. I then sum up over the three treatments, so the expected probability is 1/3. As both observed and expected probabilities of agreement are 1/3, the unweighted kappa coefficient is, therefore, 0. When we use the quadratic weighting

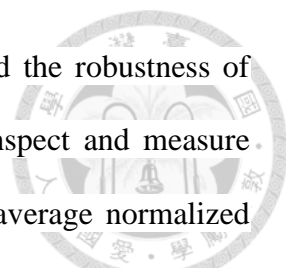
scheme, the weights assigned to total agreement is 1 for the difference of one ranking position is $3/4$, and is 0 for the difference of two ranking positions. The observed probability is $1 \times 1/3 + 3/4 \times 1/3 \times 2 = 5/6$, and the expected probability is $1 \times 1/9 \times 3 + 3/4 \times 1/9 \times 4 = 4/6$. The quadratic weighted kappa coefficient is $\frac{5/6 - 4/6}{1 - 4/6} = 0.5$.

In addition to using the average of quadratic weighted Cohen's kappa to measure the robustness of treatment ranking of the whole network, I also used the minimum and maximum values of quadratic weighted Cohen's kappa within each network to represent the worst and the best scenarios when one trial was deleted from the original dataset. Compared to the average value, the minimum and maximum values of quadratic weighted Cohen's kappa are expected to be less related to the number of treatments in the network. I classified the robustness of treatment ranking into five levels: slight (<0.2), fair (0.2-0.4), moderate (0.4-0.6), substantial (0.6-0.8), almost perfect (>0.8) agreement.

4.4.2. Treatment-level and NMA-level assessment

For treatment-level assessment, I used normalized entropy; for NMA-level assessment, the average of normalized entropies of all treatments within an NMA was taken as the uncertainty of treatment ranking for the whole network. The range of normalized entropy is from 0 to 1, and I classified the uncertainty of treatment ranking into five levels, including very high (>0.8), high (0.6-0.8), median (0.4-0.6), low (0.2-0.4), very low (<0.2).

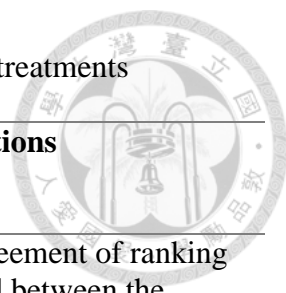
4.4.3. Association between the Uncertainty and Robustness of Ranking



To evaluate the association between the uncertainty of ranking and the robustness of ranking, the scatterplot and Pearson's correlation coefficients to inspect and measure their relationships was used. For the NMA-level assessment, the average normalized entropy was compared to the average, minimum, and maximum value of quadratic weighted Cohen's kappa of each NMA. Five levels of the average normalized entropy and five levels of the average, minimum, and maximum value of quadratic weighted Cohen's kappa also tabulated for comparisons. For the treatment-level assessment, the scatterplot and Pearson's correlation coefficients of normalized entropy for each treatment against the percentage of trials the alternate deletion of which did not change the rank of treatment was used.

The linear model for NMA-level assessment and linear mixed model for treatment-level assessment was used to further investigate the strength of association between uncertainty and robustness of ranking. Several potential factors were considered, including the number of included trials, total participants, number of treatments, and interventions assessed (Table 5). For the NMA-level analysis, the average quadratic weighted Cohen's kappa as the dependent variable was used, and the average normalized entropy and the other four NMA-level variables (number of included trials, the number of total participants, the number of treatments, and the type of interventions assessed) as explanatory variables. For the treatment-level analysis, the random intercept model with NMA as random effect was used; the percentage of studies, the deletion of which did not change the rank of a specific treatment, was the dependent variable, and the normalized entropy of that treatment and the other four NMA-level variables were explanatory variables.

Table 5. Variables used for regression analysis of 60 NMAs and 348 treatments



Outcome / Explained variable	Variables	Level of variables	Definitions
Outcome variable	Agreement of ranking (average quadratic weighted cohen's kappa)	NMA-level	the agreement of ranking derived between the complete dataset and the dataset with one trial removed
Outcome variable	Percentage of studies that did not change rank	Treatment-level	For each treatment, the percentage of trials within a NMA that did not change ranking by removing one trial out
Explanatory variable	Average Normalized Entropy	NMA-level	average normalized entropy of each treatment in the study
Explanatory variable	Normalized Entropy	Treatment-level	normalized entropy of each treatment in the study
Explanatory variable	Number of included trials	NMA-level	the total number of trials included in the study
Explanatory variable	Number of total participants	NMA-level	the total number of participants included in the study
Explanatory variable	Number of treatments	NMA-level	The total number of treatments included in the study
Explanatory variable	Type of interventions assessed	NMA-level	non-pharmacological vs any; pharmacological vs pharmacological; pharmacological vs placebo

CHAPTER 5: Results



This chapter presents the differences between current and proposed methods in real examples and simulations in the first two sections. The distribution of ranking uncertainty of published NMA and the graphs that were used to present rank with ranking uncertainty were shown in the third section. The associations between robustness and uncertainty of ranking were presented in the fourth section.


5.1 Proposed Methods

The results of comparing Normalized Entropy to other indices using four examples were described in the first section, and the results of comparing Normalized Entropy to Variance were described in the second section.

5.1.1 Comparing Normalized Entropy, Rankogram, and the Width of 95% CI of SUCRA

The ranking probabilities, the width of 95% CI of SUCRA, and Normalized Entropy for each treatment in four examples were presented in Table 6.

Example 1: 5-aminosalicylates, immunomodulators, and biologics for the prevention of postoperative recurrence in Crohn's disease



In Example 1, the relative effectiveness of the four treatments can easily be distinguished by the ranking probabilities (Biologics ranked first with 100%, Immunomodulators ranked second with 98.3%, 5-ASA ranked third with 93.3%, and Placebo ranked fourth with 94.9%). All the four treatments also achieved a short width of 95% CI of SUCRA and low Normalized Entropy (Biologics: 0.33 vs 0.15; Immunomodulators: 0.33 vs 0.20; 5-ASA: 0.00 vs 0.00; Placebo: 0.00 vs 0.06). The rankogram for this example is shown in Figure 8 (A).

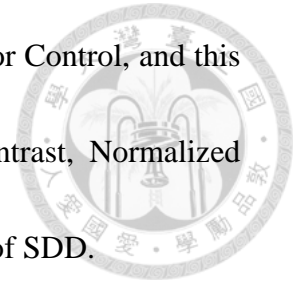
Example 2: erythropoiesis-stimulating agents for CKD patients with anemia

In Example 2, the distribution of ranking probabilities was relatively flat for most interventions except for placebo/no treatment. The magnitude of uncertainty for each treatment was high except placebo/no treatment with a relatively short width of 95% CI of SUCRA (0.20) and low Normalized Entropy (0.22) among all treatments. The rankogram for this example is shown in Figure 8 (B).

Example 3: selective digestive decontamination (SDD), selective oropharyngeal decontamination (SOD), and topical oropharyngeal chlorhexidine for prevention of death in general intensive care

In Example 3, I observed that ranking probabilities distribution for Control is more peaked than that for SDD, but the width of 95% CI of SUCRA for SDD (0.33) is smaller

than that for Control (0.67), indicating higher uncertainty of ranking for Control, and this is inconsistent with the distribution of ranking probabilities. In contrast, Normalized Entropy for Control was 0.21, smaller than 0.42, Normalized Entropy of SDD.



Example 4: low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients

In Example 4, I observed that the width of 95% CI of SUCRA for the three interventions, namely placebo, fondaparinux, and unfractionated heparin (UFH), was identical. However, their distributions of ranking probabilities were quite different, especially that for fondaparinux, which showed a very high probability of being the best (0.888). In contrast, the Normalized Entropy of fondaparinux was 0.28, which was much smaller than those of the other two interventions (placebo: 0.78, UFH: 0.85).

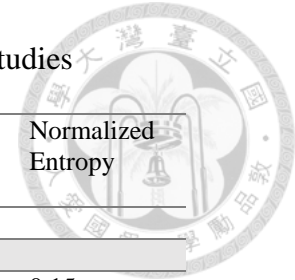
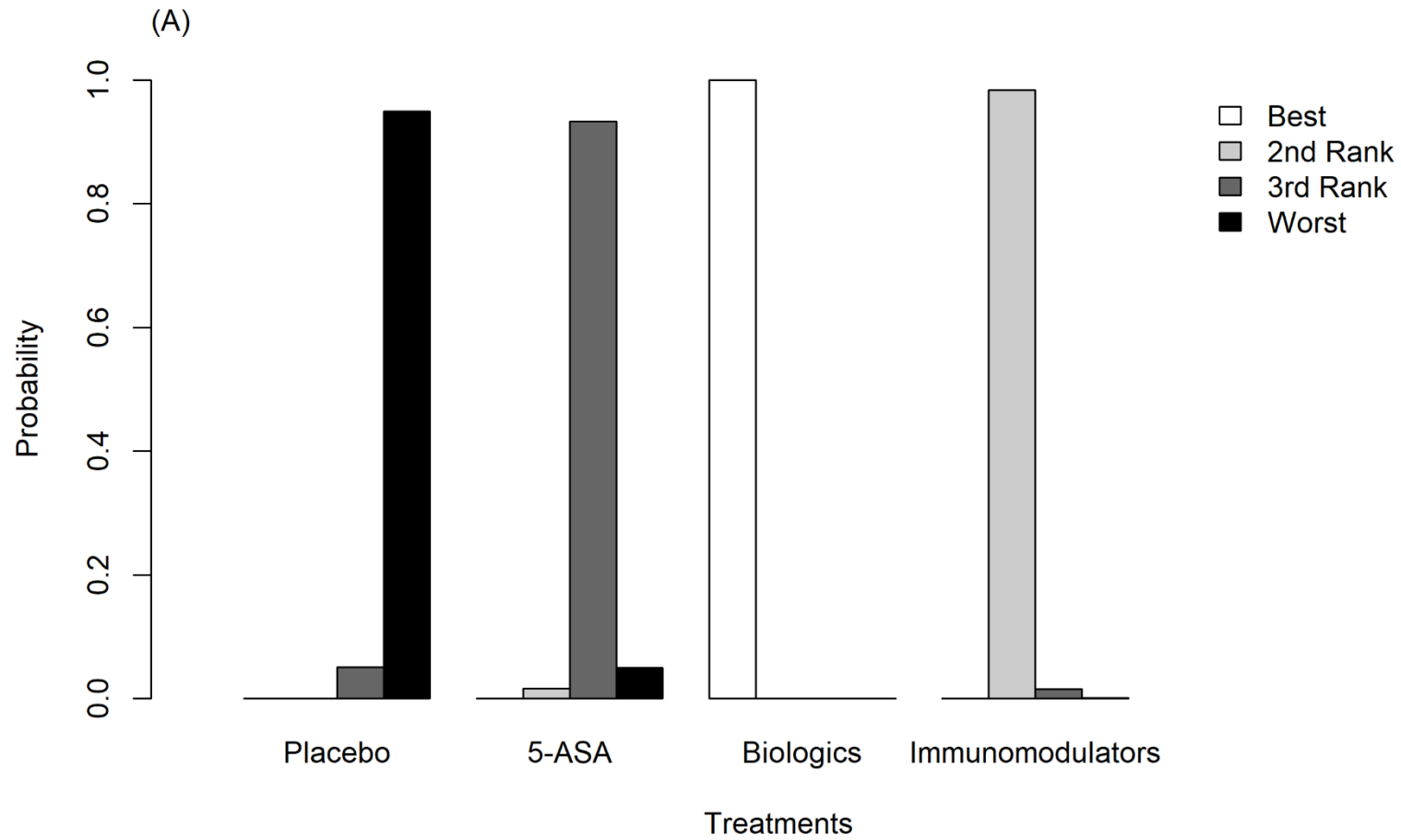
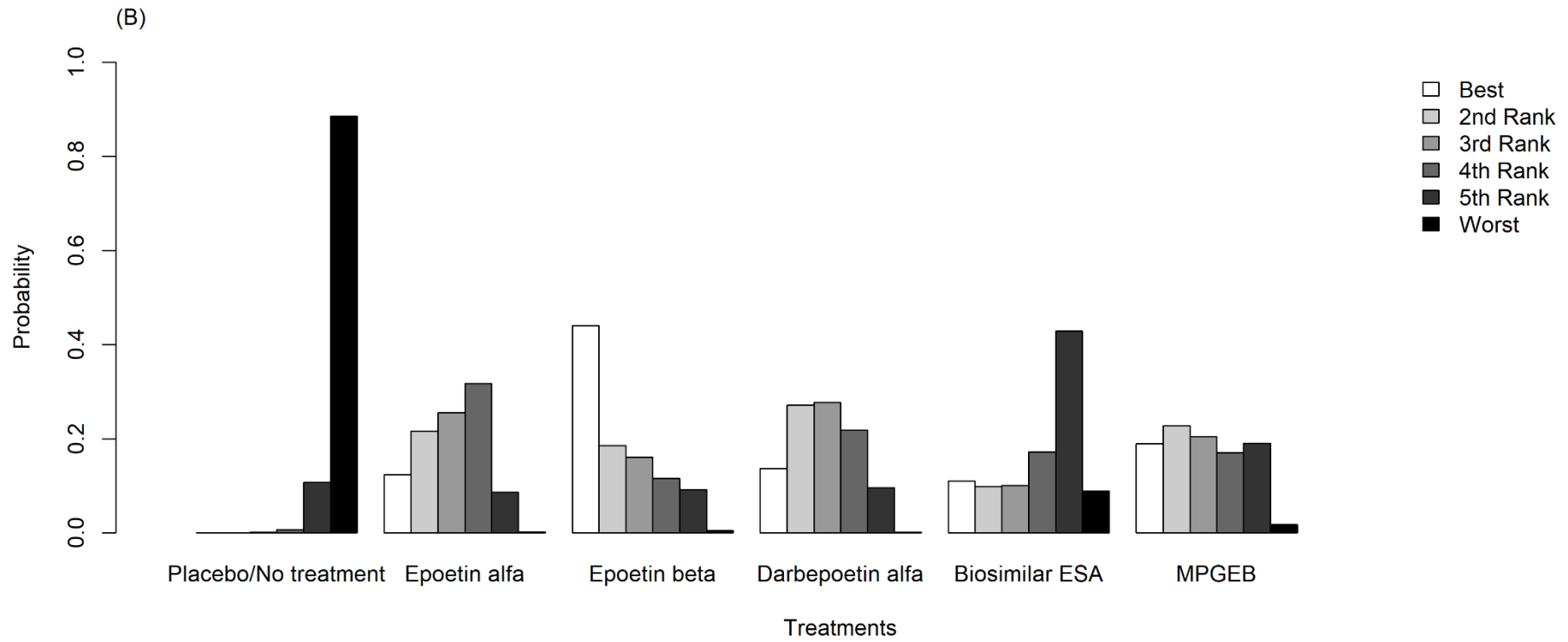


Table 6. Ranking probabilities, the width of 95% CI of SUCRA, and Normalized Entropy of the four network meta-analysis studies

Treatments	Ranking probabilities							The width of 95% CI of SUCRA	Normalized Entropy
	1 st	2 nd	3 rd	4 th	5 th	6 th	7 th		
Example 1: agents for the prevention of postoperative recurrence in Crohn's disease									
Placebo	0.0%	0.0%	5.1%	94.9%				0.33	0.15
5-ASA	0.0%	1.7%	93.3%	5.0%				0.33	0.20
Biologics	100.0%	0.0%	0.0%	0.0%				0.00	0.00
Immunomodulators	0.0%	98.3%	1.6%	0.1%				0.00	0.06
Example 2: erythropoiesis-stimulating agents for CKD patients with anemia									
Placebo/no treatment	0.0%	0.0%	0.1%	0.7%	10.7%	88.5%		0.20	0.22
Epoetin alfa	12.4%	21.6%	25.5%	31.7%	8.6%	0.2%		0.80	0.85
Epoetin beta	44.0%	18.6%	16.1%	11.6%	9.2%	0.5%		0.80	0.82
Darbepoetin alfa	13.7%	27.1%	27.7%	21.8%	9.6%	0.1%		0.80	0.86
Biosimilar ESA	11.0%	9.9%	10.1%	17.2%	42.9%	8.9%		1.00	0.88
methoxy polyethylene glycol-epoetin beta	18.9%	22.8%	20.5%	17.0%	19.0%	1.8%		0.80	0.93
Example 3: selective digestive decontamination (SDD), selective oropharyngeal decontamination (SOD), and topical oropharyngeal chlorhexidine for prevention of death in general intensive care									
Control	0.0%	5.1%	92.4%	2.5%				0.67	0.23
SDD	76.3%	23.7%	0.0%	0.0%				0.33	0.40
SOD	23.7%	70.8%	4.9%	0.6%				0.67	0.55
Chlorhexidine	0.0%	0.4%	2.7%	96.9%				0.33	0.11
Example 4: low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients									
Placebo	0.1%	2.4%	8.2%	26.0%	35.8%	22.2%	5.3%	0.67	0.78
Enoxaparin	1.2%	20.8%	28.9%	21.8%	16.7%	7.1%	3.5%	0.83	0.86
Certoparin	7.7%	33.0%	6.0%	6.1%	6.7%	11.1%	29.4%	1.00	0.87
Fondaparinux	88.8%	6.1%	1.6%	1.0%	0.7%	0.9%	0.9%	0.67	0.26
Nadroparin	0.5%	6.5%	10.1%	9.0%	12.7%	27.6%	33.6%	0.83	0.84
Dalteparin	1.0%	13.7%	15.2%	11.4%	11.5%	21.9%	25.3%	0.83	0.92
Unfractionated heparin (UFH)	0.7%	17.5%	30.0%	24.7%	15.9%	9.2%	2.0%	0.67	0.84

Figure 8. Rankograms for the (A) Example 1 and (B) Example 2





5.1.2 Comparing Normalized Entropy, Normalized Variance, and Normalized Standard Deviation



Table 7 presented the results of comparing using Normalized Entropy, Normalized Variance, and Normalized Standard Deviation to quantify the uncertainty of ranking for four NMA studies.

In Example 1, Normalized Entropy and Normalized Variance showed similar patterns, but the value of Normalized Variance is around half that of the Normalized Entropy. Normalized Entropy and Normalized Standard Deviation are very similar in pattern and value.

In Example 2, MPGEB treatment has the highest Normalized Entropy but not the Normalized Variance. The treatment with the highest normalized variance (0.36) and normalized standard deviation in example 2 is Biosimilar ESA because its distribution of ranking probabilities is more skewed, with a probability of 42.9% being the 5th. The ranking probabilities of MPGEB are quite evenly distributed among ranks of 1 to 5, thereby showing greater uncertainty measured by its large Normalized Entropy. In addition, the ranking probabilities distribution for Biosimilar ESA is slightly bimodal and cannot be detected by Normalized Entropy but can be detected by normalized standard deviation.

In Example 3, results of the two indices are very similar, but the range of Normalized Entropy (0.11 to 0.55) is much wider than the range of Normalized Entropy (0.02 to 0.12). The pattern and value of Normalized Standard Deviation is very similar to Normalized Entropy.

In Example 4, the Normalized Entropy of Enoxaparin (0.86) and Certoparin (0.87) is very close; however, because Certoparin has a bimodal distribution, its Normalized Variance (0.58) is almost third times that of Enoxaparin (0.21) and Normalized Standard Deviation of Certoparin (0.76) is also much larger than Enoxaparin (0.46). The Normalized Entropy of Placebo (0.78) is about three times that of Fondaparinux (0.26) because Fondaparinux has a large probability of being the best treatment (88.8%); however, the Normalized Variances of Placebo and Fondaparinux are similar (0.14 and 0.09, respectively), and Normalized Standard Deviation of Placebo and Fondaparinux are also similar (0.37 and 0.30, respectively), which are not consistent with their distributions of ranking probabilities.

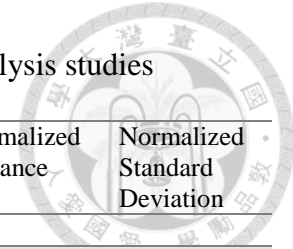


Table 7. Ranking probabilities, Entropy/Normalized Entropy, and Variance/Normalized Entropy of the four network meta-analysis studies

Treatments	Ranking probabilities							Entropy	Normalized Entropy	Variance	Normalized Variance	Normalized Standard Deviation
	1 st	2 nd	3 rd	4 th	5 th	6 th	7 th					
Example 1: agents for the prevention of postoperative recurrence in Crohn's disease												
Placebo	0.0%	0.0%	5.1%	94.9%				0.30	0.15	0.05	0.02	0.15
5-ASA	0.0%	1.7%	93.3%	5.0%				0.40	0.20	0.07	0.03	0.21
Biologics	100.0%	0.0%	0.0%	0.0%				0.00	0.00	0.00	0.00	0.00
Immunomodulators	0.0%	98.3%	1.6%	0.1%				0.12	0.06	0.02	0.01	0.11
Example 2: erythropoiesis-stimulating agents for CKD patients with anemia												
Placebo/no treatment	0.0%	0.0%	0.1%	0.7%	10.7%	88.5%		0.57	0.22	0.13	0.02	0.23
Epoetin alfa	12.4%	21.6%	25.5%	31.7%	8.6%	0.2%		2.20	0.85	1.39	0.22	0.75
Epoetin beta	44.0%	18.6%	16.1%	11.6%	9.2%	0.5%		2.12	0.82	1.91	0.31	0.87
Darbepoetin alfa	13.7%	27.1%	27.7%	21.8%	9.6%	0.1%		2.22	0.86	1.41	0.23	0.75
Biosimilar ESA	11.0%	9.9%	10.1%	17.2%	42.9%	8.9%		2.28	0.88	2.27	0.36	0.95
Methoxy polyethylene glycol-epoetin beta (MPGEB)	18.9%	22.8%	20.5%	17.0%	19.0%	1.8%		2.40	0.93	2.08	0.33	0.91
Example 3: selective digestive decontamination (SDD), selective oropharyngeal decontamination (SOD), and topical oropharyngeal chlorhexidine for prevention of death in general intensive care												
Control	0.0%	5.1%	92.4%	2.5%				0.46	0.23	0.08	0.03	0.22
SDD	76.3%	23.7%	0.0%	0.0%				0.80	0.40	0.18	0.08	0.35
SOD	23.7%	70.8%	4.9%	0.6%				1.10	0.55	0.28	0.12	0.43
Chlorhexidine	0.0%	0.4%	2.7%	96.9%				0.22	0.11	0.04	0.02	0.17
Example 4: low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients												
Placebo	0.1%	2.4%	8.2%	26.0%	35.8%	22.2%	5.3%	2.19	0.78	1.22	0.14	0.37
Enoxaparin	1.2%	20.8%	28.9%	21.8%	16.7%	7.1%	3.5%	2.41	0.86	3.67	0.21	0.46
Certoparin	7.7%	33.0%	6.0%	6.1%	6.7%	11.1%	29.4%	2.44	0.87	5.18	0.58	0.76
Fondaparinux	88.8%	6.1%	1.6%	1.0%	0.7%	0.9%	0.9%	0.73	0.26	0.81	0.09	0.30
Nadroparin	0.5%	6.5%	10.1%	9.0%	12.7%	27.6%	33.6%	2.36	0.84	2.59	0.29	0.54
Dalteparin	1.0%	13.7%	15.2%	11.4%	11.5%	21.9%	25.3%	2.58	0.92	4.86	0.37	0.61
Unfractionated heparin (UFH)	0.7%	17.5%	30.0%	24.7%	15.9%	9.2%	2.0%	2.36	0.84	3.61	0.17	0.41

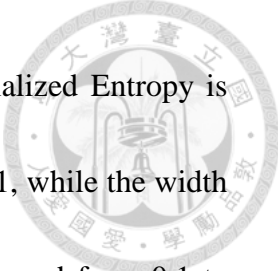


5.2 Simulations

Results of two simulations that compared Normalized Entropy to 95% CI of SUCRA and P(Best), respectively, were presented in the following two sections.

5.2.1 Comparing Normalized Entropy and 95% CI of SUCRA

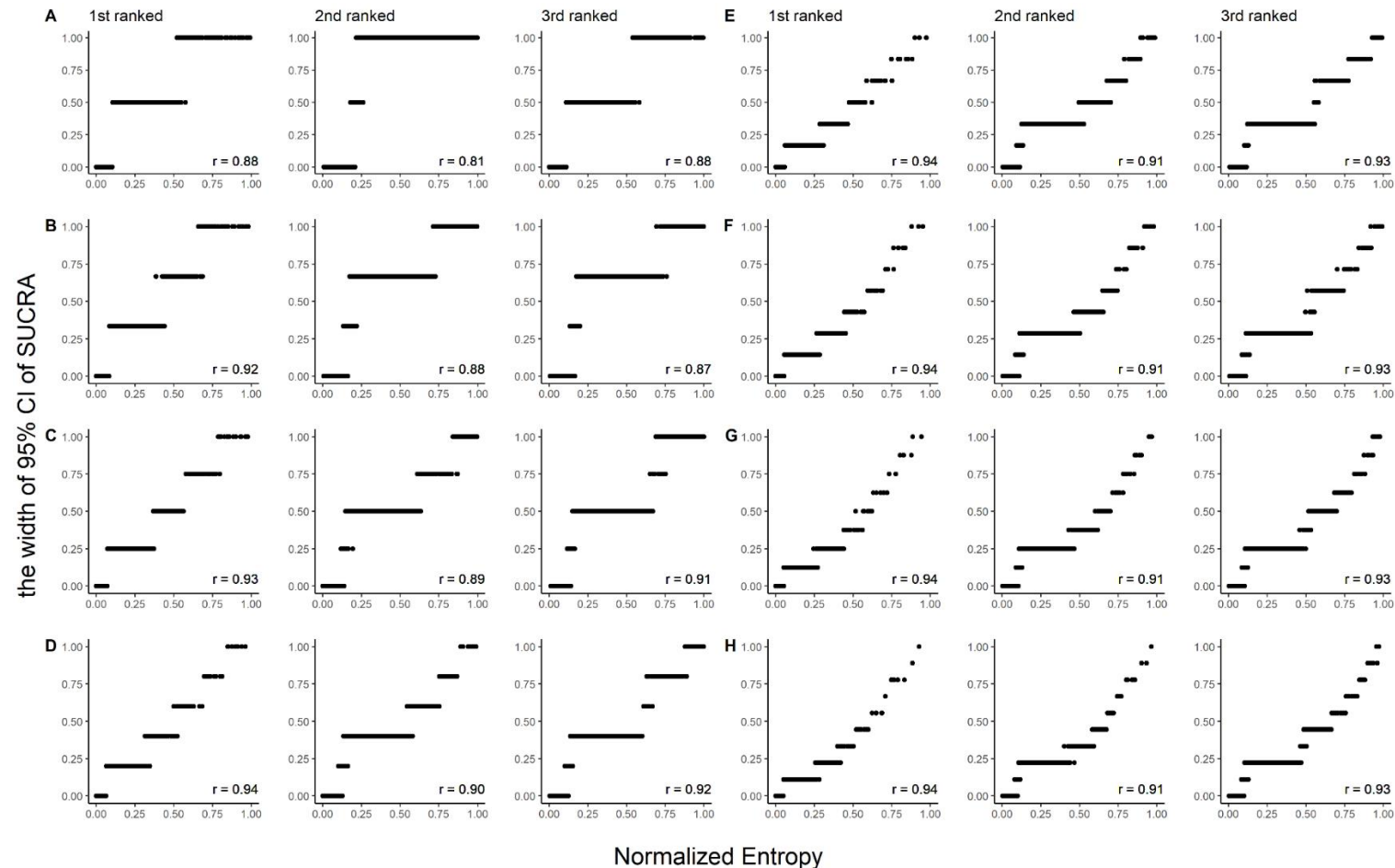
The relationship between Normalized Entropy and the width of 95% CI of SUCRA for top 3 ranked treatments within networks with 3 to 10 treatments were presented in Figure 9(A) to Figure 9(H). Overall, Normalized Entropy is positively correlated with the width of 95% CI of SUCRA under different information fractions. However, I observed that the possible values for the width of 95% CI of SUCRA equal the number of treatments included in the network. For instance, in a network with three treatments, the possible values for the width are 1, 0.5 and 0. Suppose in one simulation, treatments A, B, and C are the best, second best and worst treatment, respectively. Their SUCRA values will therefore be 1, 0.5 and 0, respectively. If treatment A is always the best, its width of 95% CI of SUCRA will be zero. If it is sometimes the second best, the width is likely to $1 - 0.5 = 0.5$. If it falls to the third place in some simulations, the width will be $1 - 0 = 1$. This applies to treatments B and C as well. For a network with k treatments, the possible widths are $1, \frac{k-2}{k-1}, \frac{k-3}{k-1}, \dots, \frac{1}{k-1}, 0$.



While the widths of 95% CI of SUCRA are discrete values, Normalized Entropy is continuous. For instance, Normalized Entropy increased from 0 to 0.1, while the width of 95% CI of SUCRA stayed at 0; when Normalized Entropy increased from 0.1 to 0.55, the width of 95% CI of SUCRA jumped to and stayed at 0.5; when Normalized Entropy increased from 0.5 to 1, the width of 95% CI of SUCRA jumped to 1. Due to its discrete nature, the width of 95% CI of SUCRA is less informative than Normalized Entropy in assessing the uncertainty of treatment ranking, particularly for networks with only a few treatments and for those treatments in the middle positions. On the other hand, for a small range of Normalized Entropy, the corresponding width of 95% CI of SUCRA may jump from one level of uncertainty to a much higher level. For example, the width of 95% CI of SUCRA for the second-best treatment in Figure 9(A) could jump from 0 to 1, when Normalized Entropy increased from 0.20 to 0.25.



Figure 9. Simulation results for relationship between Normalized Entropy and the width of 95% CI of SUCRA for top 3 ranked treatments within network with 3 to 10 treatments



A: 3 treatments, B: 4 treatments, C: 5 treatments, D: 6 treatments, E: 7 treatments, F: 8 treatments, G: 9 treatments, H: 10 treatments; r presented in each figure is the correlation coefficient between Normalized Entropy and the width of 95 CI of SUCRA

5.2.2 Comparing Normalized Entropy and P(Best)

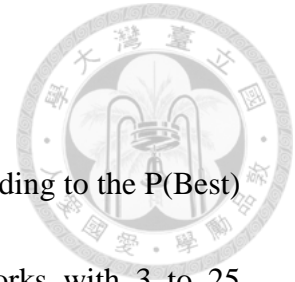


Figure 10 showed the possible range of Normalized Entropy corresponding to the P(Best) equal to 0.50, 0.70, or 0.90. The results were presented for networks with 3 to 25 treatments.

Generally, the range of Normalized Entropy is smaller for P(Best), equal to 0.90 (Figure 10 (C)) than that for P(Best), equal to 0.50 (Figure 10 (A)). For the probability of being the best equal to 0.90, the highest Normalized Entropy is around 0.20, and the lowest is around 0.1. For the probability of being the best equal to 0.7, the highest Normalized Entropy is around 0.60 and the lowest is around 0.20. For the probability of being the best equal to 0.50, the highest Normalized Entropy is around 0.80, and the lowest is around 0.20. This indicates that if the Normalized Entropy of treatment is smaller than 0.20, its level of ranking uncertainty may be interpreted as that of treatment with P(Best) equal to 0.90 or higher. If Normalized Entropy for treatment is around 0.40, its level of ranking uncertainty may be interpreted as that of treatment with P(Best) equal to 0.70. If Normalized Entropy for treatment is around 0.80 or above, its level of ranking uncertainty may be interpreted as that of a treatment with P(Best) equal to or less than 0.5. When the number of treatments increases, the same P(Best) indicates a more precise

treatment ranking and tends to attain a lower Normalized Entropy value. This demonstrates that Normalized Entropy accurately measures the uncertainty of treatment ranking, and its values are comparable across different networks.

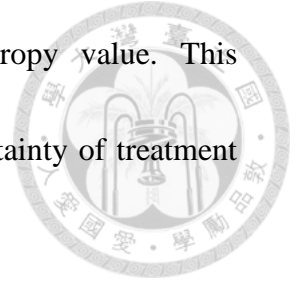
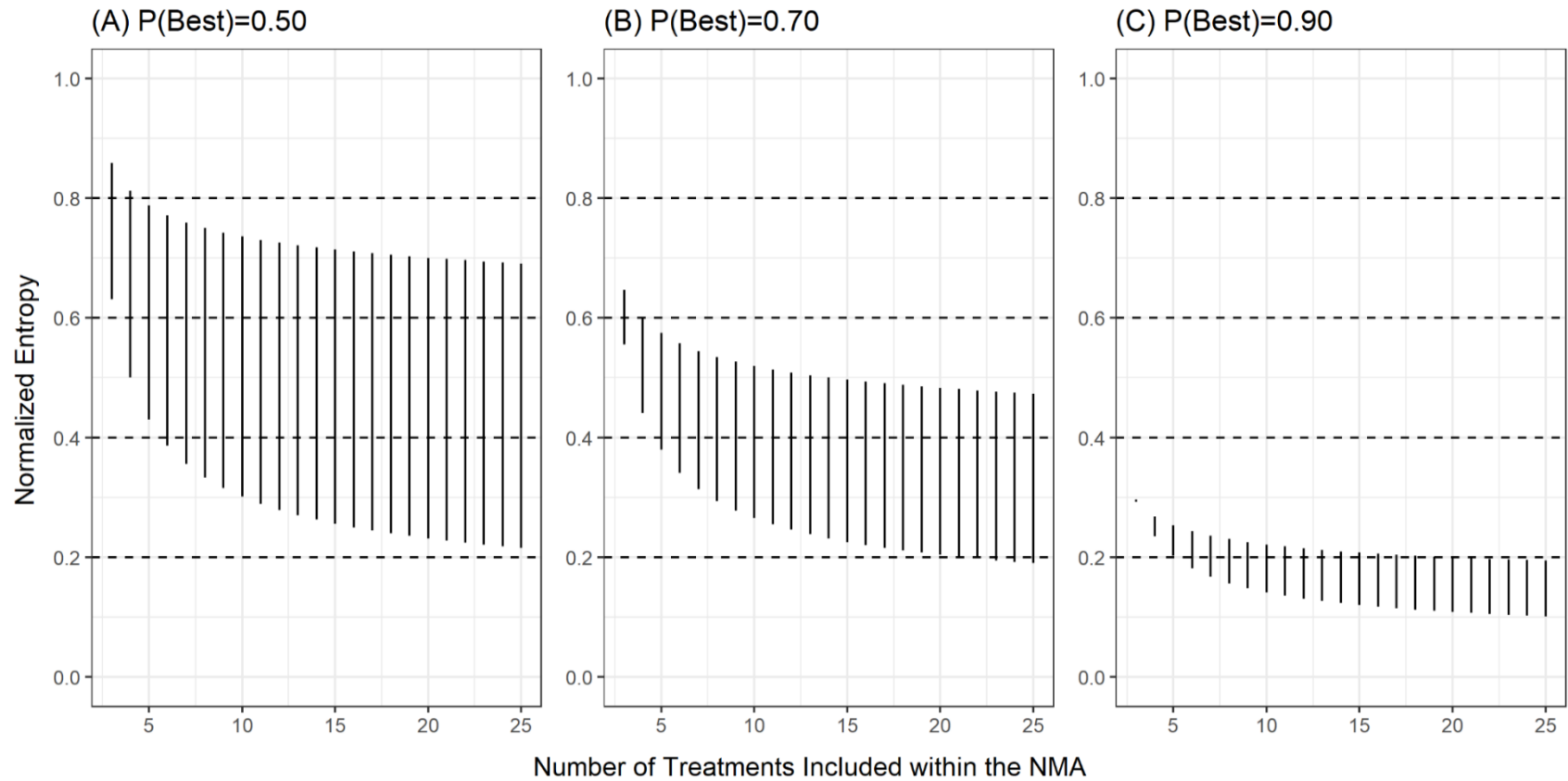




Figure 10. Range of Normalized Entropy corresponding to the probability of being the best equal to (A) 0.5, (B) 0.7 or (C) 0.9 for network with 3 to 25 treatments included



5.3 Ranking Uncertainty of Published NMA



The results of the ranking uncertainty for published NMAs are described in this section.

5.3.1 The Distribution of Ranking Uncertainty for Published NMAs

There are 453 NMAs in the catalog of the database, and 278 NMAs flagged as verified datasets. After excluding NMAs data with the outcome measures that are not odds ratio or mean difference, and the data format is not arm-based (Figure 11), 157 NMAs were included. There are 118 studies using odds ratios and 39 studies using mean difference as outcome measures.

The features of 157 NMAs were summarized in Table 8. The median number of interventions included within the network was 7 (Q1-Q3: 5-9), and over 40% of studies compared four to six interventions within the network. The median number of trials included within the network was 21 (Q1-Q3: 12-36). Near half (44.6%) NMAs included fewer than 20 trials in the network, 32.5% NMAs included 20-40 trials, 13.4% NMAs included 40-60 trials, 4.5% NMAs included 60-80 trials and 5.1% NMAs included more than 80 trials. Regarding the type of interventions assessed in the network, 63.7% NMAs were pharmacological vs. placebo, 19.1% NMAs were non-pharmacological vs. any, and 17.2% NMAs were pharmacological

vs pharmacological. The distribution of ranking uncertainty levels for 157 published NMA is shown in Figure 12. More than two-thirds of NMAs have a high (50.6%) or very high (18.6%) level of ranking uncertainty.

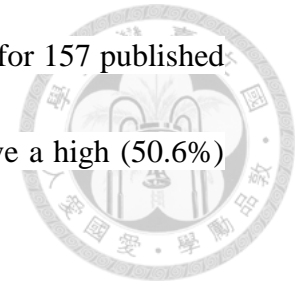


Figure 11. Flowchart of empirical dataset selection

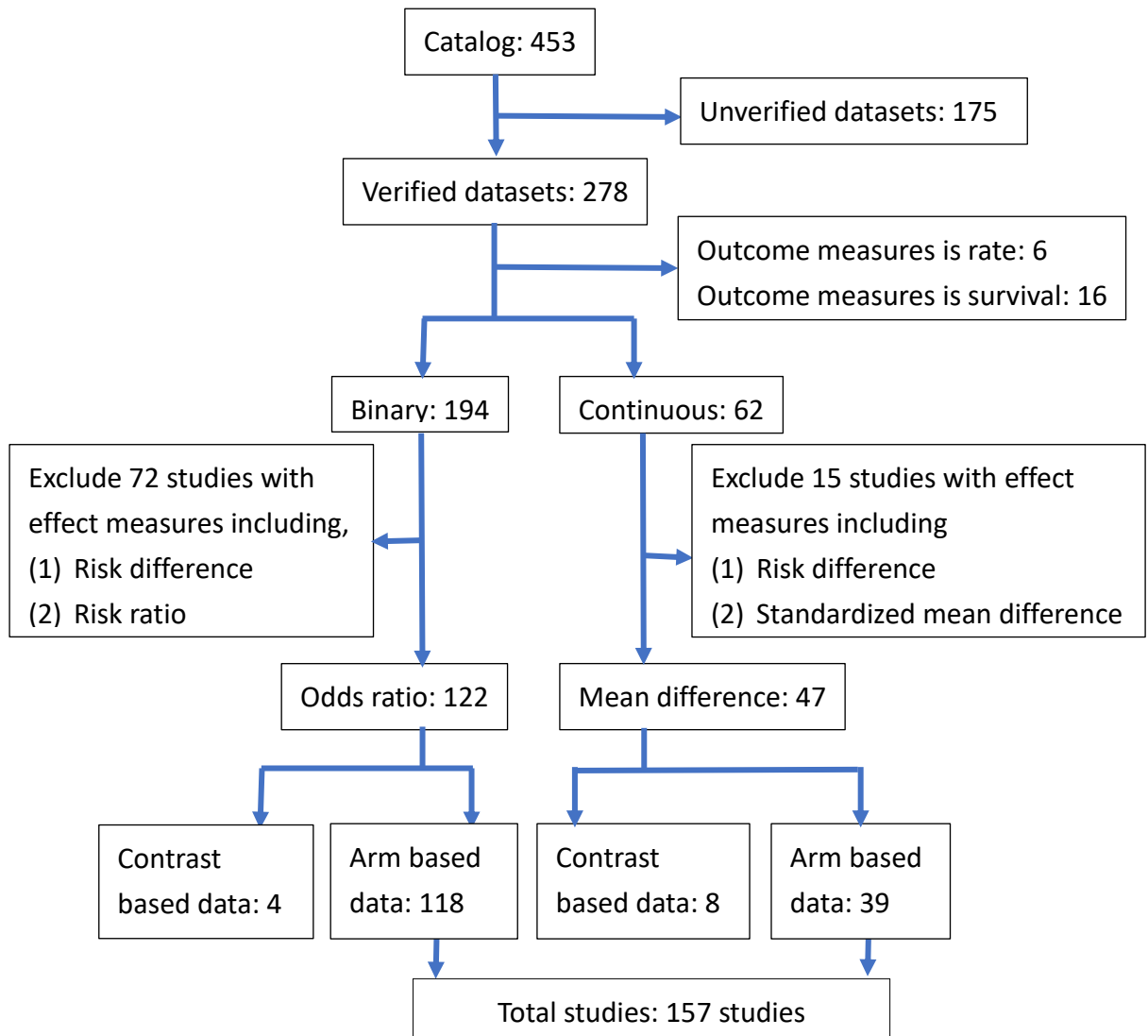


Table 8. Summary of the 157 NMAs

Characteristics	Value
Interventions, n	7 (5-9)*
Four	23 (14.6%)
Five	33 (21.0%)
Six	17 (10.8%)
Seven	15 (9.6%)
Eight	18 (11.5%)
Nine	17 (10.8%)
Ten	6 (3.8%)
More than Ten	28 (17.9%)
Trials, n	21 (12-36)*
<20	70 (44.6%)
20-40	51 (32.5%)
40-60	21 (13.4%)
60-80	7 (4.5%)
>80	8 (5.1%)
Type of interventions assessed, n	
non-pharmacological vs any	30 (19.1%)
pharmacological vs pharmacological	27 (17.2%)
pharmacological vs placebo	100 (63.7%)

*median (1st and 3rd quantile)

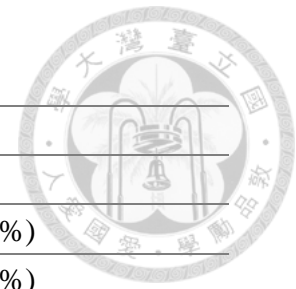
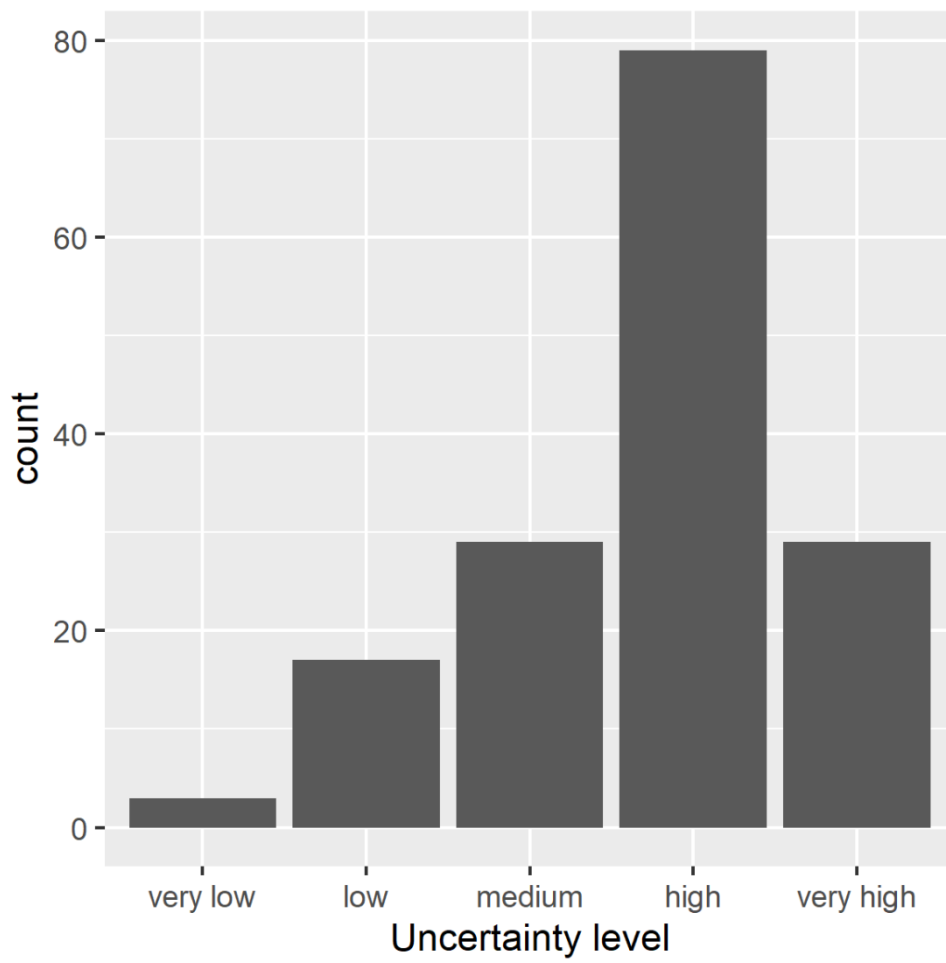


Figure 12. The distribution of ranking uncertainty levels for 157 published NMAs



Among 157 NMAs, the 95% CI range of SUCRA for one NMA cannot be estimated.

Therefore, 156 NMAs were included for exploring the association between robustness and uncertainty of ranking. The association of ranking uncertainty in

Normalized Entropy and 95% CI range of SUCRA for each treatment of 156

published NMAs by the number of treatments included in the NMA was presented

in Figure 13. When the number of treatments is less than 10, there is an apparent

trend that 95% CI of SUCRA would be affected by the number of treatments

included within the NMA. On the other hand, for those NMAs with more than ten

treatments, there is a trend that while Normalized Entropy is very high, the range of

95% CI of SUCRA could be low.

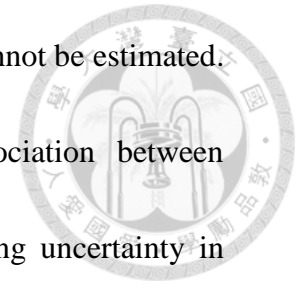
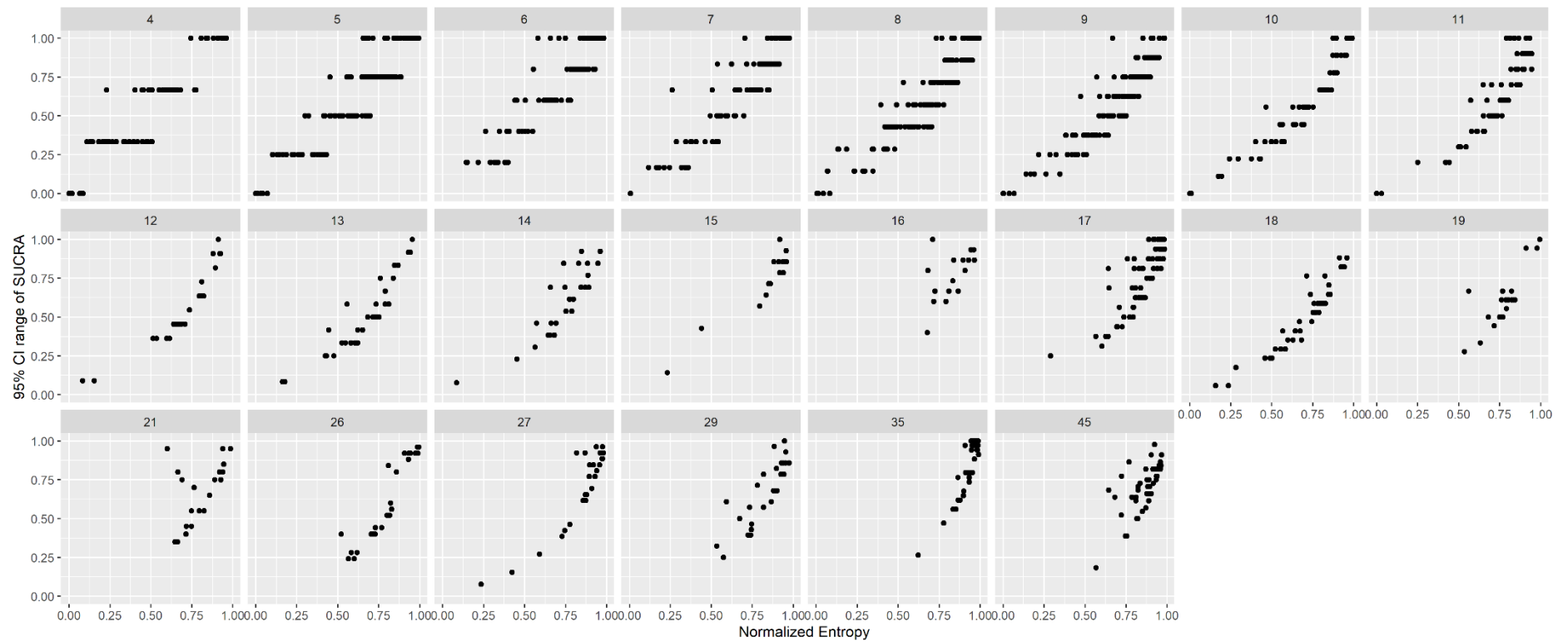




Figure 13. The association of ranking uncertainty in Normalized and the range of 95% CI of SUCRA for each treatment of 156 published NMAs by number of treatments included in the NMA



5.3.2 Two Illustrative Examples



Example 5: Treatment of Latent Tuberculosis Infection

When we compare treatments, we often do not just evaluate one outcome. We usually look for treatments that are more effective and have lower side effects. An NMA compared the efficacy of 16 treatments for preventing latent tuberculosis infection (LTBI) and 10 treatments for hepatotoxicity⁸¹. The study included 61 studies.

We presented the ranking and uncertainty levels for two outcomes of Example 5 in Figure 14. When there are two outcomes in a study, some studies may use clustered ranking plot to present two rankings simultaneously. However, treatments that do not rank both outcomes will be neglected. In addition, the clustered ranking plot cannot present rankings for three or more outcomes. They also cannot present the uncertainty of ranking.

In Figure 14, we presented ranking for each outcome and used colors to distinguish the five levels of ranking uncertainty. We found that although INH-EMB 12 months, RFB-INH, and RFB-INH (high) ranked best, third, and fourth, respectively, there is no information for prevention of active TB hepatotoxicity. RMP-INH-PZA ranked

second (uncertainty level is high), but the hepatotoxicity ranked seven (uncertainty level is very high); RMP ranked fifth (uncertainty level is very high) and have the least hepatotoxicity (uncertainty level is very low). Readers can decide the priority of treatments based on which outcome they value more, and they also can incorporate other possible considerations, such as cost, route of medicine, etc., into the figure for the overall judgement.

Example 6: Fluid resuscitation in Sepsis

Uncertainty of ranking can also be applied to determine treatment nodes while maintaining their interpretability. When performing NMAs, treatments can be lumped into the same treatment node or separated into different nodes. Treatments may be split into different groups based on dose, duration, procedures etc. However, it may not be optimal to separate all treatments because each treatment may not have sufficient power to generate robust estimates. However, if all treatments are lumped into the same nodes, the results may be useless in clinical practice⁸².

An NMA study compared mortality among different fluids for resuscitation in sepsis presented their results by using three ways (6-node, 4-node, and 2-node) to classify treatments⁸³. The study divided fluids into crystalloid and colloids for 2-node comparison, and divided colloids into albumin, gelatin, and hydroxyethyl starch

(HES) for 4-node comparison. For 6-node comparison, HES is further divided into low-molecular-weight HES (L-HES) and high-molecular-weight HES (H-HES), and crystalloids is divided into balanced and unbalanced (saline) solutions. The study includes 14 studies (18,916 patients). The three classification methods have different numbers of studies included in the analysis (2-node comparison: 12 studies; 4-node comparison: 13 studies; 6-node comparison: 13 studies).

I presented the ranking and its uncertainty for outcome with 6-node, 4-node, and 2-node, using scatter plots in Figure 15 (A)-(C). I can found that the uncertainty of treatments are lower in the 4-node analysis, and higher for only 2 nodes included in the analysis. Therefore, it showed that lumping more treatments together does not mean the uncertainty of ranking would be lower. In this Example 6, when I group treatments into 2 nodes, the uncertainty of both treatments are close to 1, which means that the ranking of these two categories are not meaningful.



Figure 14. Rankings for two outcomes with uncertainty in 5 levels

Treatments\Outcomes	(A) Prevention of active TB		(B) Hepatotoxicity	
	OR (95% CrI)	Rank (Uncertainty)	OR (95% CrI)	Rank (Uncertainty)
INH-EMB 12 mo	0.12 (0.02–0.54)	1		
RMP-INH-PZA	0.21 (0.11–0.41)	2	2.41 (0.25–20.02)	7
RFB-INH	0.18 (0.03–0.95)	3		
RFB-INH (high)	0.19 (0.03–0.98)	4		
RMP	0.25 (0.11–0.57)	5	0.14 (0.02–0.81)	1
INH 12-72 mo	0.31 (0.21–0.47)	6	2.72 (0.96–7.44)	8
RMP-INH 3-4 mo	0.33 (0.20–0.54)	7	0.72 (0.21–2.37)	3
RMP-PZA	0.33 (0.18–0.58)	8	3.32 (0.99–11.23)	9
RPT-INH	0.36 (0.18–0.73)	9	0.52 (0.13–2.15)	2
INH 6 mo	0.40 (0.26–0.60)	10	1.10 (0.40–3.17)	5
INH 9 mo	0.46 (0.22–0.95)	11	1.70 (0.35–8.05)	6
INH-EMB	0.54 (0.19–1.56)	12		
INH 3-4 mo	0.57 (0.31–1.02)	13		
RMP-INH 1 mo	0.65 (0.23–1.71)	14		
Placebo	0.62 (0.41–0.94)	15	4.12 (1.33–15.88)	10
No treatment	1.00 (reference)	16	1.00 (reference)	4

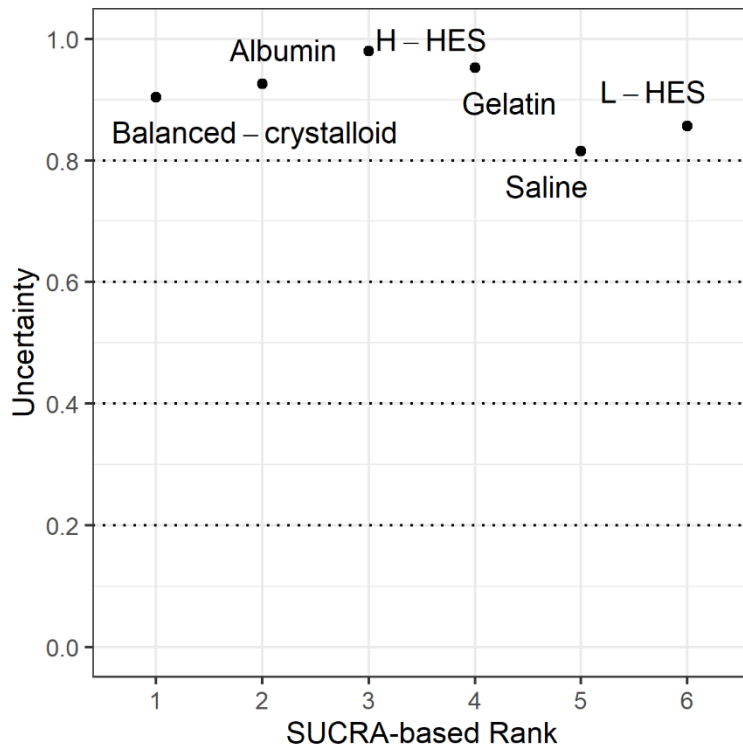
Normalized Entropy	Uncertainty
0.8-1	Very high
0.6-0.8	High
0.4-0.6	Medium
0.2-0.4	Low
0-0.2	Very low

Figure 15 (A) SUCRA-based rank with uncertainty of 6 treatments (B) 4 treatments

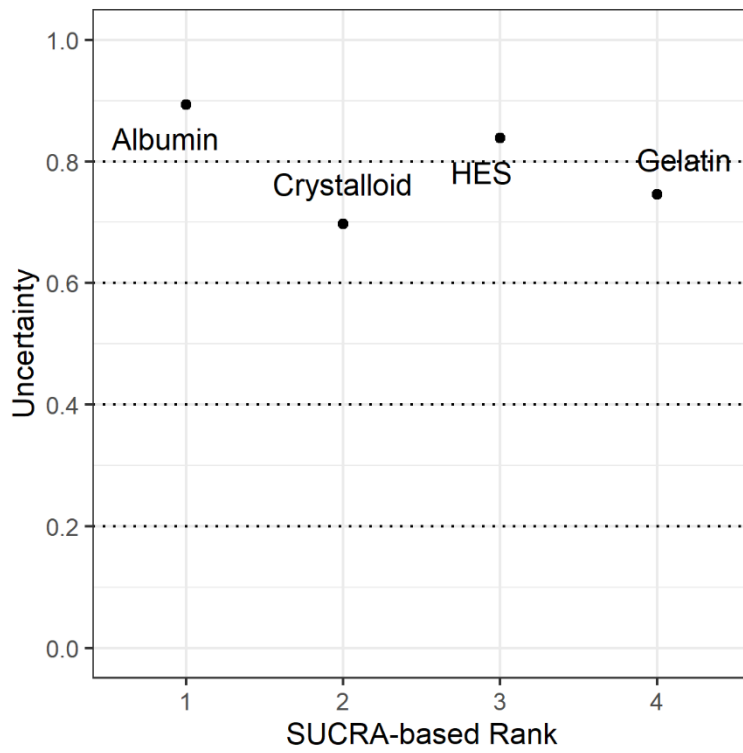
(C) 2 treatments for example



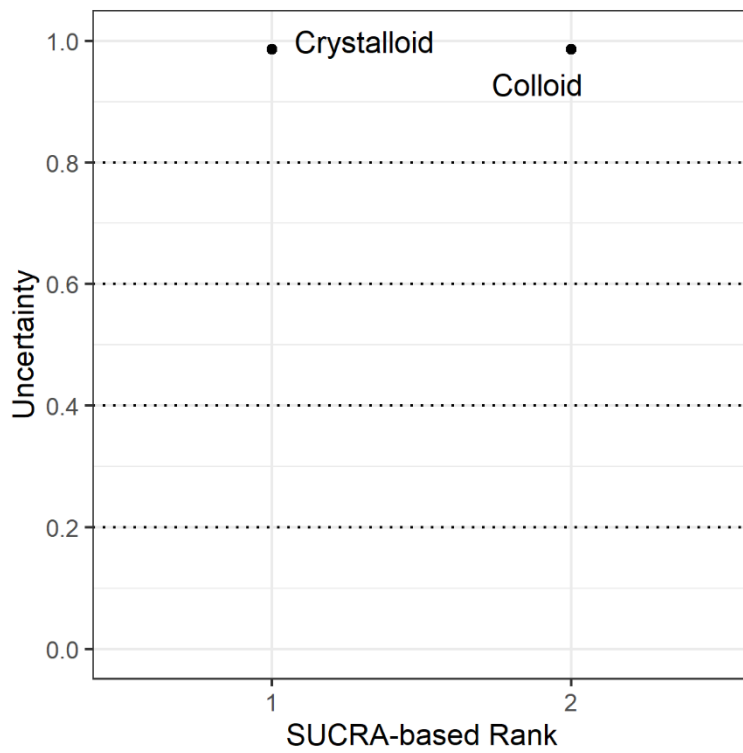
(A)



(B)



(C)



5.4 Association between Uncertainty and Robustness of Treatment Ranking

The selection process of NMAs from *nmadb* database for exploring the association between uncertainty and robustness of treatment ranking were shown in Figure 16. A

total of 60 NMAs were included. 43 NMAs reported odds ratios and 17 NMAs the mean difference among them. The basic information of the 60 NMAs was summarized in Table 9. The median number of interventions included within the network was 5 (Q1-Q3: 4-7), and over 70% of NMAs compared fewer than six interventions. The median number of trials included within the network was 26 (Q1-Q3: 17-36). More than one quarter (28.3%) of NMAs included fewer than 20 trials in the network, 50.0% NMAs included 20-40 trials, 15.0% NMAs included 40-60 trials and 6.7% NMAs included more than 60 trials. Regarding the type of interventions assessed in the network, 66.7% NMAs were pharmacological vs. placebo, 20.0% NMAs were non-pharmacological vs. any, and 13.3% NMAs were pharmacological vs. pharmacological. When one of their included trials was deleted, 50 NMAs (80.0%) treatment ranking was altered. Further information, such as condition/disease, outcome measure, and the number of trials and treatments included of each NMA, can be found in Table 10 and Table 11.

Figure 16. Flowchart of the study selection process

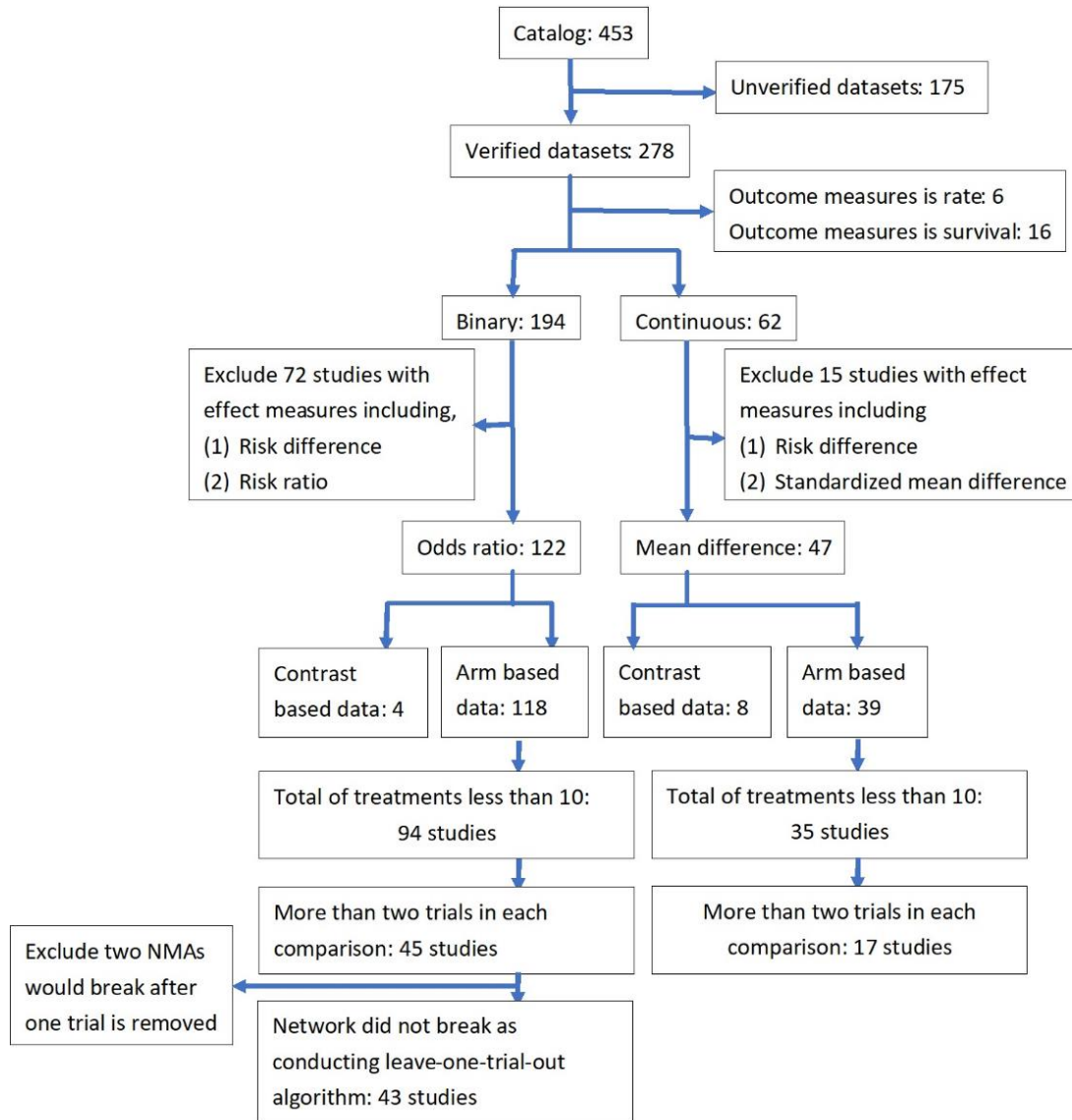
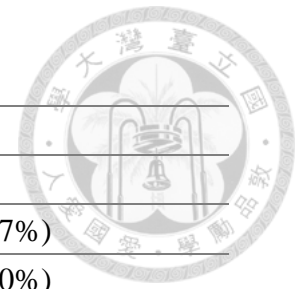


Table 9. Summary of the 60 NMAs

Characteristics	N (%)
Interventions, n	5 (4-7)*
Four	19 (31.7%)
Five	15 (25.0%)
Six	10 (16.7%)
Seven	3 (5.0%)
Eight	4 (6.7%)
Nine	6 (1.0%)
Ten	3 (5.0%)
Trials, n	26 (17-36)*
<20	17 (28.3%)
20-40	30 (50.0%)
40-60	9 (15.0%)
>60	4 (6.7%)
Type of interventions assessed, n	
non-pharmacological vs any	12 (20.0%)
pharmacological vs pharmacological	8 (13.3%)
pharmacological vs placebo	40 (66.7%)
Ranking of treatments after leave-one-trial out approach	
All remained unchanged	12 (20.0%)
Have some change	48 (80.0%)

*median (1st and 3rd quantile)



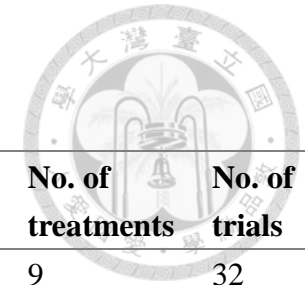
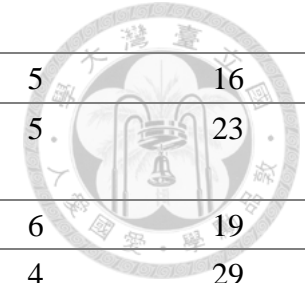
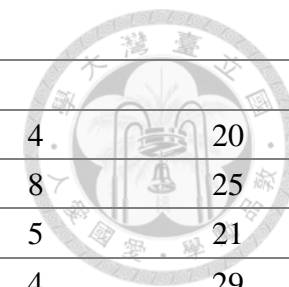


Table 10. Basic characteristics of 60 NMAs

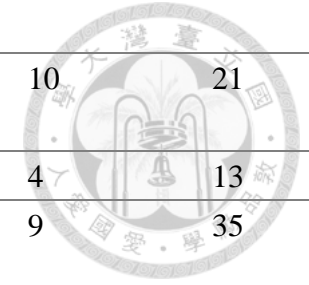
ID	First Author	Year	Condition/Disease	Outcome	No. of treatments	No. of trials
473552	Valentin	2011	rheumatoid arthritis	ACR70 improvement	9	32
479585	Singh	2013	any disease condition except human immunodeficiency disease (HIV/AIDS)	total adverse events	10	49
479600	Donahue	2012	rheumatoid arthritis	response to treatment (defined as achieving ACR50 response)	9	30
479622	Thakkinstian	2012	chronic prostatitis/chronic pelvic pain syndrome	Total symptom scores	5	13
479629	Reinecke	2015	chronic pain	pain intensity	4	22
479650	Coleman	2008	cancer	incidence of cancer	6	27
479661	Owen	2010	stroke	all strokes	4	14
479770	Greco	2015	adult cardiac surgery patients	mortality	5	46
479808	Singh	2009	rheumatoid arthritis	50% improvement in patient- and physician-reported criteria of the American College of Rheumatology (ACR 50)	7	31
479971	Loke	2014	acute coronary syndrome	adverse coronary events	4	27
480029	Xiong	2014	localized prostate cancer	all-cause mortality	8	17
480060	Tadrous	2014	primary osteoporosis	any gastrointestinal related to adverse events	5	46



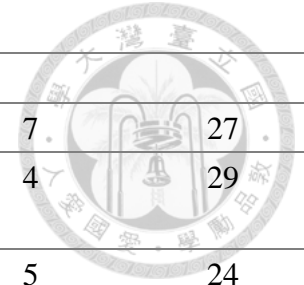
480074	Phan	2014	atrial fibrillation	incidence of sinus rhythm	5	16
480612	Kunitomi	2015	mild-to-moderate asthma	change from baseline in forced expiratory volume in 1 s (FEV1(L))	5	23
480804	Palmer	2014	chronic kidney disease	preventing blood transfusion	6	19
480851	Stowe	2010	parkinson's disease patients suffering from motor complications	off-time reduction	4	29
481107	Wang	2015	patients with cancer receiving myelosuppressive chemotherapy	Febrile neutropenia risk for all chemotherapy cycles without adjustment for relative dose intensity	5	30
481140	Kunitomi	2013	schizophrenia	symptoms of schizophrenia	4	21
481236	Zagmutt	2012	early parkinson's disease	Total Adverse Events	4	6
481384	Schoenberg	2013	laparoscopic heller myotomy	Success Rates at 12 months	4	16
481583	Sun	2014	moderate-to-severe restless legs syndrome	Change in IRLS score at the end of maintenance	5	14
481589	Desai	2012	rheumatoid arthritis	overall withdrawal	10	41
481695	Liang	2014	advanced non-small-cell lung cancer	objective response rate	6	11
481733	Yang	2014	crohn's disease recurrence	endoscopic recurrence	4	12
481734	Zhang	2013	patients treated with antihypertensive drugs	new-onset diabetes	6	28
481836	Roskell	2014	chronic obstructive pulmonary	trough forced expiratory volume in 1 second	9	16



			disease	(FEV1)		
481941	Dong	2014	rocuronium	onset time of rocuronium	4	20
482004	Shamliyan	2013	episodic migraine	prevention of episodic migraine	8	25
482006	Lin	2014	primary molar pulpotomy	clinical success of primary molar pulpotomy	5	21
482258	Price	2014	patients in general intensive care	mortality	4	29
482382	Shams	2013	hot flashes	daily frequency of hot flashes	5	9
482734	FurUnited Kingdomawa	2014	control conditions currently used in psychotherapy trials	response to treatment	4	48
501201	Baker	2009	chronic obstructive pulmonary disease	exacerbation episodes in Chronic Obstructive Pulmonary Disease (COPD \geq 1)	8	39
501226	Chang	2012	plantar fasciitis	effectiveness of focused shock wave (FSW) therapy of different intensity levels and a new alternative, radial shock wave (RSW) for managing plantar fasciitis	5	12
501235	Cooper	2011	smoke alarms	possession of a functioning alarm	7	20
501251	Dong YH	2013	chronic obstructive pulmonary disease	Risk of mortality for inhaled medications in patients with chronic obstructive pulmonary disease (COPD)	6	41
501256	Eisenberg	2008	smoking cessation	most rigorous criterion of abstinence in smoking cessation	5	61
501257	Elliott	2007	diabetes	effect of antihypertensives on incidence diabetes mellitus	6	22

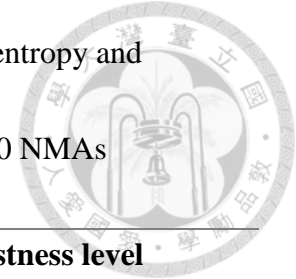


501261	Filippini G	2013	multiple sclerosis	the relative effectiveness of immunomodulators and immunosuppressants in patients with MS	10	21
501277	Goudswaard	2009	type 2 diabetes	insulin therapies in patients with type 2 diabetes	4	13
501281	A.K. Gupta	2013	actinic keratosis	The relative efficacy of eight treatments in nonimmunosuppressed participants for actinic keratosis	9	35
501297	Hutton	2012	cardiac surgery	relative risks of death between antifibrinolytics and no treatment	4	77
501300	Jansen	2006	type 2 diabetes	relative efficacy	4	12
501305	Nasreen Khan	2013	refractory partial onset seizure	Efficacy of anti-epileptic drugs	5	12
501317	Lin	2012	dentin hypersensitivity	effectiveness in resolving dentin hypersensitivity among different in-office desensitizing treatments	6	41
501321	Liu	2012	type 2 diabetes	effectiveness	9	39
501325	Mak	2012	oral antithrombotic agents	Acute coronary events comprising either MI or ACS	5	26
501332	Maund	2011	morphine-related side-effects after major surgery	morphine-related outcomes	4	58
501337	Meader	2010	opioid detoxification	Completion of treatment	4	20
501340	Middleton	2010	heavy menstrual bleeding	efficacy as second line treatment for heavy menstrual bleeding	4	20
501348	Mills	2010	short-term smoking abstinence	Smoking Abstinence	4	89
501351	Huseyin Naci	2013	statins	Harms of individual statins	8	101
501374	Puhan	2009	chronic obstructive pulmonary	exacerbation in patients with chronic obstructive	5	34



		disease	pulmonary disease		
501393	Singh	2009	rheumatoid arthritis	efficacy of biologics for rheumatoid arthritis	7 27
501399	Stowe	2010	parkinson's disease patients with motor complications	efficacy	4 29
501404	Thijs	2008	transient ischaemic attack or stroke	efficacy of antiplatelet	5 24
501408	Tropeano	2010	carotid intima-media thickness	decrease of carotid intima-media thickness (CIMT)	6 28
501414	Van de Bruel	2010	cataract surgery	protective effect of ophthalmic viscoelastic devices	6 21
501424	Wang	2010	catheter-related infections	effectiveness of venous catheters for catheter-related infections	9 43
501434	Yu	2006	coronary artery bypass graft surgery	effectiveness of inhaled anesthetics in reducing post-operative myocardial infarctions after cardiac surgery	6 14

Table 11. Uncertainty of ranking evaluated by average normalized entropy and robustness of ranking evaluated by LOTO and Cohen's kappa for 60 NMAs

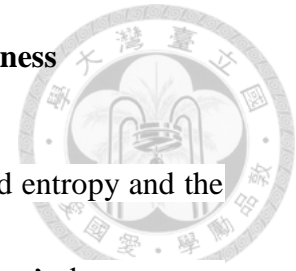


ID	Average Normalized Entropy	Uncertainty level (defined by average Normalized Entropy)	Robustness level (defined by Squared weighted Cohen's kappa)
473552	0.77	High	0.98
479585	0.89	Very High	0.98
479600	0.62	High	0.99
479622	0.62	High	0.95
479629	0.47	Median	1.00
479650	0.84	Very High	0.97
479661	0.43	Median	0.99
479770	0.76	High	0.96
479808	0.65	High	0.99
479971	0.25	Low	1.00
480029	0.87	Very High	0.95
480060	0.71	High	0.99
480074	0.50	Median	0.90
480612	0.59	Median	0.98
480804	0.76	High	0.89
480851	0.34	Low	0.99
481107	0.69	High	0.92
481140	0.11	Very Low	1.00
481236	0.44	Median	0.97
481384	0.26	Low	0.99
481583	0.44	Median	0.96
481589	0.62	High	0.99
481695	0.47	Median	0.99
481733	0.11	Very Low	1.00
481734	0.32	Low	1.00
481836	0.60	High	0.97
481941	0.50	Median	0.97
482004	0.78	High	0.95
482006	0.71	High	0.98

482258	0.32	Low	1.00
482382	0.76	High	0.96
482734	0.29	Low	1.00
501201	0.63	High	0.98
501226	0.72	High	0.96
501235	0.68	High	0.98
501251	0.58	Median	0.98
501256	0.27	Low	1.00
501257	0.37	Low	1.00
501261	0.78	High	0.97
501277	0.74	High	0.98
501281	0.70	High	0.97
501297	0.67	High	0.99
501300	0.34	Low	0.97
501305	0.71	High	0.93
501317	0.83	Very High	0.95
501321	0.60	Median	0.98
501325	0.62	High	0.98
501332	0.33	Low	1.00
501337	0.53	Median	1.00
501340	0.58	Median	0.97
501348	0.14	Very Low	1.00
501351	0.75	High	0.99
501374	0.56	Median	0.96
501393	0.66	High	0.99
501399	0.35	Low	0.99
501404	0.22	Low	1.00
501408	0.79	High	0.95
501414	0.76	High	0.96
501424	0.67	High	0.99
501434	0.83	Very High	0.99



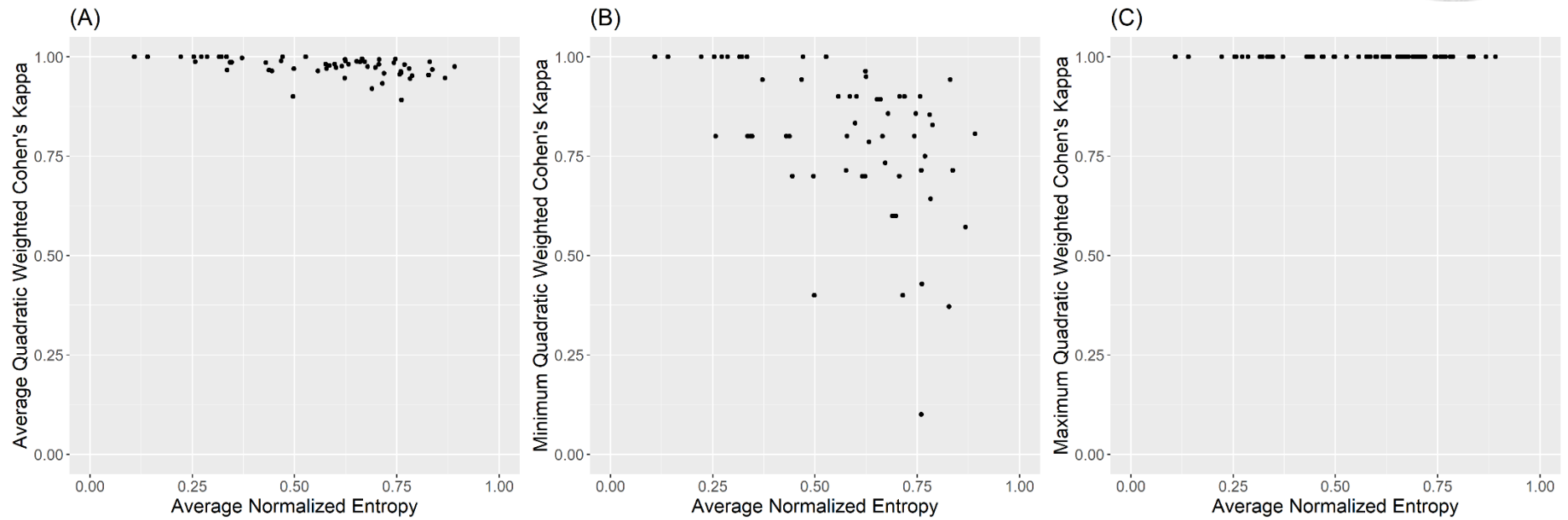
5.4.1 NMA-level Association between Uncertainty and Robustness

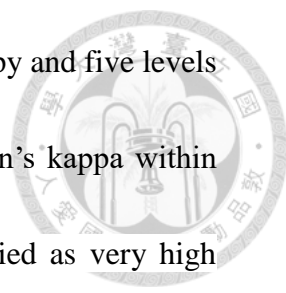


For the 60 NMAs, the associations between the average normalized entropy and the average, minimum, maximum value of quadratic weighted Cohen's kappa were presented in Figure 17. Their Pearson's correlation coefficients were -0.59, -0.50, and 0, respectively. While the average normalized entropy was less than 0.4, the average values of quadratic weighted Cohen's kappa were close to 1. When the normalized entropy increased, the variation of average quadratic weighted Cohen's kappas increased, but they remained almost greater than 0.9. The minimum value of quadratic weighted Cohen's kappa showed greater variations when the average normalized entropy was high. When the average normalized entropy was less than 0.25, the minimum values of quadratic weighted Cohen's kappa became 1, i.e., a perfect agreement. When the average normalized entropy was greater than 0.75, the minimum value of quadratic weighted Cohen's kappa ranged between 0.2 and 0.8. The higher the average normalized entropy was, the lower the minimum quadratic weighted Cohen's kappa was. However, some NMAs with a high average normalized entropy showed high minimum quadratic weighted Cohen's kappa. The maximum value of quadratic weighted Cohen's kappa were all 1 for 60 NMAs, showing that there was at least one trial the deletion of which did not change their treatment ranking.



Figure 17. Scatter plots of average normalized entropy and (A) average/ (B) minimum/ (C) maximum quadratic weighted Cohen's kappa for 60 networks



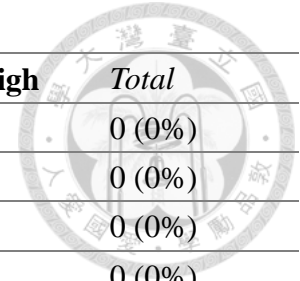


In Table 12, I compared five levels of the average normalized entropy and five levels of the average, minimum, and maximum value of quadratic Cohen's kappa within the network. The uncertainty of about half of NMAs was classified as very high (8.3%) and high (45.0%), and that of the other half was medium (21.7%), low (20.0%), and very low (5.0%). As for the robustness of ranking, both the average and maximum values of quadratic Cohen's kappa for all 60 NMAs fell into the highest level of agreement irrespective of their levels of normalized entropy. Consequently, 32 NMAs of high or very high uncertainty of ranking appeared to show good robustness of treatment ranking. In contrast, the minimum values of quadratic weighted Cohen's kappa were classified as substantial (36.7%) or almost perfect (51.7%) agreement on ranking, while a few NMAs with medium to the very high level of uncertainty fall into the groups of slight (1.7%), fair (5.0%), moderate (5.0%) agreement for the robustness of ranking.



Table 12. Comparison of five levels between uncertainty of treatment ranking quantified by the average normalized entropy and robustness of treatment ranking quantified by the (A) Average (B) Minimum (C) Maximum value of quadratic Cohen's kappa within the network

(A)		Uncertainty of treatment ranking (Average Normalized Entropy)					
		Very low	Low	Median	High	Very high	Total
Robustness of treatment ranking (Average quadratic Cohen's kappa)	Slight agreement	0	0	0	0	0	0 (0%)
	Fair agreement	0	0	0	0	0	0 (0%)
	Moderate agreement	0	0	0	0	0	0 (0%)
	Substantial agreement	0	0	0	0	0	0 (0%)
	Almost perfect agreement	3	12	13	27	5	60 (100%)
	Total	3 (5.0%)	12 (20.0%)	13 (21.7%)	27 (45.0%)	5 (8.3%)	60
(B)		Uncertainty of treatment ranking (Average Normalized Entropy)					
		Very low	Low	Median	High	Very high	Total
Robustness of treatment ranking (Minimum value of quadratic Cohen's kappa)	Slight agreement	0	0	0	1	0	1 (1.7%)
	Fair agreement	0	0	1	1	1	3 (5.0%)
	Moderate agreement	0	0	0	2	1	3 (5.0%)
	Substantial agreement	0	4	6	11	1	22 (36.7%)
	Almost perfect agreement	3	8	6	12	2	31 (51.7%)
	Total	3 (5.0%)	12 (20.0%)	13 (21.7%)	27 (45.0%)	5 (8.3%)	60
(C)		Uncertainty of treatment ranking (Average Normalized Entropy)					

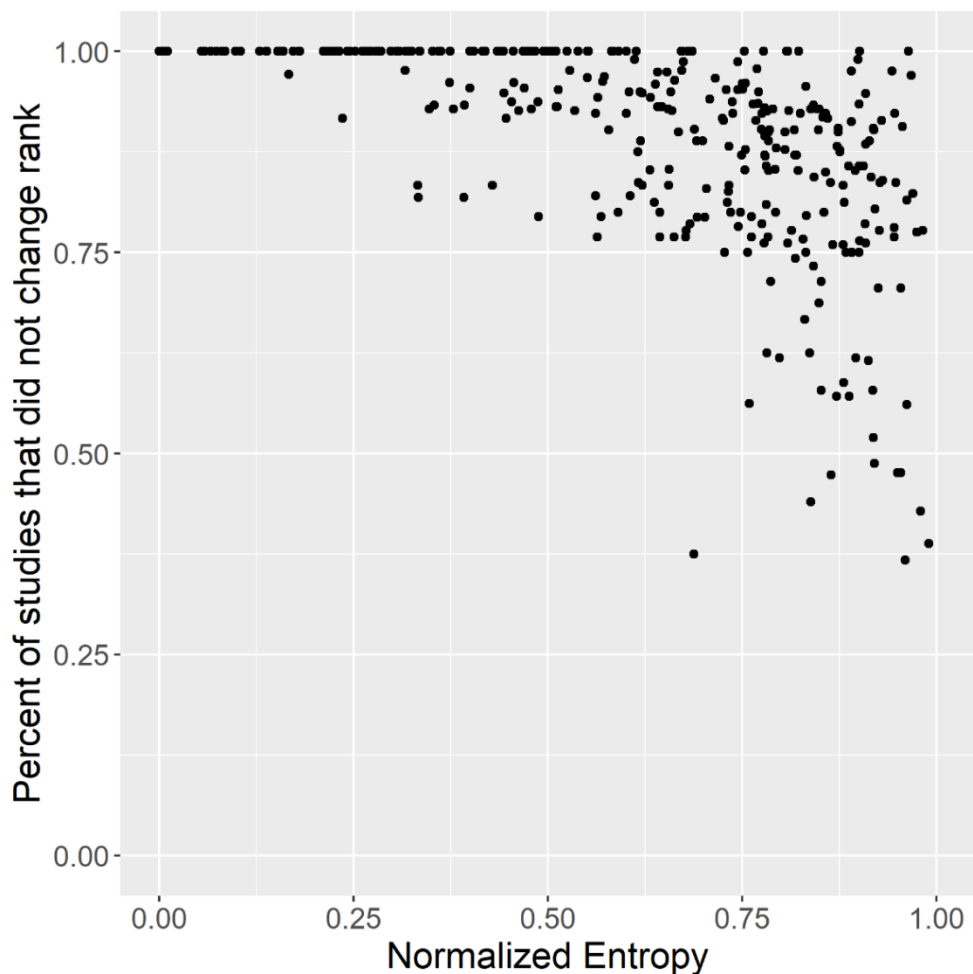


		Very low	Low	Median	High	Very high	<i>Total</i>
Robustness of treatment ranking (Maximum value of quadratic Cohen's kappa)	Slight agreement	0	0	0	0	0	0 (0%)
	Fair agreement	0	0	0	0	0	0 (0%)
	Moderate agreement	0	0	0	0	0	0 (0%)
	Substantial agreement	0	0	0	0	0	0 (0%)
	Almost perfect agreement	3	12	13	27	5	60 (100%)
	<i>Total</i>	3 (5.0%)	12 (20.0%)	13 (21.7%)	27 (45.0%)	5 (8.3%)	60

5.4.2 Treatment-level Association between Uncertainty and Robustness

The 60 NMAs included 348 treatments, and Figure 18 shows the scatterplot for the association between the normalized entropy and the percentage of treatments that did not change rank. Their Pearson's correlation coefficient was -0.59 . Each point represented a treatment. Among 348 treatments, the percentage of trials, the deletion of which did not change rank of the treatment, ranged from 37% to 100%. For those treatments whose ranks were changed by the deletion of a trial, over 25% of them are in the high and very high levels of ranking uncertainty.

Figure 18. Scatter plot of normalized entropy and percentage of treatments that did not change rank for 348 treatments within 60 NMAs



5.4.3 Regression Analysis

For the 60 NMAs, Table 13 shows that the average normalized entropy was inversely associated with the average quadratic weighted Cohen's kappa in both the univariate and the multivariable models with the adjustment of the number of included trials, the number of total participants, the number of treatments, and the type of interventions assessed (Table 5). The inverse association was greater (-0.45 in the univariate model and -0.59 in the multivariate model) when I used the minimum value of quadratic weighted Cohen's kappa. For the 348 treatments, the normalized entropy was also inversely associated with the percentage of trials, the deletion of which did not change rank, in both univariate and multivariable models.

Table 13. Results of each model to explore the association between the robustness of treatment ranking and uncertainty of treatment ranking for 60 NMAs and 348 treatments

	Univariate model	Multivariate model
NMA-level assessment (average quadratic weighted Cohen's kappa)		
Average Normalized Entropy for 60 NMAs	-0.06*	-0.07*†
NMA-level assessment (minimum value of quadratic weighted Cohen's kappa)		
Average Normalized Entropy for 60 NMAs	-0.45*	-0.58*†
Treatment-level assessment		
Normalized Entropy	-0.27*	-0.25*†

*p-value<0.05 †model adjusted for number of included trials, number of total participants, number of treatments, and type of interventions assessed

CHAPTER 6: Discussion and Conclusions



This chapter described the main findings, discussed the strengths and limitations of the analyses, and made the conclusion of the dissertation.

6.1 Using Normalized Entropy to Measure Uncertainty of Rankings

This study demonstrates that Normalized Entropy summarized ranking probabilities into a single measure to compare the uncertainty of treatment ranking, either within the same network or across different networks. I also observed that Normalized Entropy is a more accurate index for the uncertainty of treatment ranking and is more likely to distinguish subtle differences in the levels of ranking uncertainty compared to the width of 95% CI of SUCRA. A more accurate assessment of ranking uncertainty is crucial for interpreting treatment ranking, especially when we are making recommendations for treatments.

Rankograms, cumulative rankograms, the confidence/credible intervals of the mean rank or SUCRA, or IQR of median rank have been used to display the uncertainty of treatment ranking. However, these approaches either did not produce a single index value, or their values may be related to the total number of treatments within the network. Therefore, they cannot distinguish differences in the levels of ranking uncertainty for treatments within the same network meta-analysis, nor can they compare the uncertainties of treatment ranking across different network meta-analyses. Learning the uncertainty of treatment ranking across several network meta-analyses is useful for comparing the level of evidence produced by these network meta-analyses. For treatments with low uncertainty of ranking in a network, sufficient evidence on their efficacy may have been accrued, so we have reasonable confidence in making a

recommendation on the priority of their use. For treatments or networks with high uncertainty of ranking, they may be given greater priority to access research resources to obtain more evidence on their efficacy. Normalized Entropy provides a simple index to compare the uncertainties between treatments within a network or across different networks. Moreover, the calculation of Normalized Entropy required only the ranking probabilities of each treatment, which are usually provided in published network meta-analyses.

6.2 Strengths and Limitation of Normalized Entropy

To quantify the uncertainty of ranking, Normalized Entropy has several advantages over the width of 95% CI of SUCRA. First, the 95% CI of SUCRA requires the study-level data input and can only be estimated through the simulation approach. In contrast, Normalized Entropy can be computed by using the reported ranking probabilities matrix. Second, previous empirical studies showed that the 95% CI of SUCRA is not very informative. In more than one-third NMAs, its width for the top 3 treatments ranged from 0 to 1¹⁵. Example 4 showed that treatments with quite different distributions of ranking probabilities could have the same width of the 95% CI of SUCRA. Due to the widths of 95% CI of SUCRA's discrete nature, the width of 95% CI of SUCRA is less informative than Normalized Entropy in assessing the uncertainty of treatment ranking, particularly for networks with only a few treatments and for treatments in the middle positions. Thirdly, the range of the Normalized Entropy is always from 0 to 1, not affected by the number of treatments. Therefore, its values can be compared across networks with different treatments, while the width of the 95% CI of SUCRA is discrete and affected by the numbers of treatments within networks.

The limitation of Normalized Entropy is that like ranking measures such as SUCRA, it is hard to define how much difference in normalized entropy is large enough to suggest an important difference in ranking uncertainty. The Normalized Entropy is also used as a statistical index in other fields^{69,71,84}, such as decision tree in machine learning, model selection in latent class analysis, and classification for logistic regression. Generally, they use 0.8 as the threshold for 1 minus normalized entropy⁷³, which indicates a good separation of classes when Normalized Entropy is lower than 0.2. The other study suggested to divide the 1 minus Normalized Entropy into four groups: perfect (between 0.8 and 1), high (between 0.6 and 0.8), medium (between 0.4 and 0.6), and low (less than 0.4)⁷².

6.3 Is Providing Uncertainty Intervals in Treatment Ranking Helpful?

Since SUCRA is equal to P-score, defined as one-sided *P*-values, a recent study questioned the usefulness of measuring the uncertainty of ranking statistics⁴⁹. However, we are quantifying the uncertainty of treatment ranking, not for the uncertainty of SUCRA or P-score *per se*. While SUCRA is derived from ranking probabilities and used as an index for determining the ranking of a treatment, Normalized Entropy transforms the distribution of the ranking probabilities into a single index for the uncertainty of the ranking of a treatment. Note that while the uncertainty of the estimates of relative effects is high, the ranking uncertainty can still be low. Suppose that the confidence intervals of differences between the reference treatment A and three other treatments B, C and D are very wide (i.e., the uncertainty of estimates is high), but they do not greatly overlap with each other (Figure 19 (A)); consequently, the uncertainty of treatments ranking is low, and it is therefore straightforward to determine which treatments should be recommended and which

should be avoided. In contrast, when the confidence intervals are narrow (i.e., the uncertainty of estimates is low) but greatly overlapped (Figure 19 (B)), i.e., the differences in point estimates are small, the uncertainty of treatment ranking might still be high.

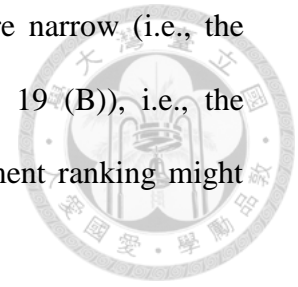
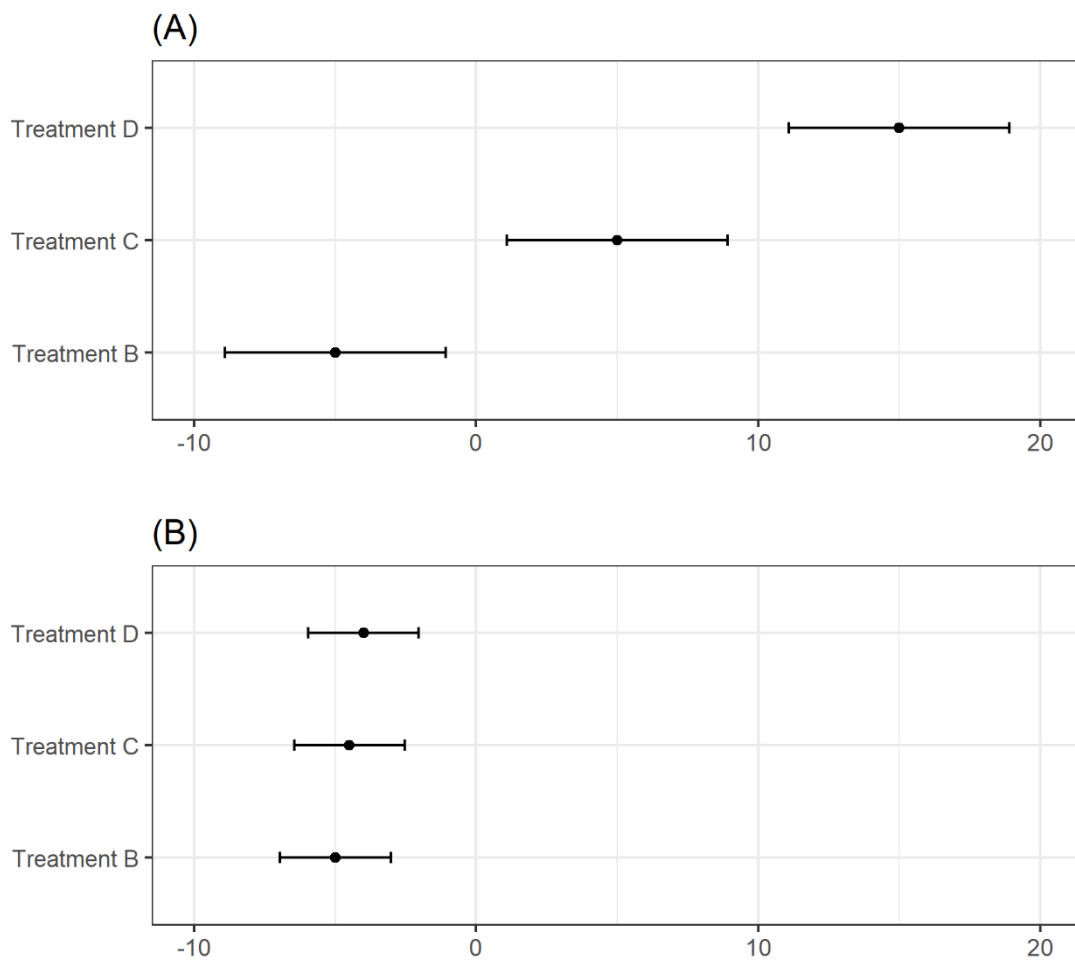


Figure 19. A hypothetical example for the relationship between the precision of estimates and uncertainty of the ranking



6.4 How is Normalized Entropy Related to Variance?

I proposed to use Normalized Entropy in this study, but Variance is another simple index for quantifying uncertainty. The main difference between these two indices is that Variance considers the bimodal distribution of ranking probabilities the most

uncertain scenario, while Entropy considers uniform distribution the most uncertain one. The bimodal distribution of ranking probabilities is uncommon; yet, it may be seen when two treatments have similar effect sizes, but one has a much wider confidence interval, thereby yielding greater distribution of ranking probabilities⁴⁶. As the bimodal distribution is considered the most uncertain scenario, the range of the Normalized Variance is quite limited for unimodal distribution of ranking probabilities, thereby being less able to distinguish different levels of uncertainty than Normalized Entropy under most scenarios. Instead, Normalized Standard Deviation can detect the bimodal distribution and have a similar pattern with Normalized Entropy when the uncertainty is high. However, when the uncertainty is low, which means that the ranking probability is concentrated in a specific rank, it is still hard to distinguish its uncertainty levels by using Normalized Standard Deviation.

6.5 High Robustness Does Not Always Imply Low Uncertainty of Treatment Rankings

In the empirical study analyzing 60 NMAs, as I expected, the treatment ranking of an NMA with low uncertainty of ranking is unlikely to be altered by subtle changes of the database; however, when the uncertainty of ranking is high, the robustness of ranking showed a wide range. Therefore, the high robustness of ranking does not always correspond to the low uncertainty of ranking, indicating that robustness and uncertainty are two correlated but distinctive concepts.

With the rapid growth in the number of publications of NMAs, a careful evaluation of the reliability of the treatment rankings is crucial for applying results from an NMA to making clinical decision^{85,86}. However, most NMAs neither evaluated the uncertainty nor the robustness of the ranking⁸⁷, or only use one of them to evaluate the reliability

of ranking⁸⁸⁻⁹⁰. Furthermore, the good robustness of treatment ranking has often been interpreted as high reliability of ranking. In contrast, the results showed that robustness and uncertainty of ranking are not perfectly correlated. When there is no outlying trial within the network, the robustness of ranking may be high, but the ranking can still be of great uncertainty. Therefore, the evaluation of the reliability of ranking could be conducted in two steps. First, evaluate the uncertainty of ranking. If the uncertainty is low, we could expect the ranking to be reliable. If the uncertainty of ranking is high, then the robustness of ranking can help see whether a single outlying trial influenced the overall ranking.

6.6 Evaluation at NMA-level, Treatment-level, and Trial-Level

The squared weighted Cohen's kappa was recommended to quantify the agreement between treatment rankings²⁰. It measures the changes in ranking by assigning a greater penalty to a greater difference in ranking position. However, Cohen's kappa is an NMA-level statistic and cannot be used for the evaluation of ranking robustness at the treatment level. We, therefore, used the percentage of trials, the deletion of which does not change the rank of treatment, to assess the robustness of ranking for individual treatments. At the NMA-level, we computed the minimum, average, and maximum values of the quadratic weighted Cohen's kappa to represent the robustness of ranking for an NMA. The maximum and average value of quadratic weighted Cohen's kappa are less useful since we want to know the maximum impact caused by deletion a trial within the NMA. Therefore, we recommend using the minimum value of weighted Cohen's kappa to represent the overall robustness of ranking at the NMA level.

The Normalized Entropy we used to quantify ranking uncertainty can provide

treatment or NMA-level information while ranking robustness can additionally provide trial-level information. Different levels of information are all needed when we evaluate the ranking of NMAs. We may want to find out which NMA or treatment may need to gather more evidence and which trial may affect ranking the most and is needed to flag out for further investigation.

6.7 Limitations of the Study of Robustness and Uncertainty of Ranking

There are some limitations to this analysis. Firstly, NMAs included in this study are those with two or more trials in each arm, i.e., the selected NMAs contained more data. Since evaluating the robustness of ranking needs to remove each trial in turn, NMAs were excluded if removing a trial would break the network. Therefore, alternative approaches are required to assess the robustness of ranking for NMAs excluded from our evaluation. Secondly, I only included those NMAs using odds ratio and mean difference as outcome measures. Further analysis can be conducted to compare these two metrics for other outcome measures.

6.8 Presentation of Uncertainty with Ranking

By visually displaying the rankings and their uncertainties in a colored table or scatter plot, the information can be simplified, but its interpretability can still be maintained. The advantage of using normalized entropy as an index for uncertainty of ranking is that the calculation does not require study-level data and is easy to incorporate into the current coding process. Normalized Entropy, therefore, provides an objective assessment of the ranking uncertainty and whether the evidence is now sufficient to make recommendations on the relative efficacy of treatments or more evidence is still required.

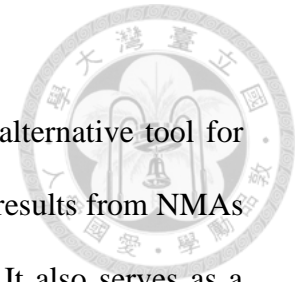
6.9 Conclusions

In this dissertation, I demonstrate that Normalized Entropy is an alternative tool for measuring the uncertainty of ranking, improving the translation of results from NMAs to clinical practice and avoid naïve interpretation of the ranking. It also serves as a tool for identifying which treatments require more evidence to reduce the uncertainty, and it can also be used to compare the uncertainty of treatment ranking across the different networks. I, therefore, recommend Normalized Entropy to be included in the presentation and interpretation of results, such as GRADE's summary of finding table⁸⁵, for future NMAs.

Moreover, this dissertation also showed that good robustness of ranking does not always correspond to low uncertainty. Therefore, although the robustness of the ranking can find the trial that has the greatest impact on the ranking, the high robustness does not mean that the ranking would not easily change when new trials are added in the future.

6.10 Future work

Although Normalized Entropy is proposed to quantify the uncertainty of ranking, it can be explored whether this indicator has other possible uses. For example, could it be used to classify treatment nodes?^{91,92} According to the NMA methodology review, the node-making process currently is still lacking clear guidance, and only 10% of 116 NMAs discussed the concept of node making process⁹³. When there are questions about whether similar but not identical interventions should be lumped together or split into different nodes, Normalized Entropy might be a useful indicator to facilitate decision making.



REFERENCE

1. Fleetwood K, Glanville J, McCool R, et al. A Review of the Use of Network Meta-Analysis in Nice Single Technology Appraisals. *Value Health* 2016; **19**(7): A348-A.
2. Kanters S, Ford N, Druyts E, Thorlund K, Mills EJ, Bansback N. Use of network meta-analysis in clinical guidelines. *B World Health Organ* 2016; **94**(10): 782-4.
3. Laws A, Tao R, Wang SS, Padhiar A, Goring S. A Comparison of National Guidelines for Network Meta-Analysis. *Value Health* 2019; **22**(10): 1178-86.
4. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004; **23**(20): 3105-24.
5. Tu YK. Use of Generalized Linear Mixed Models for Network Meta-analysis. *Med Decis Making* 2014; **34**(7): 911-8.
6. White IR. Network meta-analysis. *Stata J* 2015; **15**(4): 951-85.
7. Schünemann HJ, Vist GE, Higgins JP, et al. Interpreting results and drawing conclusions. *Cochrane handbook for systematic reviews of interventions* 2019: 403-31.
8. Schünemann HJ, Oxman AD, Higgins JP, Vist GE, Glasziou P, Guyatt GH. Presenting results and 'Summary of findings' tables. *Cochrane handbook for systematic reviews of interventions* 2008; **5**: 0.
9. Yepes-Nunez JJ, Li SA, Guyatt G, et al. Development of the summary of findings table for network meta-analysis. *J Clin Epidemiol* 2019; **115**: 1-13.
10. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011; **64**(2): 163-71.
11. Wang Z, Carter RE. Ranking of the most effective treatments for cardiovascular

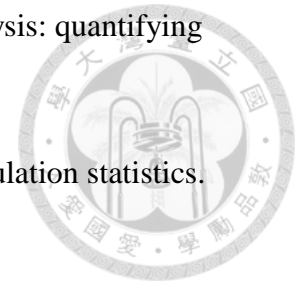


disease using SUCRA: Is it as sweet as it appears? : SAGE Publications Sage UK: London, England; 2018.



12. Bafeta A, Trinquart L, Seror R, Ravaud P. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. *BMJ* 2013; **347**: f3675.
13. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. *Ann Intern Med* 2015; **162**(11): 777-84.
14. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision-making: John Wiley & Sons; 2018.
15. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016; **164**(10): 666-+.
16. Faltinsen EG, Storebo OJ, Jakobsen JC, Boesen K, Lange T, Gluud C. Network meta-analysis: the highest level of medical evidence? *BMJ Evid Based Med* 2018; **23**(2): 56-9.
17. Mbuagbaw L, Rochweg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev-London* 2017; **6**.
18. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomized trials. *Annals of internal medicine* 2016; **164**(10): 666-73.
19. Wu Y-C, Shih M-C, Tu Y-K. Using Normalized Entropy to Measure Uncertainty of Rankings for Network Meta-analyses. *Med Decis Making* 2021; **41**(6): 706-13.
20. Daly CH, Neupane B, Beyene J, Thabane L, Straus SE, Hamid JS. Empirical

evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa. *BMJ open* 2019; **9**(9): e024625.



21. Simpson RJS, Pearson K. Report on certain enteric fever inoculation statistics. *Brit Med J* 1904; **1904**: 1243-6.
22. Glass GV. Primary, Secondary, and Meta-Analysis of Research 1. *Educational researcher* 1976; **5**(10): 3-8.
23. Shah HM, Chung KC. Archie Cochrane and His Vision for Evidence-Based Medicine. *Plast Reconstr Surg* 2009; **124**(3): 982-8.
24. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002; **21**(16): 2313-24.
25. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials. *Med Decis Making* 2013; **33**(5): 607-17.
26. Jansen JP, Fleurence R, Devine B, et al. 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 Health* 2011; **14**(4): 417-28.
27. Jones B, Roger J, Lane PW, et al. Statistical approaches for conducting network meta-analysis in drug development. *Pharm Stat* 2011; **10**(6): 523-31.
28. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. *Health Technol Asses* 2005; **9**(26): 1-+.
29. Piepho HP, Williams ER, Madden LV. The Use of Two-Way Linear Mixed Models in Multitreatment Meta-Analysis. *Biometrics* 2012; **68**(4): 1269-77.
30. Higgins JP, Welton NJ. Network meta-analysis: a norm for comparative

effectiveness? *The Lancet* 2015; **386**(9994): 628-30.

31. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997; **50**(6): 683-91.

32. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med* 1996; **15**(24): 2733-49.

33. Higgins JP, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* 2012; **3**(2): 98-110.

34. White IR, Barrett JK, Jackson D, Higgins JP. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods* 2012; **3**(2): 111-25.

35. Bafeta A, Trinquart L, Seror R, Ravaud P. Reporting of results from network meta-analyses: methodological systematic review. *BMJ* 2014; **348**: g1741.

36. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical Tools for Network Meta-Analysis in STATA. *Plos One* 2013; **8**(10).

37. Schwarzer G, Carpenter JR, Rücker G. *Meta-analysis with R*: Springer; 2015.

38. Jackson D, Barrett JK, Rice S, White IR, Higgins JP. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Statistics in medicine* 2014; **33**(21): 3639-54.

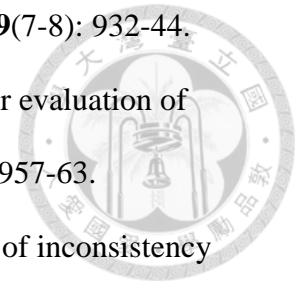
39. Tu Y-K. Using generalized linear mixed models to evaluate inconsistency within a network meta-analysis. *Value Health* 2015; **18**(8): 1120-5.

40. Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc* 2006; **101**(474): 447-59.

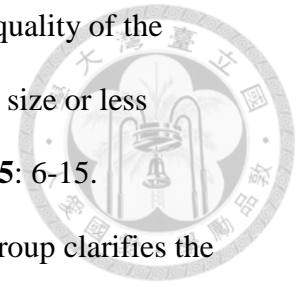
41. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed



- treatment comparison meta-analysis. *Statistics in Medicine* 2010; **29**(7-8): 932-44.
42. Yu-Kang T. Node-splitting generalized linear mixed models for evaluation of inconsistency in network meta-analysis. *Value Health* 2016; **19**(8): 957-63.
43. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. *Int J Epidemiol* 2013; **42**(1): 332-45.
44. Rochweg B, Alhazzani W, Sindi A, et al. Fluid Resuscitation in Sepsis A Systematic Review and Network Meta-analysis. *Ann Intern Med* 2014; **161**(5): 347-+.
45. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *Bmc Med* 2013; **11**.
46. Rucker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *Bmc Med Res Methodol* 2015; **15**.
47. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998; **316**(7146): 1701-4; discussion 5.
48. Langville AN, Meyer CD. Who's# 1?: the science of rating and ranking: Princeton University Press; 2012.
49. Veroniki AA, Straus SE, Rucker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *Journal of clinical epidemiology* 2018; **100**: 122-9.
50. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J Roy Stat Soc a Sta* 1996; **159**: 385-409.
51. Group GW. Grading quality of evidence and strength of recommendations. *BMJ: British Medical Journal* 2004; **328**(7454): 1490.



52. Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *Journal of clinical epidemiology* 2016; **75**: 6-15.
53. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *Journal of clinical epidemiology* 2017; **87**: 4-13.
54. Zhang Y, Coello PA, Guyatt GH, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences—inconsistency, imprecision, and other domains. *Journal of clinical epidemiology* 2019; **111**: 83-93.
55. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 2011; **64**(4): 383-94.
56. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016; **164**(10): 666-73.
57. Goldstein H. Living by the evidence. *Significance* 2020; **17**(1): 38-40.
58. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis. *Statistics in medicine* 2015; **34**(19): 2695-707.
59. Phillippo DM, Dias S, Welton NJ, Caldwell DM, Taske N, Ades AE. Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. *Ann Intern Med* 2019; **170**(8): 538-+.
60. Caldwell DM, Ades AB, Dias S, et al. A threshold analysis assessed the credibility of conclusions from network meta-analysis. *Journal of Clinical Epidemiology* 2016; **80**: 68-76.

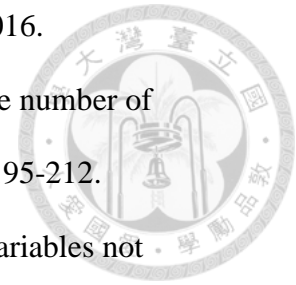


61. Norton EC, Miller MM, Wang JJ, Coyne K, Kleinman LC. Rank Reversal in Indirect Comparisons. *Value Health* 2012; **15**(8): 1137-40.
62. van Valkenhoef G, Ades AE. Evidence Synthesis Assumes Additivity on the Scale of Measurement: Response to "Rank Reversal in Indirect Comparisons" by Norton et al. *Value Health* 2013; **16**(2): 449-51.
63. Gao YN, Wu YC, Lin SY, Chang JZC, Tu YK. Short-term efficacy of minimally invasive treatments for adult obstructive sleep apnea: A systematic review and network meta-analysis of randomized controlled trials. *J Formos Med Assoc* 2019; **118**(4): 750-65.
64. Nikolakopoulou A, Mavridis D, Furukawa TA, et al. Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. *BMJ* 2018; **360**: k585.
65. Nikolakopoulou A, Mavridis D, Egger M, Salanti G. Continuously updated network meta-analysis and statistical monitoring for timely decision-making. *Stat Methods Med Res* 2018; **27**(5): 1312-30.
66. Shannon CE. A Mathematical Theory of Communication. *At&T Tech J* 1948; **27**(3): 379-423.
67. Kumar U, Kumar V, Kapur JN. Normalized Measures of Entropy. *Int J Gen Syst* 1986; **12**(1): 55-69.
68. Wu J, Tan YJ, Deng HZ, Zhu DZ. Normalized entropy of rank distribution: a novel measure of heterogeneity of complex networks. *Chinese Phys* 2007; **16**(6): 1576-80.
69. Kulisiewicz M, Kazienko P, Szymanski BK, Michalski R. Entropy Measures of Human Communication Dynamics. *Sci Rep-Uk* 2018; **8**.
70. Wickrama KK, Lee TK, O'Neal CW, Lorenz FO. Higher-order growth curves

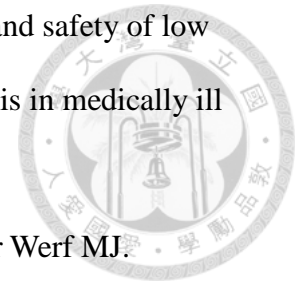


and mixture modeling with Mplus: A practical guide: Routledge; 2016.

71. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification* 1996; **13**(2): 195-212.
72. Clark SL, Muthén B. Relating latent class analysis results to variables not included in the analysis. 2009.
73. Cummings KD, Petscher Y. The fluency construct: Curriculum-based measurement concepts and applications: Springer; 2015.
74. Salanti G, Ades A, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of clinical epidemiology* 2011; **64**(2): 163-71.
75. Papakonstantinou T. nmadb: network meta-analysis database API. 2019. <https://cran.r-project.org/package=nmadb> (accessed May 16 2019).
76. Petropoulou M, Nikolakopoulou A, Veroniki A-A, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of clinical epidemiology* 2017; **82**: 20-8.
77. Bansal M, Farrugia A, Balboni S, Martin G. Relative survival benefit and morbidity with fluids in severe sepsis - a network meta-analysis of alternative therapies. *Curr Drug Saf* 2013; **8**(4): 236-45.
78. Palmer SC, Saglimbene V, Mavridis D, et al. Erythropoiesis-stimulating agents for anaemia in adults with chronic kidney disease: a network meta-analysis. *Cochrane Db Syst Rev* 2014; (12).
79. Price R, MacLennan G, Glen J, Su DC. Selective digestive or oropharyngeal decontamination and topical oropharyngeal chlorhexidine for prevention of death in general intensive care: systematic review and network meta-analysis. *BMJ* 2014; **348**: g2197.



80. Dooley C, Kaur R, Sobieraj DM. Comparison of the efficacy and safety of low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients. *Curr Med Res Opin* 2014; **30**(3): 367-80.
81. Zenner D, Beer N, Harris RJ, Lipman MC, Stagg HR, Van Der Werf MJ. Treatment of latent tuberculosis infection: an updated network meta-analysis. *Annals of internal medicine* 2017; **167**(4): 248-55.
82. Xing A, Lin L. Effects of treatment classifications in network meta-analysis. *Research Methods in Medicine & Health Sciences* 2020; **1**(1): 12-24.
83. Rochweg B, Alhazzani W, Sindi A, et al. Fluid resuscitation in sepsis: a systematic review and network meta-analysis. *Ann Intern Med* 2014; **161**(5): 347-55.
84. Larose C, Harel O, Kordas K, Dey DK. Latent class analysis of incomplete data via an entropy-based criterion. *Statistical methodology* 2016; **32**: 107-21.
85. Yepes-Nuñez JJ, Li S-A, Guyatt G, et al. Development of the summary of findings table for network meta-analysis. *Journal of Clinical Epidemiology* 2019; **115**: 1-13.
86. Carroll K, Hemmings R. On the need for increased rigour and care in the conduct and interpretation of network meta-analyses in drug development. *Pharm Stat* 2016; **15**(2): 135-42.
87. Bafeta A, Trinquart L, Seror R, Ravaud P. Reporting of results from network meta-analyses: methodological systematic review. *Bmj-Brit Med J* 2014; **348**.
88. Zhang J, Yuan Y, Chu H. The Impact of Excluding Trials from Network Meta-Analyses - An Empirical Study. *PLoS One* 2016; **11**(12): e0165889.
89. Noma H, Goshu M, Ishii R, Oba K, Furukawa TA. Outlier detection and influence diagnostics in network meta-analysis. *Res Synth Methods* 2020.
90. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis.



Stat Med 2015; **34**(19): 2695-707.

91. Del Giovane C, Vacchi L, Mavridis D, Filippini G, Salanti G. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Stat Med* 2013; **32**(1): 25-39.

92. Rochwerg B, Alhazzani W, Gibson A, et al. Fluid type and the use of renal replacement therapy in sepsis: a systematic review and network meta-analysis.

Intensive care medicine 2015; **41**(9): 1561-71.

93. James A, Yavchitz A, Ravaud P, Boutron I. Node-making process in network meta-analysis of nonpharmacological treatment are poorly reported. *Journal of clinical epidemiology* 2018; **97**: 95-102.

