SOLUTIONS TO HIGH-PRIORITY CHALLENGES IN SYSTEMATIC REVIEWS:

Network meta-analysis and integrating randomized and non-randomized evidence

SOLUTIONS TO HIGH-PRIORITY CHALLENGES IN SYSTEMATIC REVIEWS:

Network meta-analysis and integrating randomized and non-randomized evidence

By JUAN JOSE YEPES-NUÑEZ, M.D, M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2018)

Hamilton, Ontario (Health Research Methodology program)

TITLE: Presentation and methods issues in network meta-analysis and pairwise comparisons

AUTHOR: Juan José Yepes-Nuñez, M.D., M.Sc. (McMaster University, Hamilton, Canada)

SUPERVISOR: Professor Holger J Schünemann MD, PhD, MSc, FRCP(C)

NUMBER OF PAGES: xxvii, 292

LAY ABSTRACT

Systematic reviews (SR) are a summary of studies that address a particular clinical question. Frequently, SR are complemented with a statistical aggregation of results of individual studies to produce a single estimate. Summary of findings (SoF) tables are designed to present the most relevant information of systematic reviews and meta-analysis. However, it is unknown how to present network meta-analysis (NMA) findings in SoF tables. Another challenge relates to the integration of randomized controlled trials (RCTs) and non-randomized (NRS) studies. Methodological challenges in systematic reviews need to be addressed through careful research.

In our study, we appraised how NMA were conducted, and how they presented their main findings. We designed two versions of SoF tables to present NMA findings. Moreover, we conducted two systematic reviews that included RCTs and NRS to address potential challenges in analyzing and presenting their findings.

ABSTRACT

Systematic reviews (SR) and meta-analysis (MA) of randomised controlled trials (RCT) are the trustworthy sources of evidence. However, most systematic reviews focus on pair-wise comparisons. Network-meta-analysis (NMA) offers quantitative methods of integrating data from all the available comparisons of many different treatments for each outcome. In a systematic review of interventions, Summary of Findings (SoF) tables present the main findings of a review in a transparent and simple form. However, it is unknown how to present NMA findings in a tabular format. Moreover, systematic reviews and meta-analysis of interventions can summarize bodies of evidence from randomized and non-randomized studies (NRS). Integrating both sources of evidence in a single study can be challenging particularly in the context of assessing the certainty of the evidence, as well as presenting findings of both RCTs and NRS sources of evidence.

In our study, we described how 276 NMA were conducted and how authors reported their main findings. We also conducted 32 interviews with users of NMAs and we designed two final NMA-SoF tables. Furthermore, we conducted two systematic reviews that included RCTs and NRS to address methodological challenges.

Based on our results, we developed two NMA-SoF table formats to report the main findings of NMAs. The final format was appealing for users and allowed them to better understand NMA findings. Assessment of quality of individual NRS remains challenging and further research is needed to increase its appropriateness in systematic reviews of NRS. We determined that quality assessment of individual NRS was particularly challenging to implement due to the complexity of NRS evaluation tools. Our evaluation revealed that effect estimates of RCTs and NRS were better presented separately.

ACKNOWLEDGEMENTS

For my parents, who are my inner strength and my mental support. Mom and Dad, I love you both!

Achieving my PhD degree was possible not only due to personal mental and physical strength but also it was achievable because of the companionship, thoughts, and energy of many people around me. For all my family, academic colleagues and friends who supported me all along this way, all my love and thankfulness.

To my supervisor, Dr. Holger Schünemann, who has taught me how to improve myself and encouraged me during my PhD. His supervision and example has shown me there is no limit in the process of learning and building academic knowledge. Many thanks for all your support, patience, and understanding throughout all these years.

To my committee members, Dr. Gordon Guyatt, Dr. Jan Brozek, and Dr. Joseph Beyene, who with their knowledge, kindness, and unconditional support taught me to work with others, to trust in others, but most importantly to believe in others. A special acknowledgement to Dr. Guyatt who also helped me to improve myself by teaching me how to think about and resolve academic problems, who offered me his company in moments of crisis, and most importantly, who helped me to appreciate human beings and life in a broader and more complex manner.

To my colleagues and friends in the Health Research Methodology program and in all Canada. They showed me there is no difference between being academic colleagues and being friends. Real friendships are built on the basis of trustworthiness and support, and all of you always facilitated my understanding of academic concepts and also of living a full life of happiness and care.

To my partner, you always have been there, beside me. You listen to me and advise me. Having you with me in this process has helped to keep me up during moments of crisis, and you also keep reminding me there is a life outside of an academic career.

To my sister, your support during difficult times was valuable to get strength to continue my work during my PhD.

To my family. *Madre y padre*, you are all my strength and the reason for moving forward in my life. Your everyday unconditional support, despite any issues, has showed me the magnificence of love and care for a son. I have no words to show how grateful I am for everything you have given me. This is not only my own achievement, this is OUR achievement. To both of you, always love and gratitude!

Juan José Yepes-Nuñez

Hamilton, Canada. July 2018

PREFACE

This dissertation has been conducted as a “sandwich thesis” and includes four individual manuscripts submitted for peer review and publication in scientific journals. These are:

1. **Chapter 1:** Introduction of the thesis**.**
2. **Chapter 2:** Evaluation of 276 network meta-analysis publications: statistical approaches, presentation of findings, certainty of evidence assessments and interpretation of findings partially tackled.
3. **Chapter 3:** Development of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Summary of Findings (SoF) Table for Network Meta-analysis.
4. **Chapter 4:** Vitamin D supplementation in primary allergy prevention: Systematic review of randomized and non-randomized studies.
5. **Chapter 5:** Pharmacological thromboprophylaxis in patients undergoing neurosurgical interventions for preventing venous thromboembolism: a systematic review of randomized and non-randomized studies**.**
6. **Chapter 6:** Conclusion.

COPYRIGHT PERMISSION

Dr. Yepes-Nuñez has provided written permission to include copyright material in this Ph.D. thesis from the copyright holder. The permission includes a grant of an irrevocable, non-exclusive license to McMaster University and to Library and Archives Canada to reproduce the material as part of the thesis.

TABLE OF CONTENTS

LAY ABSTRACT iii

ABSTRACT iv

ACKNOWLEDGEMENTS vi

PREFACE viii

COPYRIGHT PERMISSION ix

LIST OF FIGURES xvii

LIST OF TABLES xviii

LIST OF APPENDICES xxi

LIST OF ABBREVIATIONS AND SYMBOLS xxiv

DECLARATION OF ACADEMIC ACHIEVEMENT xxvi

CHAPTER 1. INTRODUCTION 1

Evidence synthesis: From Meta-analysis to Network-Meta-analysis 2

Summary presentation of findings of Network-Meta-analysis 2

Methodological and presentational issues in systematic reviews and meta-analysis of RCTs and NRS 4

Goals and scope 5

Thesis overview 6

REFERENCES 7

CHAPTER 2. EVALUATION OF 276 NETWORK META-ANALYSIS PUBLICATIONS: STATISTICAL APPROACHES, PRESENTATION OF FINDINGS, CERTAINTY OF EVIDENCE ASSESSMENTS AND INTERPRETATION OF FINDINGS ARE ONLY PARTIALLY TACKLED 12

PREFACE TO CHAPTER 2 13

ABSTRACT 16

HIGHLIGHTS 18

INTRODUCTION 19

METHODS 20

Eligibility criteria 20

Search strategy 21

Network Meta-analysis selection 21

Data extraction and synthesis 22

RESULTS 23

Description of included systematic reviews 23

Statistical analysis in selected NMAs 24

Statistical methods for heterogeneity and transitivity evaluations 25

Statistical methods for incoherence (or inconsistency) evaluation 26

Subgroup analysis and meta-regression in selected NMA 27

Sensitivity analysis in selected NMA 27

Reporting NMA findings in selected NMA primary outcomes 28

Presentation of NMA findings 29

Assessing the certainty (or quality) of the evidence in NMA 31

Interpretation of Findings in selected primary outcomes 33

DISCUSSION 34

Main findings 34

Strength and limitations 34

Our results in the context of previous research 38

Implications for practice and research 46

Conclusions 46

ACKNOWLEDGEMENTS 47

REFERENCES 47
FIGURES 53
TABLES 55
APPENDIXES 61
CHAPTER 3. DEVELOPMENT OF A SUMMARY OF FINDINGS (SOF) TABLE FOR NETWORK META-ANALYSIS 73
PREFACE TO CHAPTER 3 74
ABSTRACT 77
HIGHLIGHTS 79
INTRODUCTION 80
METHODS 81
Development of initial NMA SoF table format 82
Brainstorming meetings 82
User testing 82
Participants 84
Ethical considerations 88
RESULTS 88
Initial NMA-SoF table formats 88
Brainstorming meetings 89
User testing 90
Final NMA-SoF table 97
Considerations of benefit and harms and continuous outcomes in NMA-SoF tables 98
A Second final NMA-SoF table 99
DISCUSSION 99
Main findings 99

The evolution of NMA-SoF formats 100
An alternative to our final NMA-SoF table 102
Strengths and limitations 103
Conclusion, and further research 104
ACKNOWLEDGEMENTS 104
REFERENCES 105
FIGURES 111
TABLES 112
APPENDIXES 122
CHAPTER 4: VITAMIN D SUPPLEMENTATION IN PRIMARY ALLERGY PREVENTION. A SYSTEMATIC REVIEW OF RANDOMIZED AND NON-RANDOMIZED STUDIES 151
PREFACE TO CHAPTER 4 152
ABSTRACT 156
HIGHLIGHTS 158
INTRODUCTION 159
METHODS 160
Inclusion and Exclusion Criteria 160
Search methods 162
Study Selection and Data Extraction 163
Certainty in the body of evidence (confidence in the estimates of effects or quality of evidence) 163
Data analysis 164
Data Synthesis 165
Subgroup analysis 165
Sensitivity analysis 165

GRADE assessment of the overall certainty in the body of evidence by outcome 166
RESULTS 166
Risk of bias in studies included 167
Effect of interventions 168
Sensitivity analysis 173
DISCUSSION 173
Suggestions for future research 177
CONCLUSION 178
REFERENCES 179
FIGURE 190
TABLES 191
APPENDIXES 208
CHAPTER 5. PHARMACOLOGICAL THROMBOPROPHYLAXIS IN PATIENTS UNDERGOING NEUROSURGICAL INTERVENTIONS FOR PREVENTING VENOUS THROMBOEMBOLISM: A SYSTEMATIC REVIEW OF RANDOMIZED AND NON-RANDOMIZED STUDIES. 214
PREFACE TO CHAPTER 5 215
ABSTRACT 218
HIGHLIGHTS 220
INTRODUCTION 221
METHODS 223
Inclusion and exclusion criteria 223
Search methods 225
Study selection and data extraction 225

Certainty in the body of evidence (confidence in the estimates of effects or certainty of evidence) 226
Data analysis 226
Data synthesis 227
Subgroup analysis 228
Sensitivity analysis 228
GRADE assessment of the overall certainty in the body of evidence by outcome 228
RESULTS 229
Risk of bias in studies included 231
Effect of interventions 231
Subgroup analysis 235
Sensitivity analysis 235
DISCUSSION 235
Limitations 236
Findings in other publications 237
Study implications 240
Suggestions for future research 241
CONCLUSION 241
REFERENCES 242
FIGURE 247
TABLES 248
APPENDIXES 255
CHAPTER 6. CONCLUSION 280
Summary of Findings 281
Implications for researchers, guideline developers, policy makers, and clinicians 283

Implications for researchers, guideline developers, policy makers 283
Implications for clinicians and other users of evidence synthesis 285
Strengths and challenges of this work 286
Further research directions 288
Final remarks 289
REFERENCES..... 290

LIST OF FIGURES

Chapter Two

Figure 1. PRISMA flow chart of study selection: Page 52

Figure 2. Clinical areas involved in 276 NMA publications: Page 53

Chapter Three

Figure 1. Phases conducted to develop the GRADE NMA-SoF table: Page 110

Chapter Four

Figure 1. PRISMA flow chart of study selection: Page 189

Chapter Five

Figure 1. PRISMA flow chart of study selection: Page 247

LIST OF TABLES

Chapter Two

Table 1. Characteristic of 276 NMA publications included from 2014 to 2015: Page 54

Table 2. Frequency of summary effect estimates and measure of uncertainty in 276 NMA publications: Page 55

Table 3. Frequency of methods applied to report NMA rank probabilities in 156 NMA publications: Page 56

Table 4. Frequency of other tools implemented to assess risk of bias or quality of individual RCT in 67 NMA publications: Page 58

Table 5. Frequency of other tools implemented to interpret NMA findings in 127 NMA publications: Page 59

Chapter Three

Table 1. Characteristics of participants: Page 111

Table 2. NMA-SoF table final format: Page 112

Table 3. NMA-SoF table template for dichotomous outcomes: Page 114

Table 4. NMA-SoF table format reporting multiple outcome information for multiple comparison treatments: Page 119

Chapter Four

Table 1. Characteristics of included studies: Page 190

Table 2. Summary of Findings table for Vitamin D versus no vitamin D supplementation in PREGNANT women for the primary prevention of allergy in their children and other important outcomes: Page 194

Table 3. Summary of Findings table for Vitamin D versus no vitamin D supplementation in BREASTFEEDING women for the primary prevention of allergy in their children and other important outcomes: Page 198

Table 4. Summary of Findings table for Vitamin D versus no vitamin D supplementation in INFANTS for the primary prevention of allergy and other important outcomes: Page 201

Table 5. Summary of Findings table for Vitamin D versus no vitamin D supplementation in CHILDREN for the primary prevention of allergy and other important outcomes: Page 204

Chapter Five

Table 1. Characteristics of studies included: Page 248

Table 2. Summary of Finding table: Page 251

LIST OF APPENDICES

Chapter Two

Appendix A. Search strategies: Page 60

Appendix B. List of included and excluded studies: Page 71

Chapter Three

Appendix A. Protocol of user testing: Page 121

Appendix B. First NMA-SoF table formats: Page 132

Appendix C. NMA-SoF table format developed for conducting user testing rounds. Format 1: Page 138

Appendix D. NMA-SoF table format developed for conducting user-testing rounds. Format 2: Page 141

Appendix E. NMA-SoF table format. Format 3: Page 144

Appendix F. NMA-SoF table template for continuous outcomes: Page 147

Chapter Four

Appendix A. Search strategies: Page 207

Appendix B. Characteristics of excluded studies: Page 207

Appendix C. Summary of Risk of Bias (RoB) in RCTs and NRS: Page 208

Appendix D. Evidence profiles: Page 212

Appendix E. Forest plots of meta-analysis for each research question: Page 212

Chapter Five

Appendix A. Outcome definitions for ASH. -page 255

Appendix B. Search strategies for randomized and non-randomized studies. -page 259

Appendix C. List of studies excluded. -page 259

Appendix D. Summary of Risk of Bias in randomized control trials for RCT. -page 260

Appendix E. Summary of Risk of Bias in non-randomized studies. -page 261

Appendix F. Evidence profile. -page 266

Appendix G. Forest plots of meta-analysis for each outcome for each type of study design. -page 272

Appendix H. Subgroup analysis by the of neurosurgical intervention in RCTs. -page 276

LIST OF ABBREVIATIONS AND SYMBOLS

AHRQ: Agency for Healthcare Research and Quality

AMED: Allied and Complementary Medicine

AMSTAR: A MeaSurement Tool to Assess systematic Reviews

ASH: American Society of Haematology

DSR: Database of Systematic Reviews

CMR: Cochrane Methodology Register

CINAHL: Cumulative Index to Nursing and Allied Health Literature

DARE: Database of Abstract of Reviews of Effects

DVT: deep venous thrombosis

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HTA: Health Technology Assessment

IQR: interquartile rank

ISPOR: International Society for Pharmacoeconomics and Outcomes Research

LILACS: Latin American and Caribbean System on Health Sciences Information

MA: Meta-analysis

NHSEED: NHS Economic Evaluation Database

NMA: Network Meta-analysis

NRS: non-randomized studies

PE: pulmonary embolism

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RCT: randomized controlled trials

RoB: risk of bias

ROBINS: Risk of bias in non-randomized studies of Interventions

SR: Systematic reviews

SoF: Summary of findings

SUCRA: surface under the cumulative ranking curve

WAO: World Allergy Organization

DECLARATION OF ACADEMIC ACHIEVEMENT

I declare that I, jointly with my supervisor, Professor Holger J. Schünemann, played the primary role in the conception, design, and execution of the studies here included. We obtained feedback and advice from Professors Guyatt, Brozek, and Beyene, as well as from members of the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group, the World Allergy Organization (WAO), and the American Society of Haematology (ASH).

This dissertation is original research that I conducted. I am the principle contributor and first author of all the manuscripts contained in this dissertation.

I am responsible and made the following contributions in all projects included in this dissertation: design, conception, analysis, and writing of materials; I designed the search strategy, screening, and data extraction for the network meta-analysis (NMA) systematic survey; I designed the interview scripts, data extraction and analysis for recording feedback from users of NMA Summary of Finding (SoF) tables, as well as I conducted all the interviews and I designed all the NMA-SoF tables formats; I designed the search strategy, screening, and data extraction for both systematic reviews and meta-analysis. I reviewed comments and feedback from experts during meetings with topic-specific experts in NMA, allergist, haematologist, and members of the GRADE Working Group.

I conducted all analyses, designed figures and tables, and organized meetings. I wrote the manuscripts with editorial advice and supervision of Professor Schünemann, and with feedback from Professors Guyatt, Brozek, and Beyene. The authors on each paper contributed significantly with important comments and advice for the final manuscripts.

For all the four manuscripts composing this “sandwich” thesis, earlier drafts of parts of this research have been presented at international academic conferences as part of the manuscripts’ development. The first article (Chapter 2) will be submitted to the Journal of Clinical Epidemiology. The second article was submitted to Journal of Clinical Epidemiology in April 2018 and is under review. The third article was published in the Allergy journal in 2017. The fourth article will be submitted to Blood Advances.

CHAPTER 1. INTRODUCTION

Chapter 1: Introduction of the thesis

Evidence synthesis: From Meta-analysis to Network-Meta-analysis

Most systematic reviews (SR) address the merits of one intervention compared to another (e.g., placebo, or another active intervention) and investigators often combine data across studies that meet eligibility criteria in what we will term a pairwise meta-analysis (1). Compared with findings derived from a single study, SRs that employ a meta-analysis (MA) improve the power to detect differences and also facilitate examination of the extent to which there are important differences in treatment effects across eligible studies - variability that is frequently called heterogeneity (2, 3). Large, unexplained heterogeneity may reduce confidence in estimates of treatment effects (1).

Recently, another form of MA, network meta-analysis (NMA) (also known as multiple treatment comparison meta-analysis), has emerged (4, 5). The NMA approach provides estimates of effect sizes for all possible pairwise comparisons whether or not they have been compared head-to-head in randomized controlled trials (RCTs) or non-randomized studies (NRS) (1).

Summary presentation of findings of Network-Meta-analysis

Summaries of evidence for health-care professionals exist in different formats and for different purposes. Structured abstracts, sometimes the only part of a

study or review that readers view or use, were initially developed to assist readers in retrieving, selecting, and critically appraising relevant literature (6). In a systematic review of interventions, Summary of Findings (SoF) tables present the main findings of a review in a transparent and simple form. They provide key information concerning the certainty of the evidence, the magnitude of treatment effects, and the sum of the available data on all patient-important outcomes.

Similar to pairwise MA, in the context of NMA, appropriate presentation of findings for each pair-wise comparison, and for each patient-important outcome, is crucial for successful understanding and implementation of evidence-based treatments (7). Decisions should be influenced not only by the precision of the effect estimates but also by other factors such as the certainty in the estimates (certainty or quality of the evidence), and the rank ordering of interventions (1).

Authors have suggested that there are fundamental aspects that should be reported in the results of NMA such as: i) a description of the network and its geometry (i.e., how the compared interventions relate to each other), ii) the effect size estimates for pairwise comparisons between interventions, and iii) the rank ordering of interventions (or ranking probabilities) (8-14). Authors have reported on methods for assessing the certainty of NMA estimates and the rank ordering of interventions for each outcome (15).

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force has described an approach to decide, from the perspective of

decision-makers, the relevance and the credibility of NMA findings (11). The ISPOR system identifies five domains linked to the NMA findings' credibility: i) evidence base, ii) analysis, iii) reporting quality and transparency, iv) interpretation and v) conflict of interest. Four subdomains are associated with the interpretation of findings: i) conclusions related to the NMA results, ii) available evidence, iii) credibility of the analysis methods and iv) bias assessment. The ISPOR guideline does not, however, specify how these subdomains could be used concurrently.

Other authors have provided tutorial articles that focus on educating clinicians and methodologists on the fundamentals of NMA (16, 17) but the focus of these articles has not been on the presentation of findings. Thus, there is a gap in exploring the optimal presentation of NMA findings using SoF tables.

Methodological and presentational issues in systematic reviews and meta-analysis of RCTs and NRS

Authors of SR may want to include NRS in a review because they identified only a few numbers of RCTs, or because there are limitations in the RCTs identified. Although RCTs are the preferred study design to analyze the effects of healthcare interventions, sometimes a research question cannot be answered by RCTs and the inclusion of NRS in SR is justifiable (18). In addition to issues in NRS related to imbalance in prognostic factors (or confounding factors) between intervention and control groups, other methodological issues arise when NRS are

included in SR. For instance, NRS are more susceptible to bias than RCTs in inclusion of participants (selection bias) and less frequently include prior specified protocols (reporting bias) (18). Similar to RCTs, an evaluation of risk of bias (RoB) for individual NRS is needed when these studies are included in a SR.

The risk of bias in non-randomized studies of Interventions (ROBINS-I) tool is a new Cochrane tool to assess the RoB in NRS. The ROBINS-I tool evaluates risk of bias in estimates of the effects (harm or benefit) of one or more interventions from studies that did not use randomization to allocate units (individuals or clusters of individuals) to compare groups (19). The tool evaluates the RoB in several domains such as: i) bias due to confounding, ii) bias in selection of participants into the study, iii) bias in classification of interventions, iv) bias due to departures from intended interventions, v) bias due to missing data, vi) bias in measurement of outcomes, and vii) bias in selection of reported results. Moreover, there is a guidance on issues in presentation of RCTs and NRS findings when these findings are reported in the context of SR (20). However, no description of the applicability of different tools to assess the RoB in individual studies, including ROBINS-tool, has been conducted in epidemiological studies.

Goals and scope

This dissertation focuses on describing solutions to high-priority methodological challenges in evidence synthesis. The challenges I addressed include conducting

and reporting NMA findings SR that include both RCTs and NRS. Specifically, the dissertation describes:

1) I evaluated 276 NMA reports published in 2014 and 2015 and I describe issues in conducting and reporting findings in these studies.
2) To develop an NMA SoF table, I conducted a qualitative study to evaluate the user's experience when NMA findings were displayed in each of the formats.
3) As a result of the qualitative study, I produced two final NMA SoF tables that can be used in NMA publications.
4) I conducted two SR that required confronting issues in conducting and reporting evidence synthesis in SR that included evidence from both RCTs and NRS.

Thesis overview

This thesis dissertation is composed of four articles described in chapter 2 to chapter 5. Chapter 2 described the conduct of a systematic survey that presents methodological issues of 276 NMA publications regarding their conduct and reporting of findings. This analysis elucidates the limitations of current NMAs and provides highlights to recognize the main components to include in a summary of finding table for NMA findings. Chapter 3 reports the development of two final NMA SoF tables through a qualitative study. The final NMA-SoF tables were developed in an iterative process that included feedback from different NMA

users including methodologists, clinical epidemiologists and biostatisticians, and clinicians. Chapter 4 and chapter 5 present two SR that address methodological issues that arose when including both RCTs and NRS. Chapter 4 described the evidence regarding the association between Vitamin D and allergic diseases development; and Chapter 5 depicted the impact of pharmacological prophylaxis for venous thromboembolism in patients undergoing neurosurgical intervention. Issues included assessing risk of bias in individual studies, and reporting findings of both RCTs and NRS.

Thus, this dissertation addresses issues in conducting NMAs and reports the development of two final NMA SoF tables. I address issues in conducting SR of RCT and NRS, specifically in areas of assessing the risk of bias of individual NRS, and how best to report estimates of RCTs and NRS.

REFERENCES

1. Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. Jama. 2012;308(12):1246-53.

2. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. 1998;351(9096):123-7.

3. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. The New England journal of medicine. 1987;316(8):450-5.

4. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. Statistical methods in medical research. 2008;17(3):279-301.

5. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in medicine. 2004;23(20):3105-24.

6. Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. Journal of clinical epidemiology. 2010;63(6):620-6.

7. Pardo-Hernandez H, Alonso-Coello P. Patient-important outcomes in decision-making: a point of no return. Journal of clinical epidemiology. 2017.

8. Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2009;181(8):488-93.

9. Cipriani A, Higgins JP, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Annals of internal medicine. 2013;159(2):130-7.

10. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. Medical decision making : an international journal of the Society for Medical Decision Making. 2013;33(5):679-91.

11. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, 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 in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):417-28.

12. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):429-37.

13. Tan SH, Bujkiewicz S, Sutton A, Dequen P, Cooper N. Presentational approaches used in the UK for reporting evidence synthesis using indirect and mixed treatment comparisons. Journal of health services research & policy. 2013;18(4):224-32.

14. Sullivan SM, Coyle D, Wells G. What guidance are researchers given on how to present network meta-analyses to end-users such as policymakers and clinicians? A systematic review. PloS one. 2014;9(12):e113277.

15. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. PloS one. 2014;9(7):e99682.

16. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. Journal of clinical epidemiology. 2014;67(6):672-80.

17. Foote CJ, Chaudhry H, Bhandari M, Thabane L, Furukawa TA, Petrisor B, et al. Network Meta-analysis: Users' Guide for Surgeons: Part I - Credibility. Clinical orthopaedics and related research. 2015;473(7):2166-71.

18. Reeves B, Deeks J, Higgins J, Wells G. Chapter 13: Including non-randomized studies. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]: The Cochrane Collaboration; 2011.

19. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. Journal of clinical epidemiology. 2018.

20. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013;66(2):158-72.

CHAPTER 2. EVALUATION OF 276 NETWORK META-ANALYSIS PUBLICATIONS: STATISTICAL APPROACHES, PRESENTATION OF FINDINGS, CERTAINTY OF EVIDENCE ASSESSMENTS AND INTERPRETATION OF FINDINGS ARE ONLY PARTIALLY TACKLED

PREFACE TO CHAPTER 2

Chapter 2: *Evaluation of 276 network meta-analysis publications: statistical approaches, presentation of findings, certainty of evidence assessments and interpretation of findings only partially tackled* will be submitted to the Journal of Clinical Epidemiology.

Chapter 2: Evaluation of 276 network meta-analysis publications: statistical approaches, presentation of findings, certainty of evidence assessments and interpretation of findings only partially tackled

Author list

Juan José Yepes-Nuñez[1,2], Yuan Zhang[1], Behnam Sadeghirad[1], Housne Ara[1], Matthew Ventresca[1], Ivan Florez[1,2], Laura Banfield[3], Loukia Spineli[4], Nigar Sekercioglu[1], Cristiane De Cassia Bergamaschi[5], Luciane Lopes[6], Carolina Martins[7], Reed Siemieniuk[1], Yutong Fei[8], Paul Elias Alexander[1], Kelly Patricia Estrada Orozco[9], Andrea García[9], Assem Khamis[10], Joseph Beyene[1], Jan Brozek[1,11], Gordon Guyatt[1,11], Holger J. Schünemann[1,11]

Affiliations

1. Department of Health Research Methods, Evidence, and Impact (HEI), McMaster University, Hamilton, Ontario, Canada
2. Faculty of Medicine, Universidad de Los Andes, Bogotá, Colombia
3. Health Sciences Library, McMaster University, Hamilton, Ontario, Canada

4. Institut für Biometrie, Medizinische Hochschule Hannover, Hannover, Germany
5. Department of Pharmaceutical Sciences, University of Sorocaba, Sorocaba, Brazil.
6. University of Sorocaba, Sorocaba, São Paulo, Brazil.
7. Department of Paediatric Dentistry and Orthodontics, School of Dentistry, Federal University of Minas Gerais, Belo Horizonte/MG, Brazil.
8. Beijing University of Chinese Medicine, Beijing, China
9. Clinical Research Institute, National University of Colombia, Bogotá, Colombia
10. American University of Beirut Medical Center, Beirut, Lebanon
11. Department of Medicine, McMaster University, Hamilton, Ontario, Canada

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact (HEI), Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON. L8S 4K1 Canada. schuneh@mcmaster.ca

Conflict of interest: The authors declare they have no competing financial interests with respect to this manuscript, or its content, or subject matter.

ABSTRACT

Objective: To describe how network-meta-analysis (NMA) publications conduct statistical analyses, present their findings, assess the certainty of evidence, and implement strategies to interpret their findings. **Study design and setting:** We searched 14 databases to identify NMAs published in 2014 and 2015. We included NMAs of randomized controlled trials that compared three or more interventions. We abstracted data in duplicate, focusing on NMA structural characteristics and analysis, presentation of results, assessment of certainty of evidence, and interpretation of findings. **Results:** We included 276 published NMAs of which 240 (87%) conducted a Bayesian analysis, 164 (59.4%) explored heterogeneity, 26 (9.4%) explicitly addressed the transitivity assumption, 69 (23.5%%) addressed differences locally between direct and indirect estimates (incoherence or inconsistency) and 111 (40.7%) globally. For the primary outcome, NMAs included a median (interquartile range) of 4880 (2186 - 21923) patients, 6 (4-8) active treatments, 17 (10 - 32) trials, and 8 (5 - 12) direct comparisons. Of the 276 NMAs, 225 (81.5%) presented the network geometry, 213 (77.2%) assessed the risk of bias, 170 (61.6%) reported both direct and NMA estimates, 156 (56.5%) presented results relating to rank hierarchy, and 15 (5.4%) presented certainty (quality) of evidence ratings. Authors almost always (n=268, 96.7%) used NMA estimates as the primary source of evidence to interpret NMA, sometimes together with rank probabilities (n=81, 29.3%). **Conclusion:** Most recent NMAs fail to explicitly evaluate transitivity, fail to

address differences between direct and indirect estimates in individual paired comparisons, and fail to assess the certainty of the evidence. Future NMAs should deal with these important issues.

Keywords: Network Meta-Analysis, Certainty of Evidence, Systematic Reviews, GRADE.

Funding: This research was supported by two Colombian governmental doctoral scholarships provided by Colciencias and EnlazaMundos, and by the GRADE Centre at McMaster University, Canada.

HIGHLIGHTS

Key finding

- NMAs typically use Bayesian analysis, present the network geometry, and address individual study risk of bias.
- Many NMAs (in the vicinity of 50%) present results of both direct and network estimates, calculate rank probabilities and present results mostly about local incoherence (inconsistency).
- Few NMAs explicitly explored transitivity, addressed differences between direct and indirect estimates in individual paired comparisons or assessed the certainty of the evidence.

What this adds to what is known

- This study provides a cross-sectional analysis of the evolving practice of NMAs. Although assessing the certainty of evidence has become standard for conventional meta-analysis, it is almost always neglected in NMAs

What are the implications? What should change now?

- NMA authors should strive to address important aspects of their findings, including presentations of direct, indirect and network estimates, assessment of local incoherence (inconsistency), and in addition, provide necessary information regarding certainty of evidence.

INTRODUCTION

Evidence synthesis of randomized controlled trials (RCTs), such as systematic reviews and meta-analysis, support healthcare decisions on the comparative effectiveness of two interventions. Multiple treatment options are, however, often available for the same condition in clinical practice and, in such situations, conventional meta-analysis provides limited insight (1). Network meta-analysis (NMA) provides a simultaneous comparison of multiple interventions, and thus addresses a key limitation of conventional meta-analysis (2).

NMA offers a quantitative method of integrating data from direct and indirect comparisons (3), provides estimates of the relative effect of multiple interventions against one another, and thus insight into the superior and less satisfactory treatment alternatives. NMAs can complement relative effects from paired comparisons with information on intervention hierarchy including “rankograms”, probability to rank first, median or mean ranking and the surface under the cumulative ranking curve (SUCRA) (4).

Many authors have provided guidance regarding identification of the evidence, statistical aspects of conducting NMA, and critical appraisal and interpretation of NMAs (4-8). A number of articles have also highlighted limitations in the conduct of NMAs, including violation of assumptions necessary to assure credible results (9-11). The evolution of NMAs is, however, happening quickly, and these prior efforts have not been comprehensive and may be outdated.

Therefore, in order to characterize the recent practice of NMAs, and to elucidate possible limitations in the methods of evaluation, analysis, and presentation of results, we conducted a systematic survey of NMA addressing the following issues: 1) the presentation of the results of the NMA, including number of outcomes, number of comparisons, network graphs; 2) exploration of heterogeneity and incoherence (inconsistency in results between direct and indirect comparisons; 3) presentation of rank probabilities (statistical method, approach to presentation); 4) presentation of certainty of evidence in direct comparisons and NMAs; and 5) interpretation of findings based on the combination of rank probabilities and certainty of evidence.

METHODS

Eligibility criteria

To be eligible, NMAs had to be published from 2014 to 2015 that compared the effects of three or more interventions based on randomized control trials. We excluded NMAs not available in English, Spanish, Chinese, German, and Italian, conference abstracts report, editorials, letters, opinions, protocols, methodological articles, overviews of NMAs, cost-effectiveness reviews that included NMAs, NMAs of individual patient data, and NMAs not involving human participants, and those addressing dental interventions.

Search strategy

A librarian (LB) from the Health Science Library at McMaster University designed the search strategy. We searched the following clinical databases using Ovid: Allied and Complementary Medicine (AMED), Cochrane Database of Systematic Reviews (Cochrane DSR), Database of Abstract of Reviews of Effects (DARE), Cochrane Methodology Register (CMR), Health Technology Assessment (HTA), NHS Economic Evaluation Database (NHSEED), Embase, Medline, and PsycINFO, Cumulative Index to Nursing and Allied Health Literature (CINAHL), and Latin American and Caribbean System on Health Sciences Information (LILACS), Web of Science; and a mathematical database: MathSciNet. Further, we explored the following grey literature sources: Journal of Negative Results in BioMedicine (http://www.jnrbm.com/), Agency for Healthcare Research and Quality (http://www.ahrq.gov/), World Health Organization (http://www.who.int/en/), the New York Academy of Medicine (http://www.greylit.org/), and the Gray Source Index (via Open Grey or Repository of Grey Literature) (http://www.greynet.org/). We searched all databases from their inception up to December 2015. Appendix A presents details of the search strategies.

Network Meta-analysis selection

Two reviewers (JYN, YZ) independently screened the titles and abstracts identified through the electronic searches. The principal reviewer (JYN) read all

studies, and one of nine additional reviewers (YZ, BS, IF, CCB, LL, PA, NS, YF, RS) addressed the eligibility of the full text independently from each study that passed the title and abstract screening. Discrepancies between the principal reviewer and the paired reviewer were solved for a third reviewer (LS). The forms for the screening of titles and abstracts and the full text were piloted before their use and the reviewers were calibrated for full-text screening. Chance-corrected agreement for full-text eligibility assessment was calculated using the Kappa statistic.

Data extraction and synthesis

We abstracted pertinent NMA details in duplicate from all studies. Similar to the full text screening, JYN extracted data from all studies, and one of 15 other reviewers (JYN, YZ, BS, HA, MV, IF, NS, CCB, LL, CM, RS, YF, PA, KP, AG, AK) also abstracted data independently from each study. Disagreements were resolved by discussion among the reviewers; in case of discordance, a third reviewer (LS) adjudicated the issue.

From each eligible NMA, we selected one primary outcome. When the authors described more than one primary outcome, we selected the first outcome reported in the NMA methods section. We extracted the following information from each NMA included in our systematic survey: 1) patient, interventions, primary outcome description; 2) characteristics of statistical analysis; 3) risk of bias assessment; and 4) certainty in the evidence assessment. For the selected

primary outcome, we obtained the following information: 1) characteristics of the network geometry; 2) findings; and 3) interpretation of findings. We carried out descriptive analysis for all the characteristics extracted. After completing the data synthesis for all eligible NMAs, we prepared multiple tables summarizing the main findings in narrative and tabular form. We used Microsoft Excel 2011 for Mac for data management and for descriptive statistics.

Given the large number of references retrieved from our search strategy, we decided to limit our data extraction to NMA reports published in 2014 and 2015.

RESULTS

Our search strategy identified 5,699 potential relevant citations of which 1,045 citations were potentially eligible after title and abstract screening, and 276 proved eligible (Figure 1). Appendix B provides a complete list of included and excluded studies. Kappa for eligibility assessment of full-text articles retrieved was moderate (0.50; 95% confidence interval: 0.44-0.55). Disagreements between reviewers occurred due to heterogeneity in statistical NMA terms. We created a guide with NMA terms and definitions of NMA statistical procedures to harmonize the full-text screening and data abstraction for the reviewers.

Description of included systematic reviews

Table 1 presents characteristic of all NMAs published from 2014 and 2015. Figure 2 shows the clinical areas involved in the NMAs.

NMA authors used multiple terms to denote NMA. Two hundred thirty-eight (86.2%) NMAs used the term "network meta-analysis", 17 (6.2%) termed NMAs as "mixed treatment comparisons", 13 (4.7%) as "multiple treatment comparisons", 4 (1.4%) as "indirect comparison meta-analysis", 2 (0.7%) as "multiple treatment meta-analysis", and 2 (0.7%) as "adjusted indirect comparison".

Statistical analysis in selected NMAs

Statistical methods for data synthesis framework

Of the 276 NMAs, 240 (87%) used a Bayesian framework; 11 (4%) a frequentist approach, and 25 (9%) failed to specify their framework. Of the NMAs using Bayesian approaches, 142 (51.4%) used hierarchical models, 44 (15.9%) included multivariate meta-analysis (n=44,15.9%), 23 (8.3%) adjusted indirect estimates, 20 (7.2%) multivariate meta-regression, and 1 (0.4%) a graphical-theoretical method. The reminder 46 (16.7%) NMAs did not report any information of the approach for data synthesis.

Two-hundred-seventeen (78.6%) NMAs produced a direct estimate for every paired comparison. One-hundred sixteen (53.5%) NMAs applied a frequentist framework to calculate direct effect estimates, and 41 (18.9%) applied a Bayesian framework. Sixty (27.6%) NMAs did not report the statistical framework applied to determine direct effect estimates.

Statistical methods for heterogeneity and transitivity evaluations

Evaluation of transitivity or similarity

Of the 25 (9.1%) of NMAs that assessed the transitivity assumption, 14 (56%) explored patient characteristics, 6 (24%) patient characteristics and study design, and 5 (20%) intervention characteristics.

Heterogeneity evaluation for direct comparisons

For the 219 direct comparisons explored in the networks, 158 (72.1%) determined the presence of heterogeneity through I-square, and 8 (3.7%) estimated the between-study variance (tau-square). The other 53 (24.2%) NMA publications did not report any statistical method to assess heterogeneity in the direct comparisons. Twenty-five used more than one test aside from the I square or tau-square, 12 (48%) also applied the Cochrane's Q test, and 13 (52%) Chi-square.

Heterogeneity evaluation for NMA estimates

Sixty-nine (25%) NMAs addressed incoherence for the entire network (between direct and indirect effect estimates). NMAs reported to evaluate heterogeneity in the entire network with different methods. Fifty (72.5%) reported I-square, 14 (20.3%) tau-square, 3 (4.3%) Cochrane's Q test, and 2 (2.9%) Chi-square test. Two-hundred-seven (75%) NMAs did not conduct any statistical method to

assess heterogeneity in the entire network or did not describe the statistical method implemented to assess heterogeneity (n=114, 41.3%; and n=93, 33.7%, respectively).

Overall, fifty-five (19.9%) NMAs assessed heterogeneity in both direct comparisons and the entire network.

Statistical methods for incoherence (or inconsistency) evaluation

One-hundred-twelve (40.7%) NMAs tested for global incoherence. NMAs applied six statistical methods to assess global incoherence. Of the 112 NMAs, 87 (31.5%) applied thw comparison model to fit and parsimony, 12 (4.3%) design treatment interaction, 11 (4%) Lu-Ades model, 1 (0.4%) net-heat plot, and 1 (0.4%) Lumley model. Four (1.4%) NMAs implemented more than one statistical method to evaluate global incoherence. One NMA that primarily applied a design treatment interaction method also applied a Lumley model. Two out of 3 NMAs that primarily implemented a comparison model fit and parsimony method, also applied a Lu-Aedes model, and the other NMA applied a net-heat plot method.

NMAs that conducted *local* incoherence evaluation applied three different statistical methods. Fifty-three (19.2%) NMAs applied loop-specific approach or Bucher method; 43 (15.6%) node-splitting; and 4 (1.4%) back-calculation. One-hundred-seventy-six (63.7%) NMAs did not address *local* incoherence.

Overall, forty-eight (17.4%) NMAs implemented an evaluation of both local and global incoherence (inconsistency).

Subgroup analysis and meta-regression in selected NMA

Of the 166 NMAs that reported a heterogeneity evaluation for the direct comparisons, 13 (7.8%) conducted a subgroup analysis. Twelve NMAs prespecified the following moderators to explore a subgroup analysis: 5 (38.5%) patient-characteristics, 4 (30.8%) outcome-characteristics, 2 (15.4%) intervention-characteristics, and 1 (7.7%) trial design. One NMA did not prespecify the moderators to explore a subgroup analysis. Authors reported in 6 NMAs (n=46.2%) that the subgroup analysis explained the heterogeneity in their estimates by applying an interaction test.

In the 69 NMAs that explored heterogeneity only for the entire network, 8 (11.6%) reported a meta-regression analysis within the network. An effect modifier explained the heterogeneity in the entire network by meta-regression in 1 NMA.

Sensitivity analysis in selected NMA

Of the 153 (55.4%) of NMAs that conducted a sensitivity analysis the following characteristics were primary explored in the NMAs: 40 (26.1%) explored trial design, 39 (25.5%) intervention characteristics, 28 (18.3%) risk of bias assessment, 23 (15%) patient-characteristics, 21 (13.7%) fixed and random

effect models, 1 (0.7%) frequentist vs Bayesian approaches; and 1 (0.7%) both intervention characteristics and risk of bias assessment.

Sixty-nine (25%) NMAs with sensitivity analysis assessed between fixed vs. random effects models. Forty-eight (69.6%) NMAs reported similar NMA effect estimates when fixed or random effect models were carried out, 19 (27.5%) NMAs reported different estimates with both methods, and 2 (2.9%) NMAs did not report results of the exploration.

Seven (2.5%) NMAs explored whether NMA effect estimates varied when a frequentist or Bayesian framework was applied. Six NMAs reported similar NMA effect estimates when NMA authors applied both approaches.

Reporting NMA findings in selected NMA primary outcomes

NMA geometry in selected NMA primary outcomes

We retrieved information of NMA geometry for primary outcomes. Two-hundred-twenty-five (81.5%) NMAs reported an NMA geometry. Each network included a median of 6 (IQR 4-8) active treatments. The median number of included trials per network was 17 (IQR 10-32), and the median total sample was 4880 (IQR 2186-21923) participants.

One-hundred-eleven NMAs (40.2%) informed treatment nodes with different sizes. The node size represented the number of participants in most of the NMA

geometries (n=100, 36.2%), and only a few of them represented the number of trials (n=11, 4%).

Links between nodes described a different type of comparisons. Most links represented direct comparisons (n=196, 71%), 24 (8.7%) represented direct and indirect comparisons, and 5 (1.8%) indirect comparisons. The mean number of links for direct comparisons was 8 (IQR 5-12). One hundred forty-six (52.9%) NMAs reported their links with different widths to represent the network geometry. Most of the widths represented the number of trials (n=138, 50%), 7 (2.5%) represented the number of participants, and 1 (0.4%) the inverse of the variance of the direct estimate.

Eight (2.9%) indicated multi-arm trials in the NMA geometry. Of those six NMAs used a different line style and two NMAs used text to describe the presence of multi-arms trials in the geometry.

Presentation of NMA findings

Formats to present summary effect estimates

Both NMA effect estimates and direct effect estimates were reported in 170 (61.6%) NMAs in multiple formats, and 106 (38.4%) NMA publications reported only NMA estimates.

In those publications that reported NMA effect estimates and direct estimates together, the NMA effect estimates were presented as follows: 75 (44.1%) NMAs used regular tables, 42 (24.7%) NMAs reported findings in two by two tables (also named matrix table or league table), 32 (18.8%) NMAs applied forest plot that included all comparisons, 15 (8.8%) NMAs reported forest plot including only estimates of basic parameters, 4 (2.4%) NMAs embedded the NMA effect estimates in the NMA geometry, and 2 (1.2%) NMAs only reported findings by text. Direct effect estimates were reported as follows: 87 (51.2%) NMAs used forest plots, 75 (44.1%) NMAs used regular tables, 4 (2.4%) NMAs reported direct findings in league tables, and 4 (2.4%) NMAs by only text.

Of the 106 NMAs that presented only NMA estimates, 34 (32.1%) NMAs used regular tables, 30 (28.3%) NMAs reported forest plot with all comparisons, 28 (26.4%) NMAs employed league tables, 10 (9.4%) NMAs used forest plot with basic parameters, 3 (2.8%) embedded the direct effect estimates in the NMA geometry, and 1 (0.9%) NMA reported findings by text.

Summary effect estimates

NMAs reported effect estimates either as relative effects (e.g. odds ratios, hazard ratios, response rates, withdrawal rates) or absolute effects (e.g. mean differences) with either 95% confidence intervals (for frequentist methods) or 95% credible intervals (for Bayesian methods). Table 2 shows all summary effect estimates and measures of uncertainty for the NMA effect estimates.

Rank probabilities

One-hundred fifty-six (56.5%) NMAs reported rank probabilities. NMA authors applied three different statistical methods to calculate rank probabilities. One-hundred forty-four (52.2%) NMAs used Bayesian approach, 9 (3.3%) NMAs conducted frequentist approach, and 3 (1.1%) NMAs used Bootstrap approach. NMAs presented rank probabilities findings in three formats. Fifty-nine (21.4%) reported rank probabilities graphically, 56 (20.3%) NMAs presented rakings numerically (n=56, 20.3%), and a combination of numerically and graphically formats in the same NMA report was used in 41 (14.9%) NMAs. Table 3 shows each of the methods applied to report the NMA rank probabilities according to each format. None of NMAs reported uncertainty for the rank measure.

Assessing the certainty (or quality) of the evidence in NMA

Risk of bias assessment or quality assessment of individual RCT

Of the 213 (77.2%) of NMAs that evaluated risk of bias of individual RCT included in the NMAs, 146 (53%) applied the Cochrane risk of bias tool, and 67 (24%) applied other tools. Sixty-four (22.8%%) NMAs did not assess risk of bias individual RCTs. Table 4 describes other tools implemented in the evaluation of risk of bias of individual RCT.

Certainty (or quality) of the body of evidence assessment for NMAs

Of the 15 (5.4%) of NMAs that assessed the certainty of the body of evidence, 4 (1.4%) applied the official GRADE approach to NMA (12), 3 (1.1%) applied an NMA framework developed by Salanti et al. (13), 4 (1.4%) applied the GRADE framework for interventions (14), and 4 (1.4%) modified the GRADE framework for interventions to their NMAs.

NMA authors of these 15 NMAs assessed the certainty of evidence to the direct comparison, indirect or NMA effect estimates applying different approaches. NMAs that applied the official GRADE framework (12) assessed the certainty of evidence for the direct, indirect, and NMA estimates. NMAs that applied Salanti's framework (13) assessed the certainty of evidence to different resources of evidence. One NMA applied the framework to assess all three sources of evidence: direct, indirect, and NMA estimates. One NMA applied the framework to assess the certainty of evidence to indirect and NMA effect estimates, and another NMA assessed the certainty of evidence to only the NMA effect estimates. Three NMAs that applied the GRADE framework for interventions (14) assessed the certainty of evidence for the direct estimates, and one NMA only reported the assessment of the risk of bias for individual RCT. NMA that adapted used an adapted GRADE framework (14) applied the framework to assess the direct, indirect and NMA estimates, and one NMA used the framework to assess the NMA estimates.

Interpretation of Findings in selected primary outcomes

Per outcome, authors of NMAs used different components to interpret NMA findings. Two-hundred-sixty-eight (96.7%) NMAs used NMA effect estimates to interpret findings, 6 (2.2%) NMAs interpreted their findings based on the ranking probabilities, 2 (0.7%) NMAs used the GRADE assessment to interpret findings, and 1 (0.4%) NMAs used the findings of the direct comparison effect estimates to interpret results. Both NMAs that interpreted their findings based on the GRADE assessment did not apply the NMA GRADE framework to assess the certainty of the evidence. Instead, authors of these two NMAs used the GRADE framework for interventions to assess the certainty of the evidence. One hundred twenty-seven (46%) NMAs applied, in addition to the resources described before, other resources to interpret findings (table 5).

Of the 267 of NMAs that interpreted their findings using the NMA effect estimates, 81 (29.3%) included rank probabilities information in their conclusions. Authors of 27 (33.3%) NMAs described their findings using both resources of evidence and explained these findings to a clinical context. We did not find any NMA that interpreted the findings while accounting for NMA estimates, rank probabilities and certainty of evidence.

DISCUSSION

Main findings

This systematic survey provides comprehensive information about the methodological approaches and reporting characteristics of 276 NMAs published in 2014 and 2015. We also explored the frequency of evaluating different NMA's assumptions in our systematic survey.

Strength and limitations

Strengths of our study include explicit eligibility criteria, a comprehensive literature search, duplicate assessment of eligibility with moderate agreement, and a detailed iterative process for the data extraction from the NMA reports.

Our review has some limitations. The identification of studies addressing how NMA authors conducted NMAs is complex and time-consuming due to different terminology. Thus, although we identified 276 eligible NMAs between 2014 and 2015, it is possible that there are other eligible NMAs that we failed to identify. We calibrated the process of screening and extracting data with our reviewers. We expect this calibration process has decreased the chance of lacking information. We got a fair kappa agreement in our full text screening process despite of standardized this process twice. This main reason for disagreement was the nomenclature variability to NMAs. This issue has been explored in another survey on systematic reviews with NMAs (11). To ensure that relevant

studies were included, the information was double checked by the principal investigator.

In this manuscript, we have identified how NMAs were conducted from methodological and reporting point of view. The next steps in doing so would include a detailed analysis of how to report NMA findings in NMA summary of finding tables (SoF) and how to interpret NMA findings using information from NMA estimates, certainty of evidence and rank probabilities.

Key findings

The transitivity assumption holds when important pre-specified effect modifiers are similarly distributed across RCTs. However, we found that few studies addressed this assumption. NMA authors should provide explicit information on patient and study characteristics that allow readers to evaluate this assumption empirically.

Inconsistency in a treatment network can indicate lack of transitivity (6). We found a high proportion of NMAs that did not conduct an inconsistency evaluation. An imbalance of effect modifiers between direct and indirect evidence might cause inconsistency (17).

In the absence of multi-arm RCTs, the Bucher method represents a straightforward approach to assess inconsistency in an NMA (17). Our findings confirm that this approach was the most frequent method to assess local

inconsistency. However, in the context of multi-arm studies, inconsistency evaluation can be complex and challenging. Design-by-treatment interaction models represent one approach to identifying inconsistencies, or conflicts, in NMA that includes multi-arm RCTs (18). Design-by-treatment interaction models allow to assess inconsistency globally, and they address issues introduced by multi-arm RCTs in NMAs (18). Our findings showed a small proportion of NMAs applied this statistical method. Although the method is suitable for a number of statistical software packages, its relatively recent introduction could explain the low prevalence of its applicability in complex NMAs.

Different strategies can be applied to explain or assess inconsistency. These strategies include 1) removing RCTs from the evidence network (adjustment method); 2) splitting nodes in the network; 3) explaining inconsistency using study-level or individual-level covariates (subgroup analysis or meta-regression), and 4) exploring relevant inferences to explain the presence of inconsistencies (18-20). In our study, no high proportion of NMAs included in our dataset explored reasons for inconsistency.

Hierarchy of treatments in the context of NMA can be reached by applying two strategies. One approach involves treatment effect estimates and its measure of uncertainty. The second approach includes multiple statistical methods to rank all competing interventions (21). Our results suggest that more than half of NMAs determined the ranking of their interventions. Rank probabilities were mainly

calculated using Bayesian approach, and few studies reported its estimate applying a frequentist approach. A substantial degree of imprecision in 95% credible intervals has been reported recently (22); however, it is unclear if uncertainty intervals for SUCRA measures and P-scores should be reported in NMAs (21). Veroniki et al. suggested to report the magnitude of uncertainty associated with SUCRA value in the context of decision making (21). For example, it is important for healthcare providers to look at the rank probabilities, or SUCRA, together with the certainty of the evidence, and the effect size, rather than the rankings or SUCRA alone. This is because the rank probabilities information could be misleading. Clinicians usually want to know the preferential order of treatments for a typical patient. Although, ranking measures and probabilities are convenient and straightforward to understand, looking at the rankings alone does not provide accurate information. Clinicians and other decision makers should always be interested in the effect sizes and the certainty of evidence because a high rank does not necessarily imply a large or patient-important effect size (15).

The GRADE Working Group' s has laid out the complexity of rating the certainty of evidence in NMA (12)(16). GRADE proposes the following four steps to assess the certainty of evidence in the effect estimates from an NMA: 1) present direct and indirect treatment estimates for each comparison of the evidence network; 2) rate the quality of each direct and indirect effect estimate; 3) present the NMA estimate for each comparison of the evidence network; and 4) rate the quality of

each NMA effect estimate (12, 16). We found a limited number of NMAs that assessed the certainty of evidence for the direct, indirect and NMA effect estimates appropriately. Although this low frequency may be explained by the short period between the guide publication and our literature search, efforts in disseminating this approach should help in incorporating this step in as part of any NMA publication.

Health-care professionals use NMA findings. Appropriate and correct interpretation of NMA findings is critical to understand and apply their results in a health-care decision-making process. Overall, our findings showed a considerable limitation in applying statistical methods to assess heterogeneity and consistency across studies as well as a lack in assessing the certainty of evidence in the body of evidence. Mills et al. suggest interpreting NMA findings in three main areas: 1) risk of bias appraisal, 2) validity of the NMA results, and 3) applicability of findings to patient care (23). Applying this set of rules for NMAs could help health-care professionals improve their understanding and applicability of NMA findings in real-life challenges.

Our results in the context of previous research

Existing studies were conducted for purposes comparable to ours. For instance, Salanti et al. (3) in 2008 described main characteristics of network geometries

reported in 18 NMAs. Salanti's review found that some network geometries were either a star network but more complex geometries are more common. This finding suggested that placebo or no active treatment dominated the published clinical research. Our review found most of the NMAs reported an NMA network geometry (n=226, 82%) and a high proportion of the reviews focused on two pharmacological interventions (n=189, 68%).

Song et al. (24) reported the characteristics of 88 reviews regarding how transitivity and consistency assumptions were addressed. Less than half of the reviews addressed both assumptions adequately. Our results are similar. Although Song et al. reported findings for NMAs published from 2000 to 2007, the NMAs we identified in 2014 and 2015, did not largely address both transitivity and consistency assumptions.

Donegan et al. (9) analyzed 43 NMAs published from 1966 to 2008. Sixty percent and 35 percent of the reviews evaluated heterogeneity and inconsistency, respectively. Twenty-four (56%) NMAs highlighted when the result was an indirect comparison, and 25 (58%) NMAs made a distinction between indirect comparisons and direct comparisons to interpret findings. Similar findings were found in our review for the proportion of NMAs that addressed a heterogeneity and inconsistency evaluation, as well as the proportion of NMAs that reported findings for direct and indirect or NMA estimates. In our review, we found a higher

proportion of reviews that interpret findings using NMA estimates vs. what was reported by Donegan (97% vs. 58%, respectively).

A systematic survey conducted by the Agency for Healthcare Research and Quality (AHRQ) (25) reported findings of 43 NMAs published between January 1, 2006, to July 31, 2011. Methodological aspects included in this systematic survey included statistical methods applied in the NMAs as well as methods for handling of potential bias, and data presentation. Several aspects were similar to our data. Most of the studies used Bayesian methods, the main clinical area of research was cardiovascular diseases, governments or foundations funded most NMAs, and pharmacologic interventions were evaluated in the majority of networks. Heterogeneity assessment was more frequently assessed for the direct comparison than in the network, effect estimates were mainly reported as odds ratios or relative risks, and authors reported the probability of the intervention being best as the primary parameter to report rank probabilities. Our systematic survey identified similar study characteristics but we included a higher number of NMAs in our systematic survey consolidating these previous findings.

In the context of the UK Health Technology Assessment programme, Tan et al. (26) evaluated the quality of 19 health technology assessment reports that included indirect and mixed treatment comparison methods for evidence synthesis. Reports were published from 1997 to 2011. Evaluation of these reports included data presentation, statistical models and finding presentation. Tan et al.

found not a standard presentational style for reporting indirect or mixed treatment comparison and use of graphics tools was limited in the reports. They concluded that standardization of reporting NMAs is needed, including graphic presentation of findings.

One-hundred-twenty-one eligible NMAs published until July 2012 were analyzed by Bafeta et al. (5) for general characteristics and methodological components. Similar to our findings, Bafeta et al. reported a high percentage (83%) of pharmacological interventions as the main resource of compared interventions, a comparable median number of interventions assessed per NMAs (7, IQR 5-9), and the median number of RCTs included per NMAs (22, IQR 15-40). However, we found a higher proportion of NMAs that did not conduct transitivity (91% vs. 66%) and inconsistency assumptions (60% for global test and 64% for local test vs. 44%), and the risk of bias assessment was more frequently reported in our findings than by Bafeta's (77% vs. 58%). Although there are still deficiencies in assessing the transitivity and inconsistency assumptions in NMAs, the evaluation of risk of bias in individual studies has increased considerably over the years.

Veroniki et al. (27) explored the prevalence of local and global inconsistency assessment in 40 NMAs. Loop specific approach was the most frequently local inconsistency method applied to assess inconsistency between direct and indirect evidence. To evaluate inconsistency for the entire network, Veroniki et al. explored the frequency of reporting only the design treatment interaction method.

Fourteen NMAs (35%) did not report an inconsistency evaluation. Our study confirms Veronik's findings as we also found the most frequent method to assess local inconsistency was the loop specific approach, and 31% of the NMAs included in our review did not report any method to assess inconsistency in the entire network. We included in our report other methods to assess inconsistency for the entire network in addition to design treatment interaction method. We found it was the most frequent method implemented in exploring inconsistency for the entire network. Authors of NMAs should always be aware of evaluating local and global inconsistency with adequate statistical methods to detect and estimate inconsistency correctly.

Two-hundred-one NMAs that included at least three interventions and published up to June 2012 were analyzed by Lee (29). Comparable to our findings, Lee found that NMAs covered a wide range of clinical topics. Heart and circulation diseases (n=41, 22%); rheumatologic diseases (n=30,15%), and cancer (n=25,12%) were the most important clinical areas covered by the NMAs. Our findings confirmed this tendency, although we found more NMAs published in the cancer field than in the rheumatology field. We decided to include surgical procedures in a separate category to differentiate non-surgical interventions from surgical interventions. We also found most NMAs included network geometry in the reports. Different inclusion criteria could explain the difference between Lee's systematic survey and our findings.

Nikolakopoulou et al. (30) described statistical methods reported in 186 NMAs that included at least four treatments and that were published until the end of 2012. Many of the findings reported by Nikolakopoulou et al. were similar to our findings such as the median number of studies per network (21, IQR 13–40), the median number of treatments per network (6, IQR 5–9), the median sample size per network (7729, IQR 3043–24987), and the median number of studies per comparison (2, IQR, 1–4). Similarly, the Bayesian hierarchical model was the most frequent method implemented (n=111, 59%) followed by the Bucher method (n=28, 15%), and Meta-regression (n=27, 15%). Inconsistency methods data were slightly different from our findings. We extracted information for local and global test inconsistency evaluation separately while Nikolakopoulou et al. pooled findings of implementation of inconsistency altogether. We found that loop specific approach was the most frequent method implemented in NMAs to assess inconsistency as Nikolakopoulou et al. also found it. However, the design treatment interaction method was not reported by Nikolakopoulou et al. as this model was introduced in 2012. We found this method was the second most popular to assess inconsistency in the entire network after the comparison model fit and parsimony method. Thus, our findings report additional methods to assess inconsistency for the entire network as well as provide a different tendency in statistical methods implemented by NMA authors.

Petropoulou et al. (31) analyzed a total of 426 NMAs published from inception until April 14, 2015. This systematic review included NMAs with at least four

different interventions, and examined specific NMA methodological characteristics, with an emphasis on statistical analysis and reporting. Findings reported by Petropoulou et al. are similar to ours. For instance, a low frequency of NMAs addressed the transitivity assumption (n= 98, 21.9%), and half of the NMAs conducted an appropriate method to evaluate the inconsistency assumption (n=164, 49.5%). Effect estimates most frequently reported were odds ratio (n=177, 39%) and the mean difference (n=89, 20%). Authors reported, "probability of being the best" (n=137, 30%) more frequently than surface under the cumulative ranking (SUCRA) (n= 39, 9%) as we found in our systematic survey (table 3). Although Petropoulou et al. sought for similar characteristics in conducting and reporting NMAs, we explored these characteristics with greater detail and for NMAs that included three or more interventions. For instance, we included additional methods applied to report NMA rank probabilities, as well as we explored tools to assess the risk of bias in individual RCTs, and the frequency of assessing the quality for the entire body of evidence in NMAs. This additional information provides a better understanding of how the NMAs has been published and how it should be improved in the future.

Zarin et al. (11) applied A MeaSurement Tool to Assess systematic Reviews (AMSTAR) tool (32) and the ISPOR checklist for NMAs (7) to the 456 NMAs reported by Petropoulou et al. (31) to assess the quality and the validity of the analytical methods, respectively. One hundred and nine (25%) NMAs were considered with high quality, 251 (57%) NMAs were rated as moderate quality,

and 78 (18%) NMAs (n = 78) were rated as low quality. Based on the ISPOR tool, 13% (n = 57) of the NMAs discussed the heterogeneity assumption (i.e., choosing between network-specific and comparison-specific heterogeneity), and 47% (n= 238) of the NMAs did not report information about an evaluation of the inconsistency assumption. We also found a high proportion of NMAs that did not include or address heterogeneity assessments or the inconsistency assumptions correctly. Although we did not assess the quality of the NMA with a tool, our findings show high deficiencies in the process of conducting and reporting NMAs in the timeframe we evaluated.

Finally, Hutton et al. (10) reviewed the main findings of six reports that evaluated the quality of reporting of NMAs (5, 9, 24-26, 30). Similar to our findings, authors found a limitation in the NMA reports, including information of literature searching, study selection, risk of bias evaluations, applicability of NMA statistical assumptions, details of statistical models used for analyses, reporting of findings, and approaches for summarizing probability. Hutton et al. suggested the systematic application of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) tool to NMAs can address issues in conducting and presenting NMA findings. In this overview, authors did not explore the inclusion of the assessment of the certainty (or quality) of evidence for the direct, indirect, and NMA estimates as it was conducted in our approach.

Implications for practice and research

Findings of our systematic survey can be applied to multiples areas in the NMA setting. First, we provide an account of issues that researchers conducting NMA research should consider in planning their studies. Developing a protocol prior to conduct an NMA will help to include relevant statistical methods for NMA synthesis. Second, researchers conducting NMAs should consider applying a tool to improve the quality of reporting NMA. The PRISMA statement for NMA (6) could be used for this purpose. Third, as in conventional meta-analysis of interventions, clinical practice guidelines, and health technology assessments, NMA authors should consider rating the certainty of evidence using proper tools that include an assessment of the direct, indirect, and NMA effect estimates. This approach would ensure the correct inference from the results of the meta-analysis. Fourth, future NMA of interventions should incorporate intuitive presentations of the key results such as a NMA summary of findings (SoF) table.

Conclusions

This systematic survey of NMAs suggests lessons for reporting and presenting information in intuitive and transparent ways. This could be done through development of SoF tables. The development of an NMA SoF table may be based on how frequent certain approaches are used to represent findings. For instance, the frequency of reporting the NMA geometry, and rank probabilities could suggest what information should be displayed in an NMA SoF table.

However, the format and content of such a SoF table should best based on what users want rather than what researchers report.

ACKNOWLEDGEMENTS

The authors would like to thank Caitlin Daly for her contribution in clarifying statistical concepts of network meta-analysis during the screening process.

REFERENCES

1. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Bmj. 2005;331(7521):897-900.

2. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. Bmj. 2013;346:f2914.

3. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. Statistical methods in medical research. 2008;17(3):279-301.

4. Salanti G, Ades AE, 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.

5. 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.

6. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Annals of internal medicine. 2015;162(11):777-84.

7. Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17(2):157-73.

8. Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA. 2012;308(12):1246-53.

9. Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect comparisons: a review of reporting and methodological quality. PLoS One. 2010;5(11):e11054.

10. Hutton B, Salanti G, Chaimani A, Caldwell DM, Schmid C, Thorlund K, et al. The quality of reporting methods and results in network meta-analyses: an overview of reviews and suggestions for improvement. PLoS One. 2014;9(3):e92508.

11. Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, et al. Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med. 2017;15(1):3.

12. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. Bmj. 2014;349:g5630.

13. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. PLoS One. 2014;9(7):e99682.

14. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of clinical epidemiology. 2011;64(4):401-6.

15. Cipriani A, Higgins JP, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Annals of internal medicine. 2013;159(2):130-7.

16. Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. Journal of clinical epidemiology. 2018;93:36-44.

17. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Med Decis Making. 2013;33(5):641-56.

18. 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.

19. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med. 2009;28(14):1861-81.

20. Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. Journal of clinical epidemiology. 2009;62(8):857-64.

21. 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.

22. 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.

23. Mills EJ, Ioannidis JPA, Thorlund K, Schünemann HJ, Puhan MA, Guyatt G. Network Meta-analysis. In: Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd ed. New York, NY: McGraw-Hill Education; 2015.

24. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating

healthcare interventions: survey of published systematic reviews. Bmj. 2009;338:b1147.

25. Coleman CI, Phung OJ, Cappelleri JC, Baker WL, Kluger J, White CM, et al. Use of Mixed Treatment Comparisons in Systematic Reviews. AHRQ Methods for Effective Health Care. Rockville (MD)2012.

26. Tan SH, Bujkiewicz S, Sutton A, Dequen P, Cooper N. Presentational approaches used in the UK for reporting evidence synthesis using indirect and mixed treatment comparisons. J Health Serv Res Policy. 2013;18(4):224-32.

27. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol. 2013;42(1):332-45.

28. 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.

29. Lee AW. Review of mixed treatment comparisons in published systematic reviews shows marked increase since 2009. Journal of clinical epidemiology. 2014;67(2):138-43.

30. Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. Characteristics of networks of interventions: a description of a database of 186 published networks. PLoS One. 2014;9(1):e86754.

31. Petropoulou M, Nikolakopoulou A, Veroniki AA, Rios P, Vafaei A, Zarin W, 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.

32. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.

FIGURES

Figure 1. PRISMA flow chart of study selection

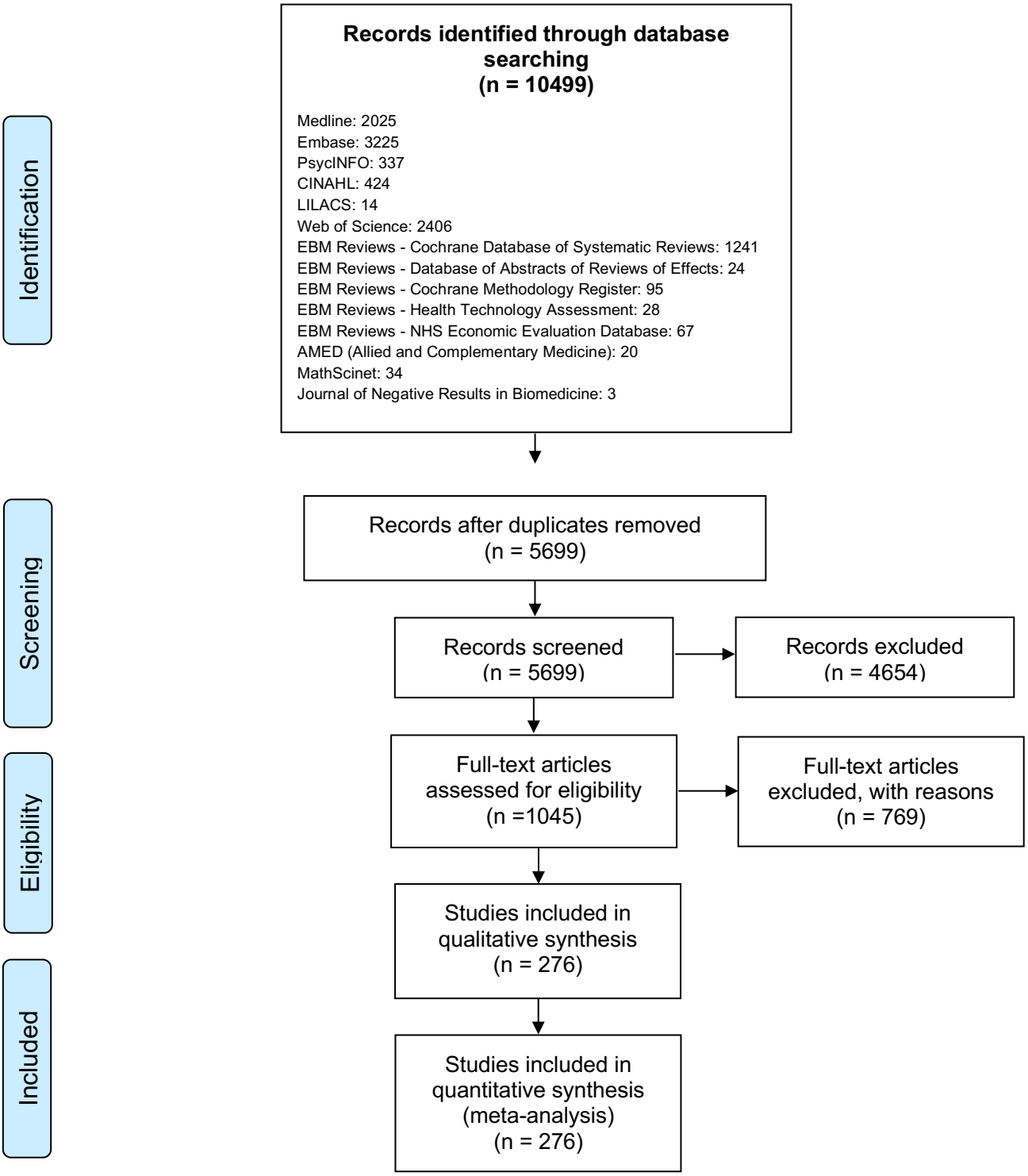
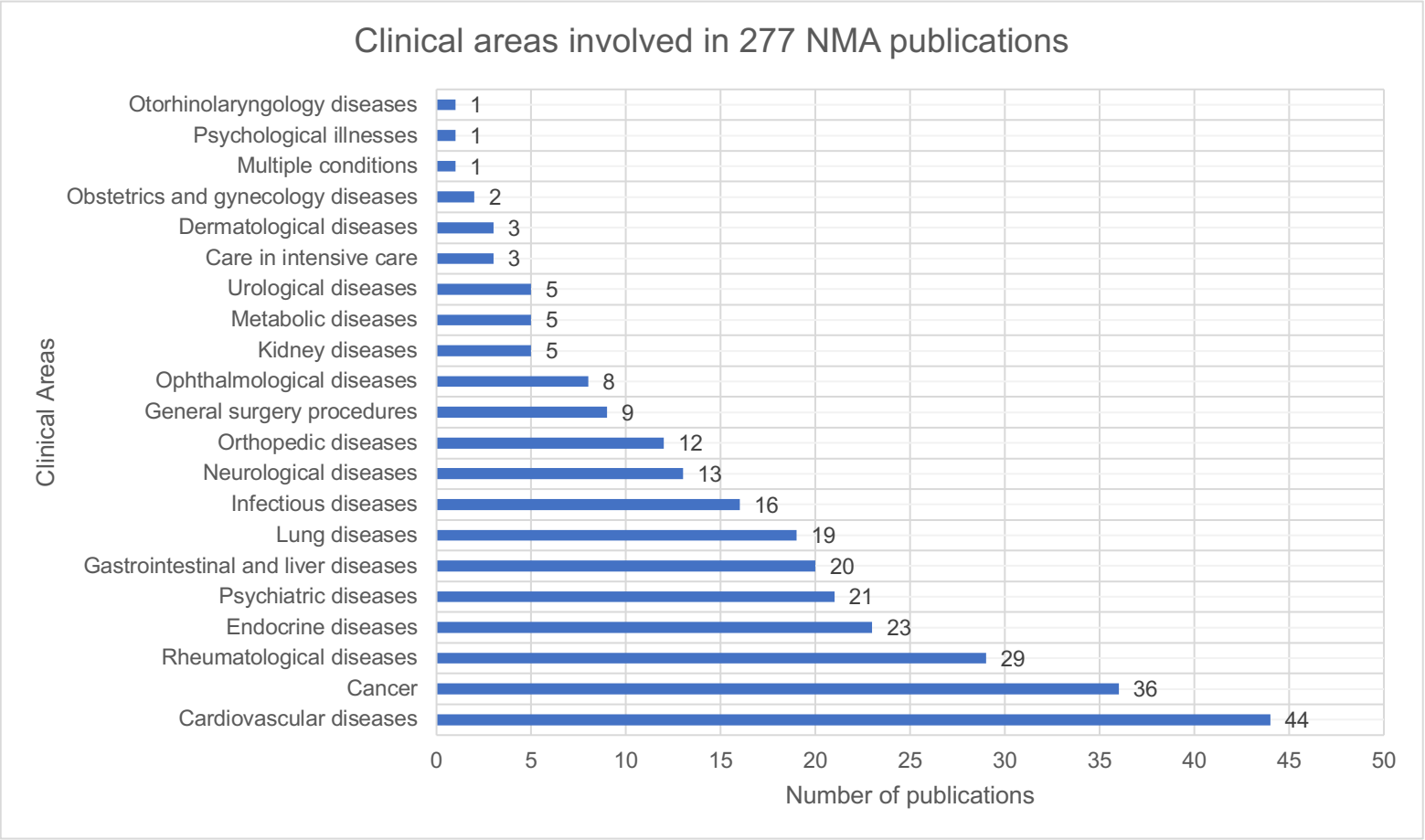


Figure 2. Clinical areas involved in 276 NMA publications



TABLES

Table 1. Characteristic of 276 NMA publications included from 2014 to 2015

Description Characteristics of NMAs	**Median**	**IQR**
Total number of participants	6477	2346-22120
Total number of active interventions	6	4-8
	n	**%**
Number of primary outcomes		
NMAs reported single primary outcome	190	68.8%
NMAs reported more than one primary outcome	86	31.2%
Type of interventions compared		
Pharmacological vs Pharmacological	188	68.1%
Non-pharmacological vs Non-pharmacological	51	18.5%
Pharmacological vs Placebo	27	9.8%
Pharmacological vs. Non-pharmacological	10	3.6%
Funding source		
Academia	10	3.6%
Government/Foundation	86	31.2%
Industry	47	17.0%
No Funding	66	23.9%
No reported	67	24.3%
Protocol reported		
Protocol not mentioned	193	60.9%
Registered	37	13.4%
Mentioned but not available	29	10.5%
Not registered but published	12	4.3%
Explicitly mentioned that there is no protocol	5	1.8%

IQR: interquartile range, NMA: network meta-analysis

Table 2. Frequency of summary effect estimates and measure of uncertainty in 276 NMA publications

Description	n	%
Relative effect measures		
OR w 95% CrI	85	30.8%
HR w 95% CrI	44	15.9%
OR w 95% CI	29	10.5%
RR w 95% CrI	29	10.5%
HR w 95% CI	12	4.3%
RR w 95% CI	8	2.9%
OR w 95% PreI	2	0.7%
Effect size w 95% CI	1	0.4%
RR w 95% PreI	1	0.4%
Rate ratio w 95% CI	1	0.4%
Rate ratio w 95% CrI	1	0.4%
Absolute effect measures		
MD w 95% CrI	34	12.3%
SMD w 95% CrI	8	2.9%
MD w 95% CI	5	1.8%
Median w 95% CrI	3	1.1%
Mean w SD	3	1.1%
SMD w 95% CI	2	0.7%
MD w SD	1	0.4%
Mean change w 95% CI	1	0.4%
Mean w SE	1	0.4%
Posterior median w IQR	1	0.4%
Posterior median w 95% CrI	1	0.4%
No reported	3	1.1%
Total	**276**	**100%**

CI: confidence interval, CrI: credible interval, IQR: interquartile range, MD: mean difference, OR: odds ratio, PreI= predictable intervals, RR: relative risk, SE: standard error, SD: standard deviation.

Table 3. Frequency of methods applied to report NMA rank probabilities in 156 NMA publications

	Description	n	%
Graphics	Rankograms	26	44.1%
	Barplots for ranking probabilities	16	27.1%
	Cumulative ranking probability plots	9	15.3%
	Cluster rank plot	2	3.4%
	Forest plots with 95% CrI	2	3.4%
	Arrow	1	1.7%
	Forest plots with 95% PreI	1	1.7%
	Histogram of rankings	1	1.7%
	SUCRA in a plot	1	1.7%
No graphics	Best-ranked treatment	43	78.2%
	SUCRA	8	14.5%
	Acceptable tolerance threshold for SMD measures	1	1.8%
	Acceptable tolerance threshold OR measures	1	1.8%
	SUCRA with 95% CI	1	1.8%
	SUCRA with 95% CRI	1	1.8%
Both numbers and graphics	Best-ranked treatment	21	25.6%
	Rankograms	14	17.1%
	SUCRA	11	13.4%
	Cumulative ranking probability plots	10	12.2%
	Barplots for ranking probabilities	9	11.0%
	SUCRA w 95% CRI	4	4.9%
	SUCRA (%)	3	3.7%
	Arrow	2	2.4%
	Circle	2	2.4%
	Mean rank w 95% CrI	2	2.4%
	Acceptable tolerance threshold OR measures	1	1.2%

Forest plots with 95% CrI	1	1.2%
Histogram of rankings	1	1.2%
Tolerance threshold for treatments	1	1.2%

PrI= predictable intervals, CrI= credible intervals. SMD= Standard Mean Difference. OR=odds ratio. SUCRA= surface under the cumulative ranking curve.

Table 4. Frequency of other tools implemented to assess risk of bias or quality of individual RCT in 67 NMA publications

Tool to assess risk of bias or quality of studies	n	(%)
Jadad score	27	40.30%
Modified Cochrane risk of bias tool.	11	16.42%
NICE	6	8.96%
Jadad score plus another tool	4	5.97%
CRD	3	4.48%
Evidence-Based Gastroenterology Steering Group.	3	4.48%
GRADE RoB	2	2.99%
Oxford scale modified	2	2.99%
CASP	1	1.49%
Clinical Trials Assessment Measure for psychological treatment	1	1.49%
Detsky scale	1	1.49%
Effective Public Health Practice Project Quality Assessment	1	1.49%
Institute for Quality and Efficiency in Health Care tool	1	1.49%
Newcastle–Ottawa Quality Assessment Scale tool modified	1	1.49%
Onychomycosis study quality assessment tool	1	1.49%
Oxford scale	1	1.49%
SIGN-50 checklist	1	1.49%
Total	**67**	**100%**

*** Studies did not apply all Cochrane risk of bias domains in individual randomized controlled trials.**

CASP: Critical Appraisal Skills Programme; CRD: Centre for Reviews and Dissemination; GRADE: Grading of Recommendations Assessment, Development and Evaluation; NMA: network-meta-analysis; NICE: National Institute for Health and Care Excellence; RCT: randomized controlled trials; RoB: Risk of Bias; SIGN: Scottish Intercollegiate Guidelines Network.

Table 5. Frequency of other tools implemented to interpret NMA findings in 127 NMA publications

Description		n	%
Ranking probabilities		81	29.3%
Direct treatment effect		19	6.9%
Direct estimates plus ranking		12	4.3%
GRADE assessment		6	2.2%
NMA treatment effect		3	1.1%
Heterogeneity		4	1.4%
Ranking plus sensitivity analysis		1	0.4%
Ranking plus inconsistency		1	0.4%
	Total	**127**	**100%**

GRADE: Grading of Recommendations Assessment, Development and Evaluation; NMA: network-meta-analysis

APPENDIXES

Appendix A. Search strategies

1. Database: Ovid MEDLINE(R)

In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to Present>

Search Strategy:

1 exp Meta-Analysis/ or meta analysis.mp. or meta analyses.mp. (109471)
2 systematic review*.mp. (74964)
3 1 or 2 (154239)
4 mixed treatment*.mp. (461)
5 multiple treatment*.mp. (2242)
6 multiple comparison*.mp. (7640)
7 mixed comparison*.mp. (12)
8 treatment network*.mp. (195)
9 comparison* network*.mp. (22)
10 mixed network*.mp. (36)
11 4 or 5 or 6 or 7 or 8 or 9 or 10 (10549)
12 3 and 11 (615)
13 network meta-analysis.mp. (834)
14 network meta-analyses.mp. (154)
15 mixed treatment* comparison*.mp. (338)
16 multiple treatment* comparison*.mp. (60)
17 mixed treatment meta analyses.mp. (2)
18 mixed treatment meta analysis.mp. (7)
19 multiple treatment* meta-analysis.mp. (51)
20 multiple treatment* meta-analyses.mp. (8)
21 network meta-regression*.mp. (6)
22 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 (1197)
23 Mixed evidence.mp. (428)

24 Indirect evidence.mp. (5855)
25 Indirect comparison*.mp. (986)
26 23 or 24 or 25 (7225)
27 Direct comparison*.mp. (12073)
28 Direct evidence.mp. (20590)
29 27 or 28 (32626)
30 26 and 29 (637)
31 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or metanalyses)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (246)
32 12 or 22 or 30 or 31 (2189)
33 remove duplicates from 32 (2025)

2. Database: Embase

<1974 to 2015 December 04>

Search Strategy:

1 exp Meta-Analysis/ or meta analysis.mp. or meta analyses.mp. (159546)
2 systematic review*.mp. (142037)
3 1 or 2 (237609)
4 mixed treatment*.mp. (804)
5 multiple treatment*.mp. (3139)
6 multiple comparison*.mp. (11597)
7 mixed comparison*.mp. (15)
8 treatment network*.mp. (336)
9 comparison* network*.mp. (29)
10 mixed network*.mp. (25)
11 4 or 5 or 6 or 7 or 8 or 9 or 10 (15848)
12 3 and 11 (931)
13 network meta-analysis.mp. (1403)
14 network meta-analyses.mp. (260)
15 mixed treatment* comparison*.mp. (636)

16 multiple treatment* comparison*.mp. (93)
17 mixed treatment meta analyses.mp. (2)
18 mixed treatment meta analysis.mp. (11)
19 multiple treatment* meta-analysis.mp. (59)
20 multiple treatment* meta-analyses.mp. (10)
21 network meta-regression*.mp. (4)
22 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 (2068)
23 Mixed evidence.mp. (488)
24 Indirect evidence.mp. (6576)
25 Indirect comparison*.mp. (1812)
26 23 or 24 or 25 (8815)
27 Direct comparison*.mp. (14673)
28 Direct evidence.mp. (21859)
29 27 or 28 (36495)
30 26 and 29 (813)
31 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or metanalyses)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword] (267)
32 12 or 22 or 30 or 31 (3314)
33 remove duplicates from 32 (3225)

3. Database: PsycINFO

<1806 to December Week 1 2015>

Search Strategy:

1 exp Meta-Analysis/ or meta analysis.mp. or meta analyses.mp. (21309)
2 systematic review*.mp. (14693)
3 1 or 2 (32329)
4 mixed treatment*.mp. (74)
5 multiple treatment*.mp. (453)
6 multiple comparison*.mp. (1730)
7 mixed comparison*.mp. (7)
8 treatment network*.mp. (112)

9 comparison* network*.mp. (5)
10 mixed network*.mp. (8)
11 4 or 5 or 6 or 7 or 8 or 9 or 10 (2373)
12 3 and 11 (125)
13 network meta-analysis.mp. (75)
14 network meta-analyses.mp. (20)
15 mixed treatment* comparison*.mp. (37)
16 multiple treatment* comparison*.mp. (13)
17 mixed treatment meta analyses.mp. (1)
18 mixed treatment meta analysis.mp. (0)
19 multiple treatment* meta-analysis.mp. (19)
20 multiple treatment* meta-analyses.mp. (2)
21 network meta-regression*.mp. (1)
22 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 (129)
23 Mixed evidence.mp. (425)
24 Indirect evidence.mp. (747)
25 Indirect comparison*.mp. (120)
26 23 or 24 or 25 (1284)
27 Direct comparison*.mp. (2095)
28 Direct evidence.mp. (2191)
29 27 or 28 (4281)
30 26 and 29 (108)
31 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or metanalyses)).mp. [mp=title, abstract, heading word, table of contents, key concepts, original title, tests & measures] (48)
32 12 or 22 or 30 or 31 (337)
33 remove duplicates from 32 (337)

4. Database – CINAHL

Interface - EBSCOhost Research Databases
Search Screen - Advanced Search
Database **- CINAHL<** December Week 1 2015**>**

Search Strategy:

S13 S7 OR S8 OR S11 OR S12 (424)

S12 (umbrella or overview) N3 ("Cochrane" or "systematic review" or meta analysis or meta analyses) (79)

S11 S9 AND S10 (57)

S10 Direct comparison* or Direct evidence (1286)

S9 Mixed evidence or Indirect evidence or Indirect comparison* (524)

S8 network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression* (266)

S7 S5 AND S6 (116)

S6 mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network* (1195)

S5 S1 OR S2 OR S3 OR S4 (52447)

S4 (MH "Systematic Review") (24519)

S3 systematic review* (41063)

S2 meta analyses or meta analysis (23985)

S1 (MH "Meta Analysis") (16968)

5. Database: LILACS

Search Strategy:

(tw:(network meta analysis)) OR (tw:(network meta analyses)) AND (instance:"regional") AND (db:("LILACS")) (14)

6. Database: Web of Science

< to December 2015**>**

Search Strategy:

32 (2,406) #31 OR #30 OR #22 OR #12

31 (270) TOPIC: (("umbrella" or "overview") NEAR/3 ("Cochrane" or "systematic review" or "meta analysis" or "metanalyses"))
30 (626) #29 AND #26
29 (44,580) #28 OR #27
28 (25,511) TOPIC: ("Direct evidence")
27 (19,115) TOPIC: ("Direct comparison*")
26 (7,813) #25 OR #24 OR #23
25 (1,091) TOPIC: ("Indirect comparison*")
24 (5,933) TOPIC: ("Indirect evidence")
23 (827) TOPIC: ("Mixed evidence")
22 (1,398) #21 OR #20 OR #19 OR #18 OR #17 OR #16 OR #15 OR #14 OR #13
21 (4) TOPIC: ("network meta-regression*")
20 (6) TOPIC: ("multiple treatment* meta-analyses")
19 (56) TOPIC: ("multiple treatment* meta-analysis")
18 (8) TOPIC: ("mixed treatment meta analysis")
17 (2) TOPIC: ("mixed treatment meta analyses")
16 (73) TOPIC: ("multiple treatment* comparison*")
15 (553) TOPIC: ("mixed treatment* comparison*")
14 (134) TOPIC: ("network meta-analyses")
13 (941) TOPIC: ("network meta-analysis")
12 (692) #11 AND #3
11 (11,580) #10 OR #9 OR #8 OR #7 OR #6 OR #5 OR #4
10 (217 TOPIC: ("mixed network*")
9 (34) TOPIC: ("comparison* network*")
8 (233) TOPIC: ("treatment network*")
7 (18) TOPIC: ("mixed comparison*")
6 (8,317) TOPIC: ("multiple comparison*")
5 (2,093) TOPIC: ("multiple treatment*")
4 (788) TOPIC: ("mixed treatment*")
3 (143,912 #2 OR #1
2 (76,250) TOPIC: ("systematic review*")
1 (94,698) TOPIC: ("Meta-Analysis" or "meta analysis" or "meta analyses")

7. Database: EBM Reviews - Cochrane Database of Systematic Reviews

<2005 to November 2015>

Search Strategy:

1 meta analys?s.mp. (7699)
2 systematic review*.mp. (7743)
3 1 or 2 (8450)
4 (mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network*).mp. (922)
5 3 and 4 (917)
6 (network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression*).mp. (192)
7 (Mixed evidence or Indirect evidence or Indirect comparison*).mp. (506)
8 (Direct comparison* or Direct evidence).mp. (867)
9 7 and 8 (239)
10 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or meta analyses)).mp. (168)
11 5 or 6 or 9 or 10 (1241)
12 remove duplicates from 11 (1241)

8. Database: EBM Reviews - Database of Abstracts of Reviews of Effects

<2nd Quarter 2015>

Search Strategy:

1 meta analys?s.mp. (19938)
2 systematic review*.mp. (33323)
3 1 or 2 (35363)
4 (mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network*).mp. (269)
5 3 and 4 (259)

6 (network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression*).mp. (303)
7 (Mixed evidence or Indirect evidence or Indirect comparison*).mp. (426)
8 (Direct comparison* or Direct evidence).mp. (450)
9 7 and 8 (145)
10 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or meta analyses)).mp. (24)
11 5 or 6 or 9 or 10 (580)
12 remove duplicates from 11 (580)

9. Database: EBM Reviews - Cochrane Methodology Register

<3rd Quarter 2012>

Search Strategy:

1 meta analys?s.mp. (2070)
2 systematic review*.mp. (2427)
3 1 or 2 (3843)
4 (mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network*).mp. (74)
5 3 and 4 (47)
6 (network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression*).mp. (46)
7 (Mixed evidence or Indirect evidence or Indirect comparison*).mp. (68)
8 (Direct comparison* or Direct evidence).mp. (62)
9 7 and 8 (25)
10 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or meta analyses)).mp. (19)
11 5 or 6 or 9 or 10 (95)
12 remove duplicates from 11 (95)

10. Database: EBM Reviews - Health Technology Assessment

<4th Quarter 2015>

Search Strategy:

1 meta analys?s.mp. (266)
2 systematic review*.mp. (1614)
3 1 or 2 (1730)
4 (mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network*).mp. (10)
5 3 and 4 (3)
6 (network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression*).mp. (9)
7 (Mixed evidence or Indirect evidence or Indirect comparison*).mp. (39)
8 (Direct comparison* or Direct evidence).mp. (94)
9 7 and 8 (15)
10 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or meta analyses)).mp. (2)
11 5 or 6 or 9 or 10 (28)
12 remove duplicates from 11 (28)

11. Database: EBM Reviews - NHS Economic Evaluation Database

<2nd Quarter 2015>

Search Strategy:

1 meta analys?s.mp. (1004)
2 systematic review*.mp. (2291)
3 1 or 2 (2813)

4 (mixed treatment* or multiple treatment* or multiple comparison* or mixed comparison* or treatment network* or comparison* network* or mixed network*).mp. (71)
5 3 and 4 (34)
6 (network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression*).mp. (44)
7 (Mixed evidence or Indirect evidence or Indirect comparison*).mp. (117)
8 (Direct comparison* or Direct evidence).mp. (138)
9 7 and 8 (15)
10 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or meta analyses)).mp. (2)
11 5 or 6 or 9 or 10 (67)
12 remove duplicates from 11 (67)

12. Database: AMED (Allied and Complementary Medicine)

<1985 to December 2015>

Search Strategy:

1 network meta-analysis.mp. (2)
2 multiple treatment* comparison*.mp. (1)
3 1 or 2 (3)
4 Mixed evidence.mp. (9)
5 Indirect evidence.mp. (24)
6 Indirect comparison*.mp. (4)
7 4 or 5 or 6 (37)
8 Direct comparison*.mp. (85)
9 Direct evidence.mp. (37)
10 8 or 9 (122)
11 7 and 10 (2)
12 ((umbrella or overview) adj3 ("Cochrane" or "systematic review" or meta analysis or metanalyses)).mp. (15)
13 3 or 11 or 12 (20)

13. Database: MathScinet

<Up to December 2015>

Search Strategy:

network meta-analysis or network meta-analyses or mixed treatment* comparison* or multiple treatment* comparison* or mixed treatment meta analyses* or mixed treatment meta analysis or multiple treatment* meta-analysis or multiple treatment* meta-analyses or network meta-regression* (34)

14. Database: Journal of Negative Results in Biomedicine

<Up to December 2015>

Search Strategy:

network meta-analysis (3)

Appendix B. List of included and excluded studies.

See the list in the following link:

https://www.dropbox.com/s/vev3icelqml0i8s/Appendix_B.docx?dl=0

CHAPTER 3. DEVELOPMENT OF A SUMMARY OF FINDINGS (SOF) TABLE FOR NETWORK META-ANALYSIS

PREFACE TO CHAPTER 3

Chapter 3: *Development of a Summary of Findings (SoF) Table for Network Meta-analysis* was submitted to the Journal of Clinical Epidemiology in April 2018 and is currently under review.

Chapter 3: Development of a Summary of Findings (SoF) Table for Network Meta-analysis

Authors

Juan José Yepes-Nuñez [1, 2], Shelly-Anne Li [3], Gordon Guyatt [1, 4], Susan M. Jack [1,5], Jan L. Brozek [1, 4], Joseph Beyene [1], M. Hassan Murad [6], Bram Rochwerg [1,4], Lawrence Mbuagbaw [1], Yuan Zhang [1], Ivan D. Flórez [1, 2], Reed A. Siemieniuk [1], Behnam Sadeghirad [1], Reem Mustafa [1, 5], Nancy Santesso [1], Holger J. Schünemann [1, 4]

Affiliations

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, ON, Canada.

2. School of Medicine. University of Antioquia, Medellín, Colombia.

3. University of Toronto, Canada

4. Department of Medicine, McMaster University, Hamilton, Canada

5. School of Nursing, McMaster University, Hamilton, Ontario, Canada

6. Mayo Clinic, Evidence-Based Practice Center, Rochester, Minnesota, United States

7. Division of Nephrology and Hypertension, University of Kansas Medical Center, Kansas City, Kansas, USA.

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact (HEI), Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON. L8S 4K1 Canada.
holger.schunemann@mcmaster.ca

Conflict of interest: The authors of this study declare no financial conflict of interest. They are members of the GRADE working group.

ABSTRACT

Objective: To develop a summary of findings (SoF) table format that displays the critical information from a network meta-analysis (NMA) including . **Study design and setting:** We applied a user experience model for data analysis based on four rounds of semi-structured interviews. **Results:** We interviewed thirty-two stakeholders who conduct or use meta-analysis (MA). Four rounds of interviews produced six candidate NMA-SoF tables. Users found highly acceptable a final NMA-SoF table that included the following components: 1) details of the clinical question (PICO), 2) a plot depicting network geometry, 3) relative and absolute effect estimates, 4) certainty of evidence, 5) ranking of treatments, and 6) interpretation of findings. **Conclusion:** Using stakeholder feedback, we developed a new NMA-SoF table that includes relevant components for understanding NMA findings and health decision-making.

Keywords: Network Meta-Analysis**,** Decision Making, Systematic Reviews, Summary of Findings, GRADE.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. It was internally supported by the McMaster GRADE centre.

HIGHLIGHTS

Key finding

- User testing methodology provided key information to develop a Grading of summary of findings table that displays network meta-analysis (NMA) findings.
- Fundamental components of the new NMA-SoF table include details of the question and interventions for a specific outcome, the relative effect estimate for each intervention, the anticipated absolute effects, the GRADE certainty of evidence, the rank probability of the intervention, and the interpretation of findings.

What this adds to what is known

- NMA publications are increasingly popular. However, the optimal presentation of their synthesized findings remains uncertain. The new NMA-SoF table provides relevant information in a simple and user-friendly format.

What are the implications? What should change now?

- NMA authors should consider presenting results in NMA-SoF tables based on the structure that our users have found informative and appealing.

INTRODUCTION

Multiple treatments are often available for patients with the same condition. Credible systematic reviews constitute the most trustworthy source of evidence to determine the effectiveness of interventions. Most systematic reviews focus on pair-wise direct comparisons of interventions, often with placebo as the comparator group (1). In the face of multiple available interventions, a lack of direct comparisons between active interventions can make determination of their relative desirability challenging (1).

When investigators have studied multiple interventions for the same condition, network meta-analysis (NMA) can provide information regarding how each option compares to the alternatives. Using direct and indirect evidence, NMA offers estimates of the relative impact of all the interventions even when head-to-head comparisons are unavailable (2).

Summaries of evidence for health professionals exist in different formats and for different purposes. The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Summary of Findings (SoF) tables present the main findings of conventional pair-wise systematic reviews in a transparent and simple form (3, 4). They provide key information concerning the certainty of the evidence and the magnitude of treatment effects for all patient-important outcomes (3-7).

Expert guidance (2, 8-15) suggests that the presentation of NMA findings in an NMA-SoF table should report: 1) the estimates of the relative and absolute effects of the interventions included in the NMA; 2) the certainty of the evidence (also known as confidence in effect estimates or quality of the evidence); 3) rank probabilities [or surface under the cumulative ranking curve (SUCRA)]; and 4) geometry of the NMA. NMA authors do not always adhere to this guidance.

Irrespective of adherence to standards, the optimal presentation of NMA remains uncertain and challenging, particularly when the alternatives are numerous (14). Although SoFs have proved extremely helpful for summarizing evidence from conventional systematic reviews, creating a similarly useful SoF for NMAs presents many challenges, and no one has thus far suggested a structure. Therefore, in response to the challenge of optimally NMA findings presentation, we developed and tested a novel format for NMA SoFs tables.

METHODS

We conducted this study in three phases: 1) an initial development of two draft NMA tables with feedback from GRADE working group members; 2) iterative modification of NMA-SoF table formats through four brainstorming meetings with a purposeful sample of researchers and; 3) user testing. The brainstorming meetings and the user testing phases were carried out alternately (Figure 1). A steering committee (JY, HJS, JLB, JB, GG) considered feedback from brainstorming meetings and user testing and made decisions regarding response

to feedback and modification of NMA-SoF table formats. Appendix A includes the protocol for the user testing.

Development of initial NMA SoF table format

Initially, a steering committee member (HJS) developed two draft versions of GRADE NMA-SoF tables (Appendix B) that he presented during a GRADE working group meeting in 2012, modified on the basis of feedback, and presented a second time at a workshop in 2013.

Brainstorming meetings

From 2013 to 2016, we conducted four brainstorming meetings with 42 health-care researchers with experience in conducting meta-analysis and NMAs. We documented the information retrieved from each meeting. At the first two brainstorming meetings, participants viewed the two-initial draft NMA-SoF tables and provided their opinions on the how they could be improved. At the third and fourth session, the participants viewed revised versions that the steering committee had developed on the basis of feedback from prior user testing rounds.

User testing

User testing (or usability testing) is a method of formative evaluation “with the goal of learning about the design from a user*′s* perspective to improve its next

iteration" in which products are tested by end users, as opposed to developers or experts (11, 14). The primary focus of our user testing was to explore the extent to which the NMA-SoF table formats meet the needs of the users and optimizes their understanding of the information reported. The principles of descriptive qualitative methods informed sampling, data collection and content analysis procedures (16).

We conducted four rounds of user testing. The first two rounds were developed after the first two brainstorming meetings. The third and fourth rounds were conducted after the third and fourth brainstorming meetings. Thus, the steering committee developed iterative versions of the NMA-SoF tables informed by feedback both from the brainstorming meeting and the user testing. Our final goal was to develop one NMA-SoF table for dichotomous outcomes and another table for continuous outcomes. We used an NMA-SoF table that displayed findings for dichotomous outcomes during the brainstorming meetings and user testing. We finally developed an NMA-SoF table for continuous outcomes based in the information retrieved from the brainstorm meetings and user testing.

Rounds

We defined "rounds" as a group of ten interviews using the same NMA-SoF table format with different users. After each round, the steering committee modified the NMA-SoF table format based on the comments and reflections provided by the

users. The new version of the NMA-SoF table was presented to a new group of users during each new round

Participants

Selection of participants

We used purposeful sampling, which represents a variety of non-probability sampling techniques, to select participants who could help us answer our research questions during the brainstorming meetings and user testing (17). We applied criterion sampling (18, 19) to ensure that the final pool of participants represented a wide range of users of the NMA-SoF tables.

Individuals were eligible if they considered themselves as Meta-analysis (MA) or NMA users. We guided potential participants in their self-classification as follows: we defined a user as someone who has used or published a MA or NMA at least once in the last two years to answer a clinical question related to patient health care, to inform the process of making recommendations for clinical practice guidelines, to inform other types of evidence-based decision-making, or for research purposes. Participants met at least one of the criteria. Two target populations participated in the user testing: 1) healthcare professionals as defined by WHO (14), and 2) researchers. We classified users as healthcare professionals if they dedicated at least 50% of their total time to clinical practice. To be considered a researcher, users had to report dedicating more than 70% of

their time to conducting research with or related to MA or NMA (including methodologists, epidemiologists, statisticians, etc.).

Setting and recruitment

We recruited users who met our definition through international networks of the study team, and from the GRADE working group. Potential users received an email invitation. The project coordinator (JY) carried out recorded interviews face to face or by video chat and voice calls (e.g. Skype) with those who provided informed consent. None of the users participated in more than one round.

Sample size

We planned to recruit approximately 20 users in three user testing rounds; we anticipated approximately half would be healthcare professionals, and half researchers and both groups would be represented during each round of interviews. We planned to conduct interviews until we had gathered enough information to fully develop the format of an NMA-SoF table and few if any insights were emerging from new interviews (sampling to redundancy) (20).

Interviews

We used a pilot tested, semi-structured format to first explore the impressions and understanding of the NMA SoF-table as a whole, and then of each NMA-SoF table elements using a user experience model (Honeycomb model) developed by

Morville (21). The Honeycomb model addresses seven separate facets of user experience: findability, accessibility, usability, usefulness, credibility, desirability, and value (21). All semi-structured interviews took approximately one hour, were audio recorded and included three sections: 1) background information of the users; 2) users' reactions to the NMA-SoF table including their first impressions, their needs, and understanding of the information, and; 3) an overall evaluation. The interviews ended with questions about how the interview process could be improved.

Data collection and transcription

The project coordinator made notes that summarized the discussion at the brainstorm meetings. In the user testing interviews, users viewed either a paper or electronic version of an NMA-SoF table. An individual with no other role in the study transcribed all audio-recorded data verbatim and, after transcription, deleted the recording.

Data Analysis

Data collection and analysis occurred concurrently, so that questions from the interview guide could be modified to allow for better capture of themes. The analysis was done using MAXQDA software (22).

We used both deductive and inductive approaches to analyze the data in duplicate. First, we used Hsieh and Shannon's (23) directed content

analysis (deductive approach) to determine the initial coding scheme based on Morville's Honeycomb model (21), a model that had been used successfully in previous research evaluating SoF tables for conventional pairwise meta-analysis (24, 25). The seven facets of the Honeycomb model served as the initial coding categories. We also included an Understandability code as it had been tested in other pairwise meta-analysis SoF studies (6, 26). Next, we developed operational definitions for each category based on the model. Coding began with reading and then highlighting all segments of data related to the categories. Next, we coded all the highlighted passages using the list of predetermined codes. Using inductive content analysis, we searched for new information that was not captured using the initial coding scheme. Any content that did not fit into any of the categories was assigned a new code according to Elo et al. (27). We used inductive content analysis because no other studies have investigated health professionals' experience with NMA-SoF tables (27).

Our analysis supported the pragmatic design goals of improving the latest NMA-SoF table after each round of user testing. For each interview transcription, two researchers (JY and SAL) independently organized the information by open coding, creating categories, and abstraction.

Ethical considerations

The Hamilton Integrated Research Ethics Board approved the research conducted in this project. All users provided both oral and written consent prior to the interview (review reference 2016-0956-GRA).

RESULTS

We conducted four brainstorming meetings and four rounds of interviews.

Initial NMA-SoF table formats

GRADE working group members and 11 clinicians and methodologists first viewed and commented on the initial two versions of the NMA SoF. They reported concerns including distracting box colours, failure to clearly highlight the certainty of evidence, omission of ranking treatment information and network geometry, failure to clearly identify the reference comparator intervention, and deficiencies in reporting the absolute effects and their interpretation in the NMA context. GRADE working group members also reported missing information regarding the direct and indirect estimates. At the end of the meeting, some users suggested seeing all the interventions compared to a single reference intervention in only one two-by-two table (or league table). Based on the above feedback, we created two new NMA-SoF table formats (Appendix C, and D), which we discussed at the brainstorming meetings.

Brainstorming meetings

We conducted the brainstorming meetings in three different countries: one in Colombia (Medellín, June 2014), two in Canada (Hamilton, July 2014 – July 2015), and a final one in South Korea (Seoul, October 2016). In total, 42 users provided information on five NMA-SoF table formats. Two of the brainstorm meetings in which participants commented on four NMA-SoF table formats (Colombian June 2014 and Canadian July 2014) took place before the first of the user testing. In the Colombian brainstorming meeting we explored format limitations in the two initial NMA-SoF tables (Appendix B). In the Canadian brainstorming meeting in 2014, we received feedback using two additional NMA-SoF table formats (Appendix C, and D). For the third and fourth brainstorming meetings, we obtained feedback on a single NMA-SoF table format (Appendix E). We used this format during the first three rounds of interviews.

The steering committee decided which elements should be included and which elements were less important to include. For instance, although some users found the initial designs appealing, the key issue was the lack of ability to present a large number of interventions in columns. We restructured the NMA-SoF table to display the NMA relevant information for one outcome, including relative and absolute effect estimates, certainty of the evidence and ranking probabilities for all the comparisons using a single comparator with the interventions displayed in rows.

User testing

We interviewed 32 users (Table 1). In addition to the seven facets included in the Honeycomb model, we identified another theme, which we labelled "suggestions".

User testing: First impressions

Users reported that the NMA-SoF table format (Appendix E) was similar to other GRADE SoF tables. Users found the relative and the absolute effects, the certainty of evidence, and the interpretation of findings the most engaging sections.

Users found inclusion of the network geometry and different colours for each intervention appealing, especially because they were able to easily identify each intervention in the network geometry and in the table. They found that displaying this information by outcome was clear and attractive considering the number of interventions reported in a single NMA-SoF table format. Specifically, the users found that presenting a single outcome per NMA-SoF table more readable than alternatives that included more than one outcome per NMA-SoF table. Users noted the drawback that, in a clinical decision context in which consideration of multiple outcomes is a necessary, the to refer to several NMA-SoF tables, one for each outcome, was a serious limitation.

User testing: Usability

Users, in general, found it easy to use the NMA-SoF table (Appendix E) when we presented the information with the clinical context. Presenting the information of NMA relative and absolute effects without the direct and indirect estimates was attractive for most users, but a minority would have preferred all three summary estimates.

"*I don't know. It is not possible to see all of the indirect or direct NMA estimate I believe, right?"*

The steering committee therefore attempted to add the direct and indirect effect estimates and their confidence intervals to the NMA-SoF table during the process design of the formats, but including this additional data added further complexity. The steering committee ultimately decided only to present the highest certainty of the available effect estimates, whether it was the network estimate, the direct estimate, or the indirect estimate.

Accessibility

Users did not report any substantial difficulties accessing the information reported in the NMA-SoF table. For example, users found the information in the NMA-SoF table (PICO information, network geometry, NMA relative effects, etc.) easily accessible:

"*Well, I think it's helpful you have the question at the very top and kind of deliberations on the actual PICO components. And at the top I think that's helpful.*"

For NMA-SoF tables with fewer than 6 comparisons, a single page was sufficient, and users worked on the table without any accessibility issues. However, in the examples with more than 6 interventions, the table did not fit onto a single page. Users needed to study these NMA-SoF tables for an appreciably longer period of time than the NMA-SoFs that fit onto a single page.

Some users expressed concerns about the chosen colours and suggested that we should consider the needs of users with color visual impairment. In response, the final NMA-SoF table format will not include red and green to avoid issues with the most common colour blindness.

Usefulness

In a clinical decision scenario, users reported that a single NMA-SoF table was useful:

"*I think it would be useful. The key challenge is always to match the patient, [that] the patient matches what is in there. I think the rank would be helpful to decide [on the option].*"

Users found all the components of the NMA-SoF table helpful, but some aspects, including the network geometry, and the certainty of evidence assessment, were particularly helpful.

Credibility

In general, users found the information reported in the NMA-SoF table trustworthy because they were able to identify information about the NMA relative effect, the absolute effect, and the certainty of evidence in a same NMA-SoF table format:

"…*So that does, why I would trust this information because I see the anticipated absolute effect. That's all relative effects and the certainty in the evidence. That's all clearly shown here.*"

Users made suggestions about how to increase the value of the table by adding some information that was not shown in the initial NMA-SoF table format. The number of RCTs included in direct estimates of a specific intervention, and the numbers of participants in a specific pairwise comparison were two additional characteristics added in response to this feedback. We added Information in the NMA relative effect estimate column that explains if the estimate was from the direct, the indirect, or both bodies of evidence.

Findability

Users did not report challenges locating the information in the NMA-SoF table that included the PICO section and the network graph that oriented users to the reference comparator. Because some users were unfamiliar with NMA terminology, we added a section with NMA definitions and explanations for all abbreviations.

Desirability and Value

Users did not report negative emotional reactions and expressed a need to have SoF tables accompanying NMA reports. Users noted that this NMA-SoF table can be part of an NMA publication but also can be used in a clinical practice guideline context:

"…*Yes, these tables would probably be in the appendix of the guideline, it would be in the original report of the NMA meta-analysis. And eventually an abbreviated version could be actually in the text but then I don't think that you would necessarily*".

Understandability

In the first round of interviews, we received feedback to improve the understanding of the information reported in the network diagram, the ranking probabilities, and the interpretation of findings. For the network diagram, we

increased the size of the circles (which represent individual interventions) and added an abbreviation label for the interventions. We kept the solid lines linking the interventions that were directly compared. We reported, for each intervention, along with the absolute effect for the population in the studies, an additional row with baseline risk information taken from an observational study. Users did not completely understand the report of two rows with baseline risk information for one intervention in the context of an NMA.

We presented two approaches to reporting NMA-SoF rankings: the probability of being the best intervention for each outcome, and SUCRA. In general, users had issues understanding the information reported about the ranking, regardless of the approach used. SUCRA was the most appealing format for most; users suggested adding the uncertainty of the information with corresponding credible intervals. The last column of the NMA-SoF table was labeled as “what happens”. This term is used in a regular current pairwise comparison GRADE SoF table. Users had some difficulty understanding the label, which we changed to “Interpretation of Findings” after the first round of interviews.

During the second round of interviews, we added information including the setting population (PICO component), the number of RCTs, as well as the number of participants for each direct pairwise comparison, the GRADE domain for downgrading the certainty of evidence, the median rank with the credible intervals for each intervention, and a section of NMA definitions in the lowest part of the

table. Users understood each of these elements of the NMA-SoF table. We presented two baseline risk information in two rows for each intervention. Users, even though they expressed their familiarity with baseline risk information reported in pairwise SoF tables, did not understand the presence of two-baseline risk information for each intervention in the context of an NMA. The NMA-specific definitions in the footnotes facilitated understanding of NMA nomenclature.

In the third round, users easily understood the source of the NMA relative effect estimate and the information regarding the anticipated absolute effects when we kept only the information for the population in the studies, that is, one baseline risk for each intervention. However, they suggested modifying the labels of the columns that display the anticipated absolute effects as: “with [reference comparator]”, “with intervention”, and “difference”.

Suggestions

Some components in the NMA-SoF table elicited additional suggestions from the users. They suggested adding explanations of the GRADE assessment judgments in the context of NMA, using plain language that facilitates optimal understanding:

“*I think generally yes they are helpful and we should include, I would propose to include footnotes asking for standard SoF but I think in this specific case it would need to be slightly more detailed because I am interested to know*

kind of the amount of heterogeneity and what is the rationale for downgrading precision for example."

Final NMA-SoF table

The final NMA-SoF table is composed of three sections (tables 2 and 3). *Upper section*

The upper section displays information regarding the PICO components. One outcome needs to be chosen for each NMA-SoF table. Also, one intervention needs to be selected as a "reference comparator" and the other interventions are listed under the "intervention" label. A network graph for the entire network is included in this section.

Middle section

This section is composed of eight columns that report the following information: 1) Interventions for a specific outcome. Below each intervention is a description of the number of participants included in the direct comparison and whether the contribution for the relative effect comes from the direct evidence, the indirect evidence or both; 2) the relative effect estimate for each intervention, which is calculated relative to the reference comparator; 3) three columns, two of them that report the anticipated absolute effects information relative to the reference comparator, for each of the interventions, and one that reports the risk difference; 4) the certainty of evidence with the rational for downgrading the body of the

evidence, 5) the ranking treatment which can be expressed as a median or SUCRA with the corresponding credible interval, 6) and the interpretation of findings that describes the level of superiority or inferiority of each intervention compared to the reference comparator after combination of the relative effect estimate, the certainty of evidence, and the rank probability components.

Lower section

This section has three cells with: 1) definitions about NMA terminology and abbreviations used in the NMA-SoF table, 2) description of each of the final GRADE certainty of evidence judgments, and 3) the explanatory footnotes that describe in detail the rationale for the certainty of evidence assessments.

Considerations of benefit and harms and continuous outcomes in NMA-SoF tables

Users received two NMA-SoF tables simultaneously during the fourth round of interviews. The purpose of presenting these two tables was to obtain feedback about both beneficial and harmful outcomes. In general, users felt it was useful to have access to beneficial and harmful outcomes. They understood the information reported in the tables and no relevant issues related with content and distribution of information emerged. Although users mentioned the two NMA-SoF tables looked clean and well organized, they also felt worried due to high number of interventions reported in each NMA-SoF table and its interpretation for

decision-making. The steering committee also developed an NMA-SoF table for continuous outcomes based on this final NMA-SoF table format (Appendix F)

A Second final NMA-SoF table

The final table displays summary of evidence for one outcome at a time. Thus, to use the information for clinical or policy decisions, stakeholders will need multiple separate SoF tables to present summary of evidence for each patient important outcomes. This constitutes an important limitation of our final table.

To overcome this challenge, the steering committee developed a second final NMA-SoF table (Table 4) using one of the first two NMA-SoF table formats as a template but dealing with its shortcomings on the basis of what we had learned from the brainstorming meetings and user testing. This format is focused on a smaller number of interventions that matter based on the highest certainty of the evidence for the benefit outcome, ideally the top three options against the chosen key comparator.

DISCUSSION

Main findings

We conducted a descriptive qualitative study to develop an NMA-SoF table format using strategies to obtain information from diverse constituencies including clinicians and researchers. The final rendering of the table looked familiar to most

users. The final colours in the NMA-SoF table produced a positive reaction. The final NMA-SoF table format includes the main aspects that NMA reports address and captures the complexity of the information reported in a NMA publication while maximizing simplicity to achieve a user-friendly presentation.

The evolution of NMA-SoF formats

We tested six different NMA-SoF table formats. The first two NMA-SoF table formats (Appendix B) provided an advantage over the final NMA-SoF table format, as they displayed multiple outcomes information in a single presentation. However, when a large number of interventions were compared, their format became unwieldy.

The third and fourth NMA-SoF tables format offered advantage of reporting all estimates of all comparison of one or more interventions to each other for a specific outcome. Users still felt, however, confused and overwhelmed due to the large amount of information from multiple interventions. We therefore included evidence for one main comparator versus other interventions for a single outcome in our final format.

The selection of the comparator remains challenging. We suggest the following options to choose the reference comparator for the NMA-SoF table: 1) a placebo intervention, 2) a “gold” standard treatment for the condition under review, 3) or the most cost-effective intervention, 4) or the least effective intervention. We

suggest presenting the interventions by row in the NMA-SoF table based on the ranking order.

We initially reported two different baseline risk estimates. Understanding risks is challenging and misinterpretation is common (28). Therefore, we ultimately included only one baseline risk. This modification enhanced accessibility of the information and increased understanding of NMA findings.

The final NMA-SoF table format (Appendix E and table 3) captures the complexity of the information reported in a NMA publication. The network graph proved understandable for users. Previous work validated the importance of presenting relative estimates in a consistent format for understanding the effect of an intervention (25, 29). Reporting absolute effect estimates is also needed to make trade-offs between interventions with beneficial and harmful effects (30, 31). We also included natural frequencies to report the difference between two absolute effects as it improves understandability of communicating findings (32).

We included ranking information in our NMA-SoF table. Ranking probabilities need to be integrated in the context of health care decision making as different factors can influence the rank probabilities such as the sample size of individual studies, network geometry, number of studies included for each treatment comparison, and the estimated treatment effects (33, 34).

Ranking can, for several reasons, also be misleading: 1) the evidence on which the rankings are based may be of very low quality; 2) a treatment that is best in one outcome (e.g. a benefit outcome) may be the worst in another outcome (e.g., a harm outcome); or 3) issues such as cost and a clinician's familiarity with use of a particular treatment may also bear consideration (35). We encourage users to interpret ranking information with caution. Based on our findings, we suggest presenting the ranking as the median ranking or the SUCRA with - given that rankings generated in an NMA can have a high degree of imprecision (36) –the associated credible intervals.

An alternative to our final NMA-SoF table

Because of the important limitation of our final tables – the necessity for multiple tables, one for each outcome, the steering committee modified one of the two initial templates to create a second final table. Thus, according to primary interest, NMA authors can produce one of two NMA-SoF table formats (or develop both). One format presents multiple tables reports on all the available interventions with one table for each outcome, while the other presents a single NMA SoF-table displaying the findings for the top three interventions across multiple outcomes.

Strengths and limitations

Our qualitative design allowed transparency for the data analysis process. Two investigators conducted the content analysis of all transcripts independently and in duplicate, which enhances the reliability of findings. We reached saturation (no new information or themes emerging from the data or data saturation) in the fourth round of interviews, suggesting that we achieved a sufficient sample size to make conclusions about users' experience. We used a triangulation of multiple data sources to identify areas of convergence. We included a broad range of users with different backgrounds and experience in research (methodologists and biostatisticians) and clinical fields. Thus, we integrated statistical and clinical approaches in a single NMA-SoF table (37).

This study has limitations. We did not test an NMA-SoF table format that displays findings of continuous outcomes. However, based on the user testing findings, we designed a format to report results of continuous outcomes from a recent NMA publication (38) (Appendix F). Although we tested our NMA-SoF table with findings from studies that analyzed data using Bayesian approach, the NMA-SoF table can also be used in the context of frequentist approach. The NMA-SoF table was not evaluated in randomized control trial.

Conclusion, and further research

Our NMA-SoF table represents a good starting point to present NMA results to a diverse group of stakeholders. In addition to the content of a pairwise comparison SoF table, we suggest to report the following elements in an NMA-SoF table:

- Network geometry
- Relative effect estimates for the highest certainty of the evidence
- Baseline risk information for only the risk of population in included studies
- Certainty of the evidence for the NMA estimates with judgments for downgrading the body of the evidence
- Ranking treatment and its uncertainty
- Include text with definitions of NMA topics (e.g. ranking, absolute effects)

Further work includes further study of how to present NMA findings for continuous outcomes and the development of interactive SoFs that can also be included in evidence-to-decision frameworks.

ACKNOWLEDGEMENTS

The authors would like to thank the participants of the interviews who helped us to create a NMA-SoF table.

REFERENCES

1. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Bmj. 2005;331(7521):897-900.

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

3. Schünemann HJ OA, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT GS, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011). Available from www.cochrane-handbook.org: The Cochrane Collaboration, 2008; 2008.

4. Schünemann HJ OA, Higgins JPT, Vist GE, Glasziou P, Guyatt GH. Chapter 11: Presenting results and 'Summary of findings' tables. In: Higgins JPT GS, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011). Available from www.cochrane-handbook.org: The Cochrane Collaboration, 2008; 2008.

5. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. Journal of clinical epidemiology. 2013;66(2):173-83.

6. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Improving GRADE evidence tables part 1: a

randomized trial shows improved understanding of content in summary of findings tables with a new format. Journal of clinical epidemiology. 2016;74:7-18.

7. Langendam M, Carrasco-Labra A, Santesso N, Mustafa RA, Brignardello-Petersen R, Ventresca M, et al. Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed. Journal of clinical epidemiology. 2016.

8. Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2009;181(8):488-93.

9. Cipriani A, Higgins JP, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Annals of internal medicine. 2013;159(2):130-7.

10. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. Medical decision making : an international journal of the Society for Medical Decision Making. 2013;33(5):679-91.

11. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, 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 in health : the journal of

the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):417-28.

12. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):429-37.

13. Tan SH, Bujkiewicz S, Sutton A, Dequen P, Cooper N. Presentational approaches used in the UK for reporting evidence synthesis using indirect and mixed treatment comparisons. Journal of health services research & policy. 2013;18(4):224-32.

14. Sullivan SM, Coyle D, Wells G. What guidance are researchers given on how to present network meta-analyses to end-users such as policymakers and clinicians? A systematic review. PloS one. 2014;9(12):e113277.

15. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Annals of internal medicine. 2015;162(11):777-84.

16. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description - the poor cousin of health research? BMC medical research methodology. 2009;9:52.

17. Miles M, Huberman A. Qualitative Data Analysis: An Expanded Sourcebook. 2 ed. Thousand Oaks: SAGE Publications; 1994

18. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. Administration and policy in mental health. 2015;42(5):533-44.

19. Patton MQ. Chapter 5. Designing qualitative studies. Qualitative Research & Evaluation Methods. 3 ed. Thousand Oaks, CA: Sage Publications; 2002. p. 230-47.

20. Creswell JW. Five quialitative approached to inquiry. In: Creswell J, editor. Quality inquiry & Research Design. 3rd ed. California: SAGE; 2013. p. 69-110.

21. Morville P. User experience design 2004 [updated June 21, 2004. Available from: http://www.semanticstudios.com/publications/semantics/000029.php.

22. VERBI Software GmbH. MAXQDA. The Art of Data Analysis In: Gerson J, editor. Berlin VERBI GmbH; 2017.

23. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qualitative health research. 2005;15(9):1277-88.

24. Rosenbaum SE, Glenton C, Cracknell J. User experiences of evidence-based online resources for health professionals: user testing of The Cochrane Library. BMC medical informatics and decision making. 2008;8:34.

25. Rosenbaum SE, Glenton C, Nylund HK, Oxman AD. User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. Journal of clinical epidemiology. 2010;63(6):607-19.

26. Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. Journal of clinical epidemiology. 2010;63(6):620-6.

27. Elo S, Kyngas H. The qualitative content analysis process. Journal of advanced nursing. 2008;62(1):107-15.

28. Reyna VF. How People Make Decisions That Involve Risk. A Dual-Processes Approach. Curr Dir Psychol Sci. 2004;13(2):60-6.

29. Lipkus IM. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. Medical decision making : an international journal of the Society for Medical Decision Making. 2007;27(5):696-713.

30. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. Bmj. 2003;327(7417):741-4.

31. Paling J. Strategies to help patients understand risks. Bmj. 2003;327(7417):745-8.

32. Trevena LJ, Davey HM, Barratt A, Butow P, Caldwell P. A systematic review on communicating with patients about evidence. Journal of evaluation in clinical practice. 2006;12(1):13-23.

33. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. Clinical epidemiology. 2014;6:451-60.

34. Veroniki AA, Straus SE, Rucker G, Tricco AC. Commentary: Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? Journal of clinical epidemiology. 2018.

35. Mbuagbaw L, Rochwerg B, Jaeschke R, Heels-Andsell D, Alhazzani W, Thabane L, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. Systematic reviews. 2017;6(1):79.

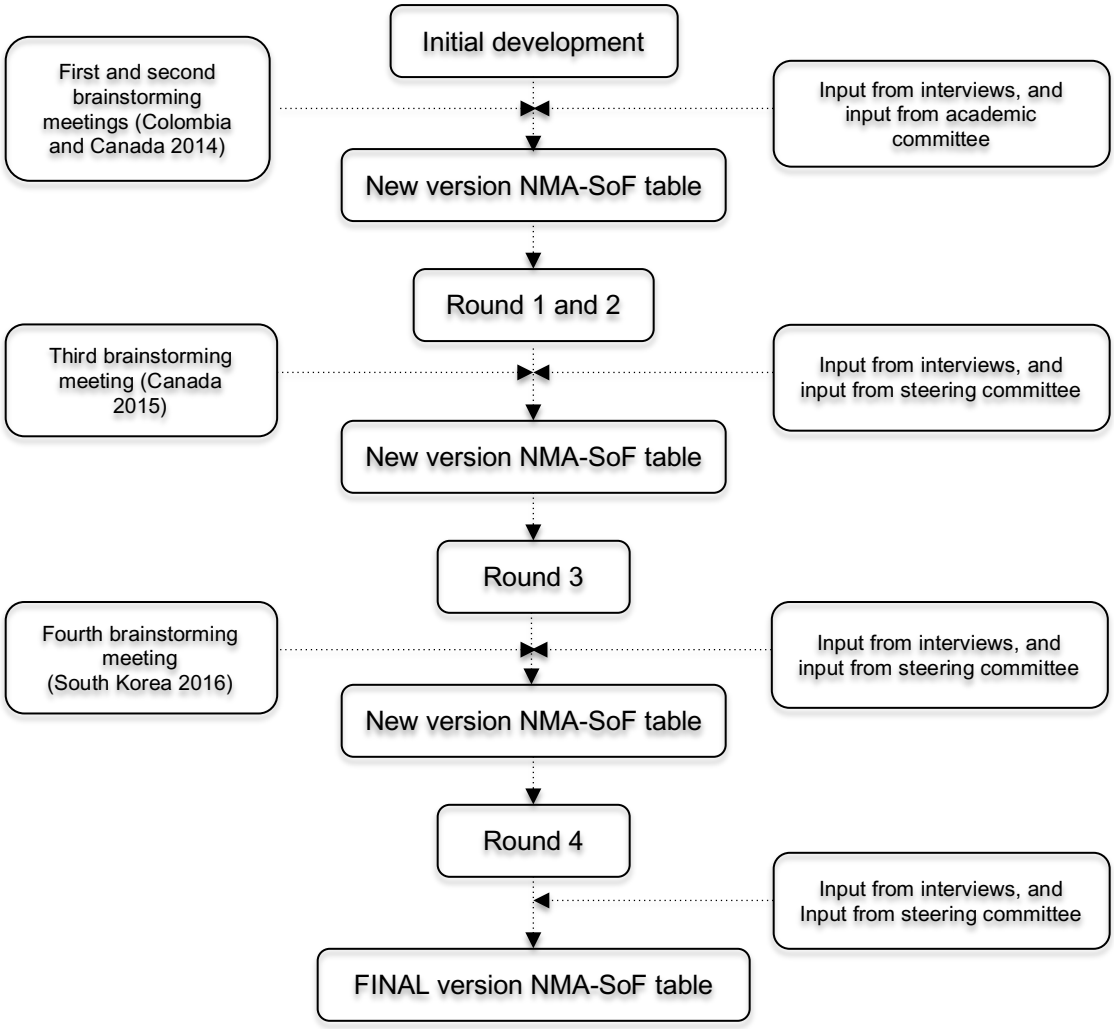
36. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. Annals of internal medicine. 2016;164(10):666-73.

37. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? Bmj. 2008;336(7659):1472-4.

38. Yuan JQ, Mao C, Wong SY, Yang ZY, Fu XH, Dai XY, et al. Comparative Effectiveness and Safety of Monodrug Therapies for Lower Urinary Tract Symptoms Associated With Benign Prostatic Hyperplasia: A Network Meta-analysis. Medicine. 2015;94(27):e974.

FIGURES

Figure 1. Phases conducted to develop the GRADE NMA-SoF table



TABLES

Table 1. Characteristics of participants

Formal education users (n=32)	
Healthcare-related professional	62.5%
Master	53.3%
PhD	59.3%
*Years of experience**	
Healthcare-related professional (n=19)	15.2 (8.7)
Researcher (n=25)	12.8 (5.5)
Frequency of using/reading scientific literature (n%)	
Never	0
Less than once per year	0
1 to 4 times per year	0
5 to 10 times per year	0
More than 10 times per year	29 (100)
Frequency in using NMA literature (decision-making or academia) (n%)	
Never	3 (10.3)
Less than once per year	2 (6.9)
1 to 4 times per year	10 (34.5)
5 to 10 times per year	6 (20.7)
More than 10 times per year	8 (27.6)
Familiarity with GRADE (n%)	
Very familiar	23 (79.3)
Somewhat familiar	2 (6.9)
A little familiar	2 (6.9)
Not familiar at all	2 (6.9)

*Mean, SD; NMA: Network Meta-analysis; GRADE: Grading of Recommendations Assessment, Development and Evaluation.

Table 2. NMA-SoF table final format

Estimates of effects, credible intervals, and certainty of the evidence for in XXXXX

Bayesian NMA-SoF table

BENEFITS

Patient or population:

Interventions:

Comparator (reference):

Outcome:

Setting:

Network Meta-analysis geometry plot

Total studies: **Total Participants**:		**Relative effect** (95% CrI)**	**Anticipated absolute effect*** (95% CrI)**			**Certainty of the evidence**	**Ranking**** (95% CrI)**	**Interpretation of Findings**
			***Without* intervention**	***With* intervention**	**Difference**			

NMA-SoF table definitions

* Lines represent direct comparisons

** Estimates are reported as odds ratio. CrI: credible interval. Results are expressed in credible intervals as opposed to the confidence intervals (9) since a Bayesian analysis has been conducted.

*** Anticipated absolute effect. Anticipated absolute effect compares two risks by calculating the difference between the risks of the intervention group with the risk of the control group.

**** Surface under the cumulative (SUCRA) ranking and credible intervals for efficacy are presented. Rank statistics is defined as the probabilities that a treatment out of *n* treatments in a network meta-analysis is the best, the second, the third and so on until the least effective treatment.

GRADE Working Group grades of evidence (or certainty in the evidence)

High quality: We are very confident that the true effect lies close to that of the estimate of the effect

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect **Very low quality:** We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect
Explanatory Footnotes

Table 3. NMA-SoF table template for dichotomous outcomes

Bayesian NMA-SoF table

BENEFITS

Estimates of effects, credible intervals, and certainty of the evidence for chemoprevention of colorectal cancer in individualswith previous colorectal neoplasia

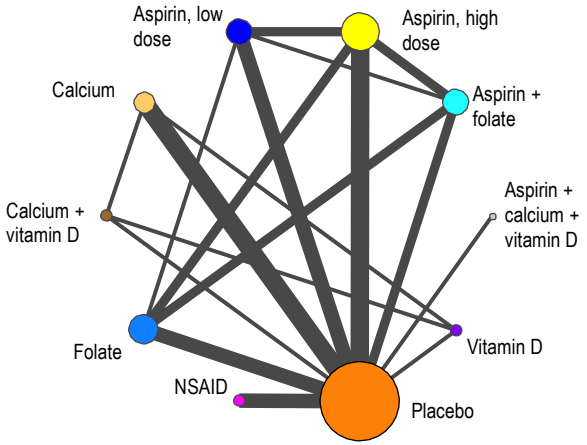
Patient or population: Individuals with previous colorectal neoplasia

Interventions: Low and high dose aspirin, nonaspirin non-steroidal anti-inflammatory drugs (NSAIDs), calcium, vitamin D, folic acid

Comparator (reference): Placebo

Outcome: Prevention of advanced neoplasia; range of follow up between three to five years

Setting: Outpatient



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[1]	53 per 1000	21 fewer per 1000 (61 fewer to 110 more)	⨁⨁◯◯ **Low** Due to Imprecision[2, 5]	3 (1 to 10)	-
●	Calcium + vitamin D (1 RCT; 1028 participants)	**OR 0.91** (0.52 to 1.63) Network estimate	74 per 1000[1]	67 per 1000	7 fewer per 1000 (36 fewer to 47 more)	⨁⨁◯◯ **Low** Due to Imprecision[2, 5]	6 (1 to 10)	-

	Aspirin + folate (2 RCT; 916 participants)	**OR 0.73** (0.43 to 1.19) Network estimate	74 per 1000[1]	54 per 1000	20 fewer per 1000 (42 fewer to 14 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 5]	4 (2 to 8)	-
	Aspirin, high dose (3 RCT; 917 participants)	**OR 0.81** (0.50 to 1.28) Network estimate	74 per 1000[1]	60 per 1000	14 fewer per 1000 (37 fewer to 21 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 5]	5 (2 to 9)	-
	Aspirin, low dose (3 RCT; 823 participants)	**OR 0.71** (0.41 to 1.23) Network estimate	74 per 1000[1]	53 per 1000	21 fewer per 1000 (44 fewer to 17 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 5]	3 (2 to 9)	-
	Nonaspirin NSAIDs (4 RCT; 3486 participants)	**OR 0.37** (0.24 to 0.53) Network estimate	74 per 1000[1]	27 per 1000	47 fewer per 1000 (56 fewer to 35 fewer)	⊕⊕⊕⊕ **High**[5]	1 (1 to 2)	-
	Vitamin D (1 RCT; 764 participants)	**OR 1.19** (0.65 to 2.15) Network estimate	74 per 1000[1]	88 per 1000	14 more per 1000 (26 fewer to 85 more)	⊕⊕◯◯ **Low** Due to Imprecision[3, 5]	9 (3 to 10)	-
	Calcium (3 RCT; 2503 participants)	**OR 1.00** (0.66 to1.52) Network estimate	74 per 1000[1]	74 per 1000	0 fewer per 1000 (25 fewer to 38 more)	⊕⊕◯◯ **Low** Due to Imprecision[4, 5]	7 (3 to 10)	-
	Folate (3 RCT; 1224 participants)	**OR 1.32** (0.85 to 2.00) Network estimate	74 per 1000[1]	51 per 1000	23 more per 1000 (11 fewer to 74 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 5]	9 (5 to 10)	-
	Placebo	Reference comparator	No estimable	No estimable	No estimable	Reference comparator	7 (4 to 9)	-

NMA-SoF table definitions

* Lines represent direct comparisons

** Estimates are reported as odds ratio. CrI: credible interval. Results are expressed in credible intervals as opposed to the confidence intervals since a Bayesian analysis has been conducted.

*** Anticipated absolute effect. Anticipated absolute effect compares two risks by calculating the difference between the risks of the intervention group with the risk of the control group.

**** Median rank and credible intervals for efficacy outcome are presented. Rank statistics is defined as the probabilities that a treatment out of *n* treatments in a network meta-analysis is the best, the second, the third and so on until the least effective treatment.

GRADE Working Group grades of evidence (or certainty in the evidence)
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect
Explanatory Footnotes
[1] Baseline risks (assumed control risk) obtained from the National Cancer Institute pooling project
[2] Very serious imprecision since 95% CrI crosses unity, and with wide credible intervals suggesting high possibility of harm.
[3] Very serious imprecision since RR>1 (suggesting greater likelihood of harm than benefit), and with wide credible intervals).
[4] Very serious imprecision since RR is one (suggesting no evidence of benefit) and wide credible intervals suggesting high possibility of harm.
[5] Conceptually, there was no significant intransitivity, with comparable distribution of plausible effect modifiers across trials of different chemopreventive agents.

Estimates of effects, credible intervals, and certainty of the evidence for chemoprevention of colorectal cancer in individuals with previous colorectal neoplasia

Bayesian NMA-SoF table

HARMS

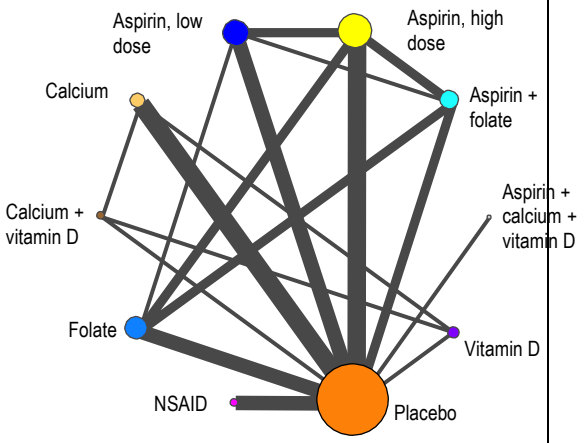
Patient or population: Individuals with previous colorectal neoplasia

Interventions: Low and high dose aspirin, nonaspirin non-steroidal anti-inflammatory drugs (NSAIDs), calcium, vitamin D, folic acid

Comparator (reference): Placebo

Outcome: Serious adverse events; range of follow up between three to five years

Setting: Outpatient



Total studies: 21 RCT	**Relative effect****	**Anticipated absolute effect*** (95% CrI)**		**Ranking****	

Total Participants: 14135		**(95% CrI)**	***Without* intervention**	***With* intervention**	**Difference**	**Certainty of**	**(95% CrI)**	**Interpretation of**
●	Aspirin + calcium + vitamin D (1 RCT; 714 participants)	**OR 0.90** (0.54 to1.51) Network estimate	187 per 1000[1]	89 per 1000	15 more per 1000 (71 more to 77 fewer)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	4 (2 to 7)	-
●	Calcium + vitamin D (1 RCT; 1125 participants)	**OR 1.11** (0.76 to 1.70) Network estimate	187 per 1000[1]	203 per 1000	16 more per 1000 (38 fewer to 94 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	2 (1 to 7)	-
●	Aspirin + folate (3 RCT; 1017 participants)	**OR 1.21** (0.83 to 1.77) Network estimate	187 per 1000[1]	218 per 1000	31 more per 1000 (27 fewer to 102 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	10 (6 to 10)	-
●	Aspirin, high dose (3 RCT; 1507 participants)	**OR 1.06** (0.76 to 1.49) Network estimate	187 per 1000[1]	196 per 1000	9 more per 1000 (38 fewer to 68 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	6 (1 to 10)	-
●	Aspirin, low dose (2 RCT; 794 participants)	**OR 0.78** (0.43 to 1.38) Network estimate	187 per 1000[1]	152 per 1000	35 fewer per 1000 (54 more to 97 fewer)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	8 (3 to 10)	-
●	Nonaspirin NSAIDs (3 RCT; 3964 participants)	**OR 1.23** (0.95 to 1.64) Network estimate	187 per 1000[1]	221 per 1000	34 more per 1000 (8 fewer to 87 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	2 (1 to 9)	-
●	Vitamin D (1 RCT; 835 participants)	**OR 1.10** (0.74 to 1.70) Network estimate	187 per 1000[1]	212 per 1000	25 more per 1000 (20 fewer to 78 more)	⊕⊕◯◯ **Low** Due to Imprecision[2, 3]	5 (2 to 10)	-
●	Calcium (4 RCT; 2669 participants)	**OR 1.38** (1.07 to 1.89) Network estimate	187 per 1000[1]	238 per 1000	51 more per 1000 (22 more to 82 more)	⊕⊕⊕⊕ **High**[3]	8 (3 to 10)	-

●	Folate (3 RCT; 1511 participants)	**OR 0.85** (0.59 to 1.22) Network estimate	187 per 1000[1]	165 per 1000	22 fewer per 1000 (21 more to 59 fewer)	⨁⨁◯◯ **Low** Due to Imprecision[2, 3]	6 (2 to 10)	-
●	Placebo	Reference comparator	No estimable	No estimable	No estimable	Reference comparator	3 (1 to 10)	-

NMA-SoF table definitions

* Lines represent direct comparisons

** Estimates are reported as odds ratio. CrI: credible interval. Results are expressed in credible intervals as opposed to the confidence intervals since a Bayesian analysis has been conducted.

*** Anticipated absolute effect. Anticipated absolute effect compares two risks by calculating the difference between the risks of the intervention group with the risk of the control group.

**** Median rank and credible intervals for harm outcome are presented. Rank statistics is defined as the probabilities that a treatment out of *n* treatments in a network meta-analysis is the best, the second, the third and so on until the least effective treatment.

GRADE Working Group grades of evidence (or certainty in the evidence)

High quality: We are very confident that the true effect lies close to that of the estimate of the effect

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect

Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Explanatory Footnotes

[1] Based on assumed control risk of 18.7% (corresponding to pooled 18.7% risk of SAEs in placebo-treated patients of included trials)

[2] Very serious imprecision since 95% CrI crosses unity, and with wide credible intervals suggesting uncertainty in the estimate.

[3] Conceptually, there was no significant intransitivity, with comparable distribution of plausible effect modifiers across trials of different chemopreventive agents.

Table 4. NMA-SoF table format reporting multiple outcome information for multiple comparison treatments

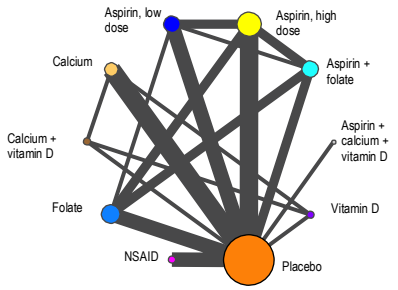
Estimates of effects, credible intervals, and certainty of the evidence for chemoprevention of colorectal cancer in individualswith previous colorectal neoplasia

Patient or population: Individuals with previous colorectal neoplasia

Interventions: Low and high dose aspirin, nonaspirin non-steroidal anti-inflammatory drugs (NSAIDs), calcium, vitamin D, folic acid

Comparison: Placebo

Settings: Outpatient, range of follow up between three to five years



Outcome	Effects and confidence in the estimate of effects						Comments
	Nonaspirin NSAIDs		**Aspirin, low dose**		**Aspirin + calcium + vitamin D**		
Prevention of neoplasia Follow up: range from 24 months to 60 months							
Placebo Comparator 74 per 1000[1] (7.4%)	**OR 0.37** (0.24 to 0.53) Network estimate	**47 fewer per 1000** (56 fewer to 35 fewer)	**OR 0.71** (0.41 to 1.23) Network estimate	**21 fewer per 1000** (44 fewer to 17 more)	**OR 0.71** (0.18 to 2.49) Network estimate	**21 fewer per 1000** (61 fewer to 110 more)	None of the ranking treatments between placebo versus other NSAIDs, calcium, vitamin D, or folic acid were highest from the ones we reported. Therefore, we did not include other comparisons in the table.
	⊕⊕⊕⊕ **High** Confidence in estimate		⊕⊕◯◯ **Low** Confidence in estimate due to Imprecision [2, 3]		⊕⊕◯◯ **Low** Confidence in estimate due to Imprecision [2, 3]		
Rank 7 (4 to 9)	**Rank**[4] 1 (1 to 2)		**Rank** 3 (2 to 9)		**Rank** 3 (1 to 10)		
	Based on 3,486 participants (4 RCT)		Based on 823 participants (3 RCT)		Based on 427 participants (1 RCT)		
Serious adverse events Follow up: range from 24 months to 60 months							
Placebo Comparator 74 per 1000[1] (7.4%)	**OR 1.23** (0.95 to 1.64) Network estimate	**34 more per 1000** (8 fewer to 87 more)	**OR 0.78** (0.43 to 1.38) Network estimate	**35 fewer per 1000** (54 more to 97 more)	**OR 0.90** (0.54 to 1.51) Network estimate	**15 more per 1000** (71 more to 77 fewer	Interventions reported for harm outcome were chosen based on the interventions included for beneficial outcome. Therefore, we did not include other comparisons in the table.

	⊕⊕⊕⊕ **Low** Confidence in estimate due to Imprecision [2, 3]	⊕⊕◯◯ **Low** Confidence in estimate due to Imprecision [2, 3]	⊕⊕◯◯ **Low** Confidence in estimate due to Imprecision [2, 3]	
Rank 4 (2 to 7)	**Rank** 2 (1 to 9)	**Rank** 8 (3 to 10)	**Rank** 4 (2 to 7)	
	Based on 3,964 participants (3 RCT)	Based on 12,098 participants (1 RCT)	Based on 714 participants (1 RCT)	

NMA-SoF table definitions

Lines in the network graphic represent direct comparisons

Estimates are reported as odds ratio. CrI: credible interval. Results are expressed in credible intervals as opposed to the confidence intervals (9) since a Bayesian analysis has been conducted.

The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; OR: Odds ratio;

GRADE Working Group grades of evidence

High quality: Further research is very unlikely to change our confidence in the estimate of effect.

Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

Very low quality: We are very uncertain about the estimate.

FOOTNOTES

[1] Baseline risks (assumed control risk) obtained from the National Cancer Institute pooling project

[2] Very serious imprecision since 95% CrI crosses unity, and with wide credible intervals suggesting high possibility of harm.

[3] Conceptually, there was no significant intransitivity, with comparable distribution of plausible effect modifiers across trials of different chemopreventive agents.

[4] Ranking is shown as median rank (rank 1-10) and 95% credible intervals.

APPENDIXES

Appendix A. Protocol of user testing.

Developing Summary of Findings Tables in Network Meta-analysis: protocol of a user testing survey

INTRODUCTION

The development of the NMA-SoF tables will apply the user-testing methodology. The starting point for the user testing is one preliminary NMA-SoF table example, which was developed by the GRADE Working Group in 2013-2014 through two brainstorm-meeting sessions. This table will be used for the current user-testing project.

METHODS AND ANALYSIS

Pilot test

The first phase (pilot test) of this qualitative study will consist of developing four interviews to potential users of NMA-SoF tables using an existing NMA-SoF table example. The purpose of this pilot test is to evaluate the interview format that will be applied during each user-testing interview. Participants for these interviews will be 1) systematic review authors, 2) methodologists, or 3) clinicians. Definitions of each of these groups are provided in the section about participant. After this phase, modifications on the interview format will be applied if

necessary. We expect to have an accurate interview format to use in different interviews during the user-testing phase.

User-testing

User testing (or usability testing) is a method of formative evaluation "with the goal of learning about the design from a *user's perspective* to improve its next iteration" where products are tested by actual users, as opposed to developers or experts (1). It can be described as "an applied form of experimentation used by developers to test whether the product they develop is usable by the intended user population to achieve their tasks" (2). User testing is carried out in a controlled laboratory-like setting and is usually scenario based. User testing is commonly employed during the development of web sites and interactive systems to help answer the question: "Does this particular design function as intended?" as well as to indicate what areas need improvement (2, 3).

We planned to develop *rounds* of interviews for the user testing. We defined *rounds* as a group of ten interviews using the same NMA-SoF table format with different participants. After each *round*, the NMA-SoF table will be modified based on the answers provided by the participants, and the feedback provided by the advisory group. The new version of the NMA-SoF table will be presented to the participants during each new *round*. We are intending to complete between one to three *rounds* of interviews. During the first *rounds* will discuss general components that should be considered to be part of the NMA-SoF table such as

the geometry of the NMA, certainty of the evidence, rank probabilities, and the plain language summary section ("what happens" column from the last version of SoF tables). In other *rounds* we will assess particular topics related to the presentation of specific components including both binary and continuous outcomes (see table 1 and 2).

Interviews

Each interview will be conducted using a semi-structured format, in order to explore the overall experience with the NMA SoF-table. Subsequently, thoughts on the seven facets from the honeycomb model such as the usefulness, usability, credibility, desirability, and value of each NMA-SoF table element will be probed. All interviews will be recorded and will take approximately one hour. Participants will be encouraged to articulate their thoughts regarding the information presented to them in one of the example NMA-SoF tables. The semi-structured interview will be divided into three sections: 1) background information of the participants, 2) participant´s reaction to the NMA-SoF table involving first impression, and 3) an overall evaluation. Tests will end with a few questions about how the testing could be improved.

Advisory group

An advisory group composed of seven to ten researchers with experience in developing NMA within the last year will be part of this group. The members of

this group will be consulted after each user-testing round to discuss relevant issues that need to be addressed through individual or group meetings.

Participants

Participants from three target populations will be included in the user testing: 1) systematic review authors, 2) methodologists, and 3) clinicians or healthcare-related professionals. We will consider a *systematic review author* as participants who have published a systematic review in the last couple years using GRADE approach, who have not yet been involved in NMA as an author but who have read one NMA report in last year. To be considered a *methodologist*, participants should declare dedicating more than 70% of their time to conducting MA or NMA research. Different methodologists could be considered such as, epidemiologists, clinical epidemiologists, statisticians, etc. Since research in NMA is something new, we expect that methodologists will be NMA (potential) authors .We will identify these participants from the GRADE Working Group (NMA project group) since they have self-identified with interest in methods of NMA. *Clinicians* will be considered potential users of NMA. We will classify *clinicians* as participants if they report at least 50% of their total time dedicated to clinical practice. At the end of this user testing we are expecting to collect information from non-NMA authors (*systematic review authors*), NMA (potential) authors (*methodologists*), and potential users of NMA (*clinicians*).

Setting and recruitment

Participants will be recruited from workshops, conferences, and other research events. All these events could be part of different Clinical Epidemiology and Evidence-Based Medicine congresses such as GRADE working group meetings, Cochrane Colloquium, Cochrane Congress, Guideline International Network (GIN) congresses, etc. We will also contact participants through networks of the co-authors who interact with researchers, and healthcare professionals.

The potential participants will be contacted by email. After they agree to participate in this survey, and subsequent to informed consent that will be obtained from each participant before beginning the interview, the interviews will be carried out face to face or by any application that specializes in providing video chat and voice calls (e.g. Skype).

Data collection and transcription

The testing will involve a prototype on paper, and the interview will be captured using an audio recorder. Following the interview, a person independent of the research group will transcribe the interview verbatim. After transcription the recording of the interview will be deleted.

Sample size

We expect to recruit at least 20 participants and conduct between one to three rounds of interviews. In total, we estimate recruiting at least five systematic review authors, five methodologists, and ten clinicians. During each round we

expect to have each type of participant represented. Each participant will take part in only one round and none of them will participate in future rounds. If at the end of the round three, we have not recruited the total number of participants required, we will carry out additional rounds until at least 20 participants have completed the user testing. We anticipate that by the third round we will have gathered enough information to fully develop (or saturate) the format of NMA-SoF table (4).

Data Analysis

We will carry out an analysis that will support the pragmatic design goals of improving this particular NMA-SoF table after each round of the user testing. For each interview transcription, two researchers will independently code the information. The interview data will be analyzed using Morville's "honeycomb" model (see figure 2). This model distinguishes between seven separate facets of user experience, including findability, accessibility, usability, usefulness, credibility, desirability, and value (5).

A brief explanation of these terms:

Findability: Can users locate what they are looking for?

Accessibility: Are there physical barriers to actually gaining access?

Usability: How easy and satisfying is this product to use?

Usefulness: Does this product have practical value for this user?

Credibility: Is it trustworthy?

Desirability: Is it something the user wants or has a positive emotional response to?

Value: Does this product advance the mission of the organization behind it?

Our analysis will occur in two phases. The aim of the first phase will be to provide an overview and a prioritization of the problems we will identify in the NMA-SoF table format. An inductive qualitative description analysis approach will be used since there are no previous studies dealing with the phenomenon of providing knowledge about the user experience of health professionals with NMA-SoF tables (6). The unit of analysis considered for this research will be the whole interviews. Using an inductive content analysis, the qualitative data will be organized by open coding, creating categories, and abstraction. Two researches will code the transcripts independently. These codes will then be compared, discussed and merged. The topics will then be rated according to the severity of the problem for the user. We will rate severity in three categories, which were created by Rosenbaum et al (7): high (show-stopper, leads to critical errors or hinders task completion), medium (creates much frustration or slows user down), or low (minor or cosmetic problems).

The second phase of analysis will be done to develop more generalizable issues underlying this study. We will re-sort the findings into the seven user-experience categories from the honeycomb model by re-reading the transcript, checking the context where the problems have come from, and evaluating which of the seven categories best fit each finding. Severity-of-problem ratings from the first phase of analysis will be kept in the second analysis. We will not evaluate accessibility, since user testing methods are not an effective way of gathering data on various aspects of this issue. Based on the results, possible changes to the format will be discussed. We will either agree on a change, or agree to try out multiple solutions in the design of the NMA-SoF tables. We expect to have at least one format example of NMA-SoF table that summarizes the information reported by users at the end of each round. Then at the end of the last round, we will have one NMA-SoF table that can be used by users of NMA. However, we understand that different formats can display different user´s points of view, therefore, after all rounds, it is possible that we will have more than one format of NMA-SoF table.

ETHICS AND DISSEMINATION

We submitted this protocol to the McMaster Research Ethics Board to obtain approval for conducting this project. We will follow the ethical conduct of research involving humans according to the *Tri Council Policy Statement: Ethical Conduct for Research Involving Humans* (8). Audio recordings of the interviews will not be used for any other purpose other transcriptions. Transcribing will take

place, for the most part, the same day as the interviews, after which recordings will be erased. We will gather names, emails or telephone numbers of participants only for the administrative purposes of sending information about the study and making interview appointments. This information will not connect in any way to the data after the interviews and will be deleted from the project folder at the end of the data collection. All participants will be provided with both oral and written information according to the guidelines of the Research Ethics Committees. This will include, among other things, the purpose of the study, how we intend to use the data, and their right to withdraw their agreement to participate at any time, both during the interview or afterwards. Everyone will sign written agreements of consent prior to testing.

CONCLUSION

Limited guidance exists for researchers on how best to present NMAs in an accessible format, especially for non-technical end-users such as policymakers and clinicians. According to Sullivan, et al (9), further developing studies for researchers on how to best present NMA-SoF tables to users of NMA should be considered in light of: 1) policymakers' preferences for evidence syntheses, 2) common principles related to the presentation of traditional meta-analysis, 3) How researchers currently present NMA, and 4) common limitations associated with conducting NMA. The optimal presentation of NMA should be considered as

only one approach to build capacity in the field of NMA and should be applied in concert with educational initiatives such as tutorials and workshops.

REFERENCES

1. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, 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 in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2011;14(4):417-28.

2. Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. Journal of clinical epidemiology. 2010;63(6):620-6.

3. Kuniavsky M. Observing the User Experience: A Practitioner's Guide to User Research. 1 ed. Kaufmann M, editor: Morgan Kaufmann; 2003.

4. Creswell JW. Five quialitative approached to inquiry. In: Creswell J, editor. Quality inquiry & Research Design. 3rd ed. California: SAGE; 2013. p. 69-110.

5. Morville P. User Experience Design: Honeycomb model. Available from: http://www.semanticstudios.com/publications/semantics/000029.php.

6. Elo S, Kyngas H. The qualitative content analysis process. Journal of advanced nursing. 2008;62(1):107-15.

7. Rosenbaum SE, Glenton C, Cracknell J. User experiences of evidence-based online resources for health professionals: user testing of The Cochrane Library. BMC medical informatics and decision making. 2008;8:34.

8. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. Bmj. 2011;342:d549.

9. Sullivan SM, Coyle D, Wells G. What guidance are researchers given on how to present network meta-analyses to end-users such as policymakers and clinicians? A systematic review. PloS one. 2014;9(12):e113277.

Appendix B. First NMA-SoF table formats

Alternative 1

Which oral anticoagulant should be used in patients with atrial fibrillation

Patient or population: patients with atrial fibrillation
Settings: primary care, community, outpatient
Interventions: Dabigatran, Rivaroxaban, and Apixaban
Comparison: Warfarin

Outcomes	Effects and confidence in the effects. Main comparator is warfarin unless specifically mentioned							Comments
	Dabigatran (150mg)		Rivaroxaban		Apixaban		Warfarin	
Death	**OR 0.89** (0.78 to 1.01)	**11 fewer per 1000 (22 fewer to 1 more)**	**OR 0.93** (0.83 to 1.04)	**7 fewer per 1000 (17 fewer to 4 more)**	**OR 0.90** (0.80 to 1.00) **OR 1.04** (0.89 to 1.23) compared to Rivaroxaban	**11 fewer per 1000 (22 fewer to 1 more)** **1 more per 1000 (4 fewer to 6 more)** compared with Rivaroxaban with moderate confidence in estimate)	**100 per 1000 (10%)**	**None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table.**
	⊕⊕⊕◯**moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 14143 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 18201 patients (1 study)			
Stroke or systemic embolism	**OR 0.65** (0.52 to 0.81) **OR 1.35** (1.03,1.79) compared to Rivaroxaban	**11 fewer per 1000 (6 fewer to 14 fewer)** **13 more per 1000 (1 more to 30 more)** compared with Rivaroxaban with	**OR 0.88** (0.74 to 1.04)	**3 fewer per 1000 (1 more to 7 fewer)**	**OR 0.80** (0.66 to 0.95) **OR 1.11** (0.87 to 1.42) compared to Rivaroxaban	**6 fewer per 1000 (1 fewer to 16 fewer)** **3 more per 1000 (3 fewer to 9 more)** compared with Rivaroxaban with	**30 per 1000 (3%)**	

		moderate confidence in estimate				moderate confidence in estimate	
	⊕⊕⊕◯**moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 14143 patients (1 study)		
Major Bleeding	**OR 0.94** (0.82 to 1.08)	**1 fewer per 1000 (3 fewer to 1 more)**	**OR 1.03** (0.89 to 1.190	**0 more per 1000 (2 fewer 3 more)**	**OR 0.70** (0.61 to 0.81)	**5 fewer per 1000 (3 fewer to 6 fewer)**	**16 per 1000 (1.6 %)**
	OR 1.10 (0.9 to 1.35) compared to Rivaroxaban	**6 more per 1000 (2 more to 44 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 1.48** (1.21 to 1.82) compared to Rivaroxaban	**8 more per 1000 (3 more to 13 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Intracranial Bleeding	**OR 0.42** (0.28 to 0.60)	**9 fewer per 1000 (6 fewer to 11 fewer)**	**OR 0.66** (0.47 to 0.92)	**5 fewer per 1000 (1 fewer to 8 fewer)**	**OR 0.42** (0.30 to 0.58)	**9 fewer per 1000 (6 fewer to 10 fewer)**	**14.9 per 1000 (1.49 %)**
	OR 1.58 (0.95 to 2.66) compared to Rivaroxaban	**5 more per 1000 (0 more to 6 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 1.56** (0.97 to 2.5) compared to Rivaroxaban	**6 more per 1000 (0 fewer to 15 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Major GI Bleeding	**OR 1.45** (1.14 to 1.86)	**9 more per 1000 (3 more to 18 more)**	**OR 1.61** (1.30 to 1.99)	**13 more per 1000 (6 more to 21 more)**	**OR 0.88** (0.68 to 1.15)	**3 fewer per 1000 (7 fewer to 3 more)**	**20.9 per 1000 (2.09%)**
	OR 1.11 (0.8 to 1.53)	**4 more per 1000**			**OR 1.83** (1.30 to 2.57)	**28 more per 1000**	

	compared to Rivaroxaban	**(1 more to 25 more)** compared with Rivaroxaban with moderate confidence in estimate			compared to Rivaroxaban	**(10 more to 53 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Myocardial Infarction	**OR 1.29** (0.96 to 1.75)	**4 more per 1000 (1 fewer to 9 more)**	**OR 0.80** (0.62 to 1.05)	**2 fewer per 1000 (5 fewer to 1 more)**	**OR 0.88** (0.66 to 1.17)	**2 fewer per 1000 (4 fewer to 2 more)**	**12.5 per 1000 (1.25%)**
	OR 0.63 (0.42 to 0.93) compared to Rivaroxaban	**5 fewer per 1000 (0 more to 11 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 0.92** (0.62 to 1.35)	**1 fewer per 1000 (4 fewer to 4 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; OR: Odds ratio;

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] FOOTNOTES

Alternative 2

Which oral anticoagulant should be used in patients with atrial fibrillation

Patient or population: patients with atrial fibrillation
Settings: primary care, community, outpatient
Interventions: Dabigatran (150mg), Rivaroxaban, and Apixaban
Comparison: Warfarin and/or Rivaroxaban

Outcome		Effects and confidence in the estimate of effects						Comments
		Rivaroxaban		Dabigatran		Apixaban		
Death								
Warfarin Comparator	**100 per 1000 (10%)**	**OR 0.93** (0.83 to 1.04)	**7 fewer per 1000** (17 fewer to 4 more)	**OR 0.89** (0.78 to 1.01)	**11 fewer per 1000** (22 fewer to 1 more)	**OR 0.90** (0.80 to 1.00)	**11 fewer per 1000** (22 fewer to 1 more)	None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table. The quality of evidence needs to be evaluated separately for the comparisons if there are differences for it.
Rivaroxaban Comparator	**29 per 1000 (2.9%)**					**OR 1.04** (0.89 to 1.23)	**1 more per 1000** (4 fewer to 6 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on **14143 patients (1 study)**		⊕⊕⊕ **Moderate** confidence in estimate **due to risk of bias** based on **12098 patients (1 study)**		⊕⊕⊕⊕ **High** confidence in estimate based on 18201 patients (1 study)		
Stroke or systemic embolism								
Warfarin Comparator	**30 per 1000 (3.0%)**	**OR 0.88** (0.74 to 1.04)	**3 fewer per 1000** (1 more to 7 fewer)	**OR 0.65** (0.52 to 0.81)	**11 fewer per 1000** (6 fewer to 14 fewer)	**OR 0.80** (0.66 to 0.95)	**6 fewer per 1000** (1 fewer to 16 fewer)	
Rivaroxaban Comparator	**38 per 1000 (3.8%)**			**OR 1.35** (1.03,1.79)	**13 more per 1000** (1 more to 30 more)	**OR 1.11** (0.87 to 1.42)	**3 more per 1000** (3 fewer to 9 more)	

		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)	
Major Bleeding							
Warfarin Comparator	**16 per 1000 (1.6%)**	**OR 1.03** (0.89 to 1.190)	**0 more per 1000** (2 fewer 3 more)	**OR 0.94** (0.82 to 1.08)	**1 fewer per 1000** (3 fewer to 1 more)	**OR 0.70** (0.61 to 0.81)	**5 fewer per 1000** (3 fewer to 6 fewer)
Rivaroxaban Comparator	**56 per 1000 (5.6%)**			**OR 1.10** (0.9 to 1.35)	**6 more per 1000** (2 more to 44 more)	**OR 1.48** (1.21 to 1.82)	**8 more per 1000** (3 more to 13 more)
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)	
Intracranial Bleeding							
Warfarin Comparator	**14.9 per 1000 (1.49%)**	**OR 0.66** (0.47 to 0.92)	**5 fewer per 1000** (1 fewer to 8 fewer)	**OR 0.42** (0.28 to 0.60)	**9 fewer per 1000** (6 fewer to 11 fewer)	**OR 0.42** (0.30 to 0.58)	**9 fewer per 1000** (6 fewer to 10 fewer)
Rivaroxaban Comparator	**8 per 1000 (0.8%)**			**OR 1.58** (0.95 to 2.66)	**5 more per 1000** (0 more to 6 more)	**OR 1.56** (0.97 to 2.5)	**6 more per 1000** (0 fewer to 15 more)
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)	
Major GI Bleeding							
Warfarin Comparator	**20.9 per 1000 (2.09%)**	**OR 1.61** (1.30 to 1.99)	**13 more per 1000** (6 more to 21 more)	**OR 1.45** (1.14 to 1.86)	**9 more per 1000** (3 more to 18 more)	**OR 0.88** (0.68 to 1.15)	**3 fewer per 1000** (7 fewer to 3 more)

Rivaroxaban Comparator	**32 per 1000 (3.2%)**			**OR 1.11** (0.8 to 1.53)	**4 more per 1000** (1 more to 25 more)	**OR 1.83** (1.30 to 2.57)	**28 more per 1000** (10 more to 53 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		
Myocardial Infarction								
Warfarin Comparator	**12.5 per 1000 (1.25%)**	**OR 0.80** (0.62 to 1.05)	**2 fewer per 1000** (5 fewer to 1 more)	**OR 1.29** (0.96 to 1.75)	**4 more per 1000** (1 fewer to 9 more)	**OR 0.88** (0.66 to 1.17)	**2 fewer per 1000** (4 fewer to 2 more)	
Rivaroxaban Comparator	**14.3 per 1000 (1.43%)**			**OR 0.63** (0.42 to 0.93)	**5 fewer per 1000** (0 more to 11 more)	**OR 0.92** (0.62 to 1.35)	**1 fewer per 1000** (4 fewer to 4 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; OR: Odds ratio;

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] FOOTNOTES

Appendix C. NMA-SoF table format developed for conducting user testing rounds. Format 1.

Estimates of effects and confidence ratings for comparison antithrombotic agents for the prevention of stroke and major bleeding in patients with atrial fibrillation

Outcome: Stroke or systemic embolism

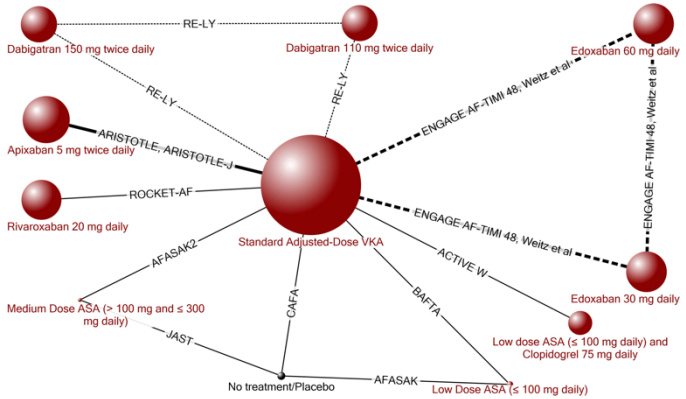
Patient or population: patients with non-valvular atrial fibrillation requiring anticoagulation

Settings: primary care, community, outpatient

Interventions: Apixaban, dabigatran, edoxaban, rivaroxaban, acetylsalicylic acid (ASA), ASA plus clopidogrel

Outcomes: Prevention of stroke or systemic embolism and major bleeding

Comparisons: Warfarin [vitamin K antagonists (VKA)]



Direct evidence ——
Indirect evidence - - -

Graphic: Evidence network for all-cause stroke or systemic embolism

Nº studies: 12 RCTs **Participants**: 82396	**Comparator (relative effects in upper third, risk difference in middle third, certainty of evidence in low third)**									**Comment**
	Warfarin * (17 per 1000) **	**Dabigatran** [a]	**Apixaban** [b]	**Edoxaban** [c]	**Rivaroxaban**	**Dabigatran** [d]	**Edoxaban** [e]	**Medium dose ASA** [f]	**Low dose ASA** [g]	
No treatment / Placebo	**OR 1.53** (0.9 to 2.63) **9 more per 1000** (2 fewer to 26 more) ⊕⊕⊕O **Moderate** [1]	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table. The certainty of evidence needs to be evaluated separately for the comparisons if there are differences for it.
Dabigatran [a]	**OR 0.66** (0.53 to 0.82)	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	

	6 fewer per 1000 (8 fewer to 3 fewer) ⊕⊕⊕O **Moderate** [1]									
Apixaban [b]	**OR 0.78** (0.65 to 0.94) **4 fewer per 1000** (6 fewer to 1 fewer) ⊕⊕⊕O **Moderate** [1]	**OR 1.19** (0.89 to 1.58)	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	
Endoxaban [c]	**OR 0.87** (0.74 to 1.02) **3 fewer per 1000** (5 fewer to 1 more) ⊕⊕⊕O **Moderate** [1]	**OR 1.31** (1.00 to 1.73)	**OR 1.11** (0.87 to 1.41)	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	
Rivaroxaban	**OR 0.88** (0.74 to 1.04) **3 fewer per 1000** (5 fewer to 1 more) ⊕⊕⊕O **Moderate** [1]	**OR 1.33** (1.01 to 1.76)	**OR 1.12** (0.87 to 1.43)	**OR 1.01** (0.8 to 1.28)	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	
Dabigatran [d]	**OR 0.91** (0.74 to 1.12) **2 fewer per 1000** (5 fewer to 2 more) ⊕⊕⊕O **Moderate** [1]	**OR 1.38** (1.11 to 1.74)	**OR 1.17** (0.88 to 1.53)	**OR 1.05** (0.8 to 1.37)	**OR 1.04** (0.8 to 1.36)	Not applicable	Not applicable	Not applicable	Not applicable	
Endoxaban [e]	**OR 1.14** (0.98 to 1.32) **3 more per 1000** (1 fewer to 6 more) ⊕⊕⊕O **Moderate** [1]	**OR 1.73** (1.32 to 2.26)	**OR 1.45** (1.15 to 1.84)	**OR 1.31** (1.13 to 1.54)	**OR 1.3** (1.04 to 1.63)	**OR 1.25** (0.97 to 1.61)	Not applicable	Not applicable	Not applicable	**Explanatory Footnotes** [a] 150 mg twice daily [b] 5 mg twice daily c 60 mg daily [d] 110 mg twice daily [e] 30 mg daily f >100 mg and ≤300 mg daily
	OR 1.35 (0.74 to 2.47)	**OR 2.04** (1.08 to 3.9)	**OR 1.72** (0.92 to 3.24)	**OR 1.56** (0.84 to 2.91)	**OR 1.54** (0.83 to 2.88)	**OR 1.48** (0.79 to 2.8)	**OR 1.18** (0.64 to 2.21)	Not applicable	Not applicable	

Medium dose ASA [f]	**6 more per 1000** (5 fewer to 24 more) ⊕⊕⊕O **Moderate** [1]									[g] ≤100 mg daily [h] ≤ 100 mg daily [i] 75 mg daily [1] Heterogeneity
Low dose ASA [g]	**OR 1.87** (1.26 to 2.8) **14 more per 1000** (5 more to 29 more) ⊕⊕⊕O **Moderate** [1]	**OR 2.84** (1.81 to 4.5)	**OR 2.39** (1.55 to 3.72)	**OR 2.16** (1.42 to 3.34)	**OR 2.14** (1.4 to 3.31)	**OR 2.05** (1.32 to 3.23)	**OR 1.64** (1.08 to 2.53)	**OR 1.39** (0.74 to 2.61)	Not applicable	
Low dose ASA [h] **& clopidogrel** [i]	**OR 1.93** (1.42 to 2.64) **15 more per 1000** (7 more to 26 more) ⊕⊕⊕O **Moderate** [1]	**OR 2.93** (2.01 to 4.3)	**OR 2.46** (1.73 to 3.54)	**OR 2.23** (1.58 to 3.17)	**OR 2.22** (1.55 to 3.14)	**OR 2.11** (1.46 to 3.07)	**OR 1.69** (1.21 to 2.4)	**OR 1.43** (0.73 to 2.81)	**OR 1.03** (0.62 to 1.7)	

Appendix D. NMA-SoF table format developed for conducting user-testing rounds. Format 2.

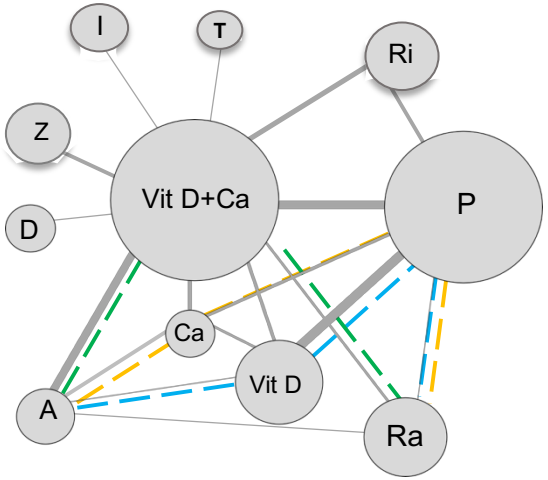
Patient or population: patients with established or at risk for osteoporosis

Settings: primary care, community, outpatient

Interventions: bisphosphonates (alendronate, risedronate, zoledronate, and ibandronate), PTH 1-34 (teriparatide), SERM such as raloxifene or bazedoxifene, denosumab, and calcium and vitamin D

Outcome: prevent osteoporotic hip fractures

Comparisons: comparison of one or more interventions to each other or to placebo



Graphic: Network of trials evaluating hip fractures

Nº studies: 40 RCTs **Participants:** 139647	**Comparator (relative effects in upper half, certainty of the evidence in lower half)**										**Comments**
	Placebo	Teriparatide [a]	Denosumab [a]	Raloxifen [b]	Zoledronate [a]	Risedronate [b]	Ibandronate	Alendronate	Vitamin D	Vitamin D + Calcium	
Teriparatide [a]	**OR 0.42 (0.10 to 1.82)** ⊕OOO **Very low** [1 2]	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table.
Denosumab [a]	**OR 0.50 (0.27 to 0.86)** ⊕⊕⊕⊕ **High**	**OR 1.17 (0.24 to 5.54)** ⊕⊕ O O **Low** [1 2]	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	
Raloxifen [b]	**OR 0.87**	**OR 2.05**	**OR 1.76**								

	(0.63 to 1.22) ⊕⊕⊕O Moderate [1]	(0.47 to 9.47) ⊕OOO Very low	(0.95 to 3.41) ⊕⊕ O O Low	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	The certainty of evidence needs to be evaluated separately for the comparisons if there are differences for it.
Zoledronate [a]	OR 0.50 (0.33 to 0.74) ⊕⊕⊕⊕ High	OR 1.18 (0.26 to 5.30) ⊕OOO Very low	OR 1.02 (0.54 to 1.93) ⊕⊕ O O Low	OR 0.57 (0.35 to 0.93) ⊕⊕ O O Low	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	
Risedronate [b]	OR 0.48 (0.31 to 0.66) ⊕⊕⊕O Moderate [3]	OR 1.12 (0.25 to 4.98) ⊕OOO Very low	OR 0.96 (0.50 to 1.78) ⊕OOO Very low	OR 0.55 (0.31 to 0.84) ⊕OOO Very low	OR 0.96 (0.56 to 1.49) ⊕⊕ O O Low	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	
Ibandronate [a]	OR 0.49 (0.21 to 1.20) ⊕OOO Very low [1 2]	OR 1.11 (0.22 to 6.42) ⊕OOO Very low	OR 0.98 (0.36 to 2.79) ⊕⊕ O O Low	OR 0.55 (0.23 to 1.42) ⊕OOO Very low	OR 0.97 (0.39 to 2.55) ⊕OOO Very low	OR 1.02 (0.43 to 2.66) ⊕OOO Very low	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	**Explanatory Footnotes** [a] With vitamin D and calcium [b] With or without vitamin D and calcium [c] With or without placebo [1] Imprecision [2] Contributing direct evidence of low or very low quality [3] Limitations (risk of bias) [4] Contributing direct evidence of certainty quality
Alendronate [b]	OR 0.45 (0.27 to 0.68) ⊕⊕⊕O Moderate [4]	OR 1.02 (0.24 to 4.82) ⊕OOO Very low	OR 0.90 (0.45 to 1.78) ⊕⊕ O O Low	OR 0.51 (0.29 to 0.87) ⊕⊕⊕O Moderate	OR 0.90 (0.51 to 1.51) ⊕⊕ O O Low	OR 0.93 (0.54 to 1.62) ⊕OOO Very low	OR 0.92 (0.34 to 2.32) ⊕⊕ O O Low	No comparisons were found reported	No comparisons were found reported	No comparisons were found reported	
Vitamin D [c]	OR 1.13 (0.94 to 1.13) ⊕⊕ O O Low [1 3 4]	OR 2.67 (0.63 to 11.97) ⊕OOO Very low	OR 2.28 (1.28 to 4.16) ⊕⊕⊕O Moderate	OR 1.30 (0.89 to 1.86) ⊕⊕ O O Low	OR 2.26 (1.50 to 3.42) ⊕⊕ O O Low	OR 2.35 (1.63 to 3.76) ⊕OOO Very low	OR 2.32 (0.92 to 5.54) ⊕OOO Very low	OR 2.54 (1.63 to 4.16) ⊕⊕⊕O Moderate	No comparisons were found reported	No comparisons were found reported	
Vitamin D plus Calcium [c]	OR 0.81 (0.68 to 0.96) ⊕⊕⊕O Moderate [4]	OR 1.92 (0.45 to 8.42) ⊕⊕ O O Low	OR 1.64 (0.97 to 2.87) ⊕⊕⊕O Moderate	OR 0.94 (0.66 to 1.31) ⊕⊕⊕O Moderate	OR 1.63 (1.16 to 2.30) ⊕⊕⊕⊕ High	OR 1.69 (1.27 to 2.54) ⊕OOO Very low	OR 1.69 (0.69 to 3.84) ⊕⊕ O O Low	OR 1.82 (1.24 to 2.90) ⊕⊕⊕O Moderate	OR 0.72 (0.57 to 0.91) ⊕⊕ O O Low	No comparisons were found reported	
Calcium [c]	OR 1.14 (0.82 to 1.59) ⊕⊕⊕O Moderate [4]	OR 2.69 (0.63 to 12.23) ⊕OOO Very low	OR 2.33 (1.25 to 4.40) ⊕⊕⊕O Moderate	OR 1.31 (0.83 to 2.06) ⊕OOO Very low	OR 2.29 (1.44 to 3.66) ⊕⊕ O O Low	OR 2.39 (1.56 to 4.04) ⊕OOO Very low	OR 2.36 (0.92 to 5.87) ⊕OOO Very low	OR 2.56 (1.57 to 4.34) ⊕⊕⊕O Moderate	OR 1.01 (0.72 to 1.44) ⊕⊕ O O Low	OR 1.40 (1.03 to 1.95) ⊕⊕ O O Low	

CI: Confidence interval; OR: Odds ratio.

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

Appendix E. NMA-SoF table format. Format 3.

Estimates of effects, credible intervals, and certainty of the evidence for comparison fluid resucitation in patients with sepsis

Bayesian NMA-SoF table

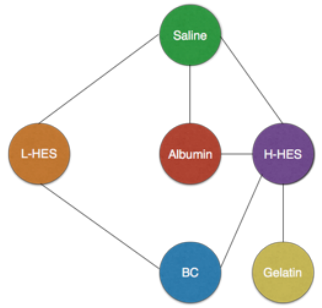
Patient or population: Critically ill patients with severe sepsis or septic shock

Interventions: Balanced crystalloid (BC), Albumin, High-molecular-weight hydroxyethyl starch (H-HES), Saline solution, Gelatin

Comparator (reference): Low-molecular weight hydroxyethyl starch (L- HES)

Outcome: Mortality; range of follow up between 24 hours to 90 days

Setting(s): Inpatient



Geometry of the Network*

Total studies: 6 RCT **Total Participants**: 8308		**Relative effect** (95% CrI)	**Anticipated absolute effect** (95% CrI)			**Certainty of the evidence**	**Ranking** (95% CrI)	**Interpretation of Findings**
			With L-HES	***With* intervention**	**Difference**			
●	Balanced crystalloid (Direct evidence; 2 RCT; 846 participants)	**0.75** (0.58 to 0.97)	180 per 1000[1]	141 per 1000	39 per 1000 fewer (from 67 fewer to 5 fewer)	⨁⨁⨁◯ **Moderate** Due to Indirectness[2]	**2.00** (1.00 to 4.00)	-
●	Albumin (No direct evidence, Indirect evidence only)	**0.79** (0.59 to 1.06)	180 per 1000[1]	148 per 1000	32 per 1000 fewer (from 65 fewer to 88 more)	⨁⨁◯◯ **Low** Due to Imprecision[3], and Indirectness[4]	**2.00** (1.00 to 5.00)	-

●	H-HES (No direct evidence, Indirect evidence only)	**0.91** (0.63 to 1.33)	180 per 1000[1]	164 per 1000	16 per 1000 fewer (from 59 fewer to 46 more)	⊕⊕◯◯ **Low** Due to Imprecision[3], and Indirectness[4]	**4.00** (2.00 to 6.00)	-
●	Saline solution (Direct evidence; 4 RCT; 7642 participants)	**1.04** (0.87 to 1.25)	180 per 1000[1]	186 per 1000	6 per 1000 more (from 20 fewer to 35 more)	⊕⊕⊕◯ **Moderate** Due to Imprecision[4], Indirectness[6], and Inconsistency[5]	**4.00** (1.00 to 6.00)	-
●	Gelatin (No direct evidence, Indirect evidence only)	**1.00** (0.44 to 2.21)	180 per 1000[1]	180 per 1000	0 per 1000 fewer (from 92 fewer to 146 more)	⊕◯◯◯ **Very Low** Due to Imprecision[3], and Indirectness[2]	**5.00** (3.00 to 6.00)	-
●	L-HES	Reference Comparator	No estimable	No estimable	No estimable	Reference Comparator	**5.00** (1.00 to 6.00)	-

NMA-SoF table definitions

* Solid lines represent direct comparisons

** Network Meta-analysis (1) estimates are reported as odds ratio. CrI: credible interval. Results are expressed in credible intervals as opposed to the confidence intervals (7) since a Bayesian analysis has been conducted.

*** Anticipated absolute effect. Anticipated absolute effect compares two risks by calculating the difference between the risk of the intervention group with the risk of the control group.

**** Median and credible intervals are presented. Rank statistics is defined as the probabilities that a treatment out of *n* treatments in a network meta-analysis is the best, the second, the third and so on until the least effective treatment.

† Information is reported from studies included in the network meta-analysis for the comparison displays.

GRADE Working Group grades of evidence (or certainty in the evidence)

High quality: We are very confident that the true effect lies close to that of the estimate of the effect

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect

Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Explanatory Footnotes

[1] Mortality is reported from a large randomized control trail where critically ill patients admitted to an intensive care unit (ICU) required fluid resuscitation with hydroxyethyl starch (HES).

[2] Serious indirectness. The indirect evidence for this comparison goes through a second order loop via heavy starch and saline.

[3] Serious imprecision. Due to wide confidence intervals in the indirect estimate.

[4] Serious indirectness. The indirect evidence for this comparison goes through a first order loop via saline and saline vs. light starch.

[5] Serious inconsistency. Due to there was significant heterogeneity in the direct comparison of light starch vs. balanced crystalloid.

[6] Serious indirectness. The indirect evidence for this comparison goes through a second order loop via balance crystalloid and heavy starch.

Appendix F. NMA-SoF table template for continuous outcomes

Bayesian NMA-SoF table

BENEFITS

Estimates of effects, credible intervals, and certainty of the evidence for improving Peak Urinary Flow Rate in patients with Benign Prostatic Hyperplasia

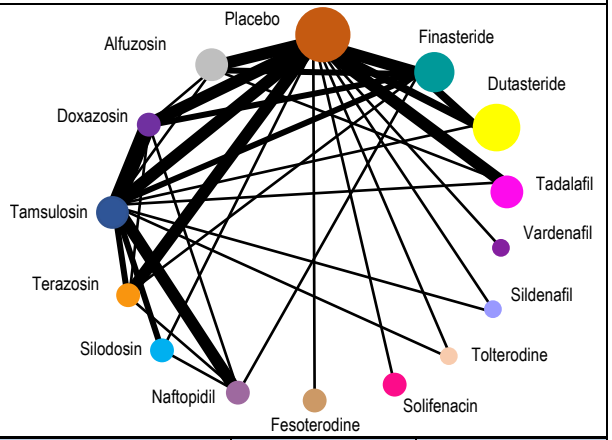
Patient or population: Patients with benign prostatic hyperplasia

Interventions: Doxazosin, Dutasteride, Terazosin, Alfuzosin, Tamsulosin, Naftopidil, Silodosin, Finasteride, Vardenafil, Sildenafil, Tadalafil, Tolterodine, Solifenacin

Comparator (reference): Placebo

Outcome: Improving Peak Urinary Flow Rate; range of follow up between 2 weeks to 4.5 years

Setting: Outpatient



Total studies: 105 RCT **Total Participants**: 45955		**Relative effect (95% CI)**	**Anticipated absolute effect (95% CI)****		**Certainty of evidence**	**Ranking*** (95% CI)**	**Interpretation of Findings**
			***Without* intervention**	***With* intervention**			
●	Doxazosin (7 RCT; 1983 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.95 higher (1.61 higher to 2.30 higher)	⨁⨁⨁◯ **Moderate** Due to Risk of bias[1]	1.12 (1.00 to 2.00)	-
●	Dutasteride (4 RCT; 9277 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.43 higher (1.05 higher to 1.82 higher)	⨁⨁⨁◯ **Moderate** Due to Risk of bias[1]	2.80 (2.00 to 6.00)	-

	Terazosin (6 RCT; 1185 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.21 higher (0.74 higher to 1.66 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	4.35 (2.00 to 9.00)	-
	Alfuzosin (9 RCT; 3195 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.07 higher (0.78 higher to 1.38 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	5.50 (3.00 to 9.00)	-
	Tamsulosin (9 RCT; 4596 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.07 higher (0.83 higher to 1.32 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	5.51 (3.00 to 8.00)	-
	Naftopidil (0 RCT; indirect evidence)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 1.05 higher (0.29 higher to 1.80 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	5.84 (2.00 to 11.00)	-
	Silodosin (2 RCT; 1479 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 0.93 higher (0.45 higher to 1.42 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	6.83 (3.00 to 10.00)	-
	Finasteride (14 RCT; 7064 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 0.85 higher (0.61 higher to 1.10 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	7.73 (5.00 to 10.00)	-
	Vardenafil (1 RCT; 214 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 0.58 higher (-1.18 lower to 2.38 higher)	⊕⊕◯◯ **Low** Due to Risk of bias[1] and Imprecision[2]	8.34 (1.00 to 13.00)	-
	Sildenafil (1 RCT; 366 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 0.52 higher (-0.50 lower to 1.54 higher)	⊕⊕◯◯ **Low** Due to Risk of bias[1] and Imprecision[2]	9.18 (2.00 to 13.00)	-
	Tadalafil (11 RCT; 4875 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was 0.43 higher (0.06 higher to 0.79 higher)	⊕⊕⊕◯ **Moderate** Due to Risk of bias[1]	10.20 (8.00 to 12.00)	-
	Tolterodine (2 RCT; 654 participants)	-	The mean difference in improving PUFR was 0.96	The mean difference in improving PUFR was -0.02 lower (-1.06 lower to 1.03 higher)	⊕⊕◯◯ **Low** Due to Risk of bias[1] and Imprecision[2]	11.61 (6.00 to 14.00)	-

●	Placebo	Reference comparator	No estimable	No estimable	Reference comparator	13.46 (12.00 to 14.00)	-

PUFR: Peak Urinary Flow Rate; MD: Mean Difference

NMA-SoF table definitions

* Lines in the network graphic represent direct comparisons

** Estimates are reported as mean difference and credible interval (CrI). Results are expressed in credible intervals as opposed to the confidence intervals since a Bayesian analysis has been conducted.

*** Ranking and confidence intervals for efficacy outcome are presented. Rank statistics is defined as the probabilities that a treatment out of *n* treatments in a network meta-analysis is the best, the second, the third and so on until the least effective treatment.

GRADE Working Group grades of evidence (or certainty in the evidence)

High quality: We are very confident that the true effect lies close to that of the estimate of the effect

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect

Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Explanatory Footnotes

[1] Serious risk of bias. Most of the studies did not report techniques for randomization (83.9%) and allocation concealment (89.5%). Ninety-seven studies were double-blinded and 9 were single-blinded. The risk of bias from incomplete outcome data was assessed as low in 110 studies.

[2] Serious imprecision since 95% CrI crosses unity, and with wide credible intervals.

CHAPTER 4: VITAMIN D SUPPLEMENTATION IN PRIMARY ALLERGY PREVENTION. A SYSTEMATIC REVIEW OF RANDOMIZED AND NON-RANDOMIZED STUDIES

PREFACE TO CHAPTER 4

Chapter 4: *Vitamin D supplementation in primary allergy prevention. A systematic review of randomized and non-randomized studies* was submitted to Allergy Journal and accepted on June 30, 2017. In this dissertation, we present the final submitted version.

Chapter 4: Vitamin D supplementation in primary allergy prevention. A systematic review of randomized and non-randomized studies.

Authors

Juan José Yepes-Nuñez [1,2], Jan L. Brożek [1,3], Alessandro Fiocchi [4], Ruby Pawankar [5], Carlos Cuello-García [1,6], Yuan Zhang [1], Gian Paolo Morgano [1], Arnav Agarwal [7], Shreyas Gandhi [7], Luigi Terracciano [8], Holger J. Schünemann [1,3]

Affiliations

1. Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario, Canada
2. University of Antioquia, School of Medicine, Medellín, Colombia
3. Department of Medicine, McMaster University, Hamilton, Ontario, Canada
4. Pediatric Hospital Bambino Gesù, Vatican City
5. Department of Pediatrics, Nippon Medical School, Tokyo, Japan

6. Tecnologico de Monterrey School of Medicine and Health Sciences, Monterrey, Mexico
7. Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
8. National Pediatric Healthcare System, Italian Federation of Pediatric Medicine, Territorial Pediatric Primary Care Group, Milan Italy.

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact (HEI), Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON. L8S 4K1 Canada.
holger.schunemann@mcmaster.ca

Conflicts of interest: The following are potential conflict of interests (within last 4 years) for authors of this work. Juan José Yepes-Nuñez declares no competing interest related to this systematic review. Part of scholarship paid by WAO funding received by HJS and JLB for developing the GLADP guidelines. Jan Brozek received research support from WAO for the development of DRACMA and GLAD-P guidelines. Alessandro Fiocchi received research support from Danone, has received travel support for meetings for presentation of DRACMA guidelines in South-East Asia from Danone Malaysia, and support for travel and honorarium for speaking from Ordesa Group. Ruby Pawankar received support

from Danone Asia Pacific for act as a speaker at a Food allergy Asia Pacific symposium 2012. Carlos Cuello-García declares no competing interest related to prebiotics. Part of his PhD scholarship and travel support for meetings is supported by WAO. Yuan Zhang declares no competing interest. Gian Paolo Morgano declares no competing interest. Arnav Agarwal declares no competing interest. Shreyas Gandhi declares no competing interest. Luigi Terracciano; Heinz-Plada Italy Medical consultant for website. Travel support from World Allergy Organization. Holger Schünemann received research support from WAO for the development of DRACMA and GLAD-P guidelines.

ABSTRACT

Background: To date, a systematic review of the evidence regarding the association between Vitamin D and allergic diseases development has not yet been undertaken. **Objective:** To review the efficacy and safety of Vitamin D supplementation when compared to no supplementation in pregnant women, breastfeeding women, infants and children for the prevention of allergies. **Methods:** Three databases were searched through 30 January 2016 including randomized (RCT) and non-randomized studies (NRS). Two reviewers independently extracted data and assessed the certainty in the body of evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. **Results:** Among the 1932 articles identified, one RCT and four NRS were eligible. Very low certainty in the body of evidence across examined studies suggests that Vitamin D supplementation for pregnant women, breastfeeding women and infants may not decrease the risk of developing allergic diseases such as atopic dermatitis (in pregnant women), allergic rhinitis (in pregnant women, and infants), asthma and/or wheezing (in pregnant women, breastfeeding women, and infants), or food allergies (in pregnant women). We found no studies of primary prevention of allergic diseases in children. **Conclusion:** Limited information is available addressing primary prevention of allergic diseases after Vitamin D supplementation and its potential impact remains uncertain.

Keywords: Allergy, GRADE, prevention, systematic review, Vitamin D

Funding information: World Allergy Organization.

HIGHLIGHTS

- To review the benefits and harms of Vitamin D supplementation when compared to no supplementation for the prevention of allergies in RCTs and NRS.
- Risk of bias (RoB) assessment of each individual RCT and NRS was deemed between moderate to very low.
- Implementation of ROBINS-I tool to assess the RoB in individual NRS was comprehensible and allowed us to better understand flaws of NRS.
- Impact of vitamin D supplementation for the primary prevention of allergic diseases is uncertain.

INTRODUCTION

Studies have established that the prevalence of asthma and allergic diseases in westernized countries has increased since the 1970s (1, 2). The hygiene hypothesis and the gut microbiome have gained prominence as explanatory factors for this increase. Other theories have associated Vitamin D with the development of asthma, wheezing, allergic rhinitis, food allergy and atopic dermatitis (3). The postulated mechanisms include decreased levels of Vitamin D at population level which increase susceptibility; genetic links between Vitamin D and the early development of lungs and asthma; and the function of immune cells and the gut microbiome related to Vitamin D (4). Support for these theories comes from a study suggesting that the active form of Vitamin D, calcitriol, modulates immune function in cell culture and animal models (5). Thus, primary prevention of asthma as well as reduction of asthma morbidity, and modulation of the severity of asthma attacks, could be influenced by Vitamin D immunomodulation on the innate and adaptive immune systems (6). Nevertheless, the understanding of the complex role of Vitamin D in immune function remains limited.

Although multiple studies have reported an association between low serum Vitamin D levels and the development of allergic diseases, this association may not be causal, and Vitamin D supplementation may have no role in the primary prevention of allergic diseases (7-10). Furthermore, two recent randomized trials

reported that increased Vitamin D supplementation in pregnant women did not confer protection against allergic diseases in their children (11, 12).

This systematic review informed the recommendations of the World Allergy Organization-McMaster guidelines for the prevention of allergies or Guideline for Allergy Disease Prevention (GLAD-P) (13). We analyzed comparative studies that examined the effects of Vitamin D supplementation versus non-Vitamin D supplementation on population and patient important outcomes in pregnant women, breastfeeding women, infants, and children.

METHODS

Inclusion and Exclusion Criteria

Types of studies

We included randomised control trials and non-randomized studies (NRS), specifically cohort and case-control studies.

Types of participants

Studies must have included one or more of the following groups of participants:

- Healthy pregnant women
- Healthy breastfeeding mothers
- Healthy infants

- Healthy children

We included studies that assessed the use and effects of Vitamin D in any age group, from newborn infants to pre-school, and school age children (up to 9 years of age).

Type of interventions

We included only studies that reported Vitamin D supplementation in isolation, irrespective of formulation (e.g. capsules, drops, suspension, etc.) or dose. We did not include studies that reported Vitamin D supplementation combined with other vitamins or nutrients, studies that compared different doses of Vitamin D supplementation and studies that reported Vitamin D intake using food-frequency questionnaires.

Type of outcomes

The outcomes were determined following a detailed guideline development process by the WAO guideline panel members (14). The following outcomes were deemed critical to the decision of whether or not to use Vitamin D for primary prevention of allergies:

- Asthma or wheezing
- Allergic rhinitis

- Eczema (including atopic dermatitis, atopic eczema, or any sort of eczema)
- Food allergy

Other outcomes were included for the panel, which were not related directly to prevention of allergies but could be critical for balancing the benefits against potential downsides vitamin D supplementation:

- Nutritional status (including weight at birth or weight gained during follow-up)
- Rickets
- Any adverse events
- Severe adverse events

Search methods

MEDLINE, EMBASE, and The Cochrane Central Register of Controlled Trials were searched from inception to January 30 2016. Restrictions on language of publication were not imposed. Search strategies for all databases are reported in the Appendix A. We checked reference lists of reviewed articles and contacted clinical experts in the speciality for additional references. We also searched The U.S. National Institutes of Health Ongoing Trials Register (www.clinicaltrials.gov) for ongoing studies.

Study Selection and Data Extraction

Six independent evaluators (JY, YZ, CC, GM, AA, SG) screened the titles and abstracts obtained through the electronic searches, to identify studies to include in the review. If a study was deemed potentially relevant we obtained the full-text and two authors (JY, CC) assessed it to make a final eligibility decision. In the case of any disagreement, it would have been discussed with a third reviewer (JLB). If additional information or clarification was needed about a study, we contacted the authors of the relevant article.

Two independent evaluators (JY, CC) read all reports of eligible studies in detail and summarised the pertinent details in a standard data extraction sheet (type of study; methodology; characteristics of participants, results and outcome measurements; as well as an evaluation of the risk of bias or study limitations). We discussed any disagreements and aimed to reach agreement by consensus with a third reviewer if necessary (JLB).

Certainty in the body of evidence (confidence in the estimates of effects or quality of evidence)

Assessment of risk of bias in included studies

Two review authors (JY, CC) independently assessed risk of bias for each randomised control trial and non-randomized studies using the criteria outlined in the Cochrane Handbook for Systematic Reviews of Intervention (15), and version

1.0 of the Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI), now called ROBINS-I (16), respectively.

Data analysis

Measure of treatment effect

We calculated relative risk (RR) and 95% confidence intervals (95%CI) for dichotomous outcomes and mean difference with associated 95% CI for continuous outcomes. For NRS, we reported the measure of the relative effect reported by authors in their studies, odds ratio (OR) or relative risk (RR).

Dealing with missing data

We contacted authors where details about study design or descriptive statistics for outcomes were not presented in original papers. We did not impute data.

Assessment of statistical heterogeneity and inconsistency

We assessed inconsistency between studies by visual inspection of forest plots, confidence intervals (CI) and its minimal or no overlap, the Q statistic (with a P value ≤0.05 as a suggestion of important statistical heterogeneity), and the I^2 value. When inconsistency was found we planned to explain it by pre-specified differences in study populations, dosages, presentation, or type of Vitamin D used.

Assessment of reporting biases

We planned to create a funnel plot to assess publication bias by visual inspection if ten or more studies were available.

Data Synthesis

We assessed the treatment effect through mean difference for continuous outcomes, and RR for dichotomous outcomes for individual studies, and used a random-effects model to pool study data. We presented all measures with 95% CI. We carried out all statistical analyses using Review Manager 5.3 (17).

Subgroup analysis

We considered four main groups for subgroup analysis, one for each clinical question based on the population where the vitamin was administered: pregnant women, breastfeeding women, infants, and children.

Sensitivity analysis

We planned analyses to determine the effect of including or excluding the studies with high risk of bias on treatment effect estimates. This analysis was planned *a priori*.

GRADE assessment of the overall certainty in the body of evidence by outcome

We then used the Grading of Recommendation, Assessment, Development and Evaluation (GRADE) methodology to rate the certainty in the body of evidence for each outcome as high, moderate, low, or very low (18). The assessment included judgments about risk of bias (19), imprecision (20), inconsistency (19), indirectness (20), and publication bias (21). We also rated the certainty in the body of evidence using a GRADE approach to observational studies (22). We crated Evidence Profile and Summary of Findings Tables for each population using GRADE's electronic tool GRADEpro GDT (www.gradepro.org). To assess the usefulness of including NRS, we applied the GRADE NRS framework for when to consider NRS (23). A senior methodologist checked all GRADE tables and ratings of the certainty in the body of evidence (HJS).

RESULTS

The search of the Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, and EMBASE yielded 1806 unique records after duplicates removed. Title and abstract screening by six review authors led to the exclusion of 1635 references. Figure 1 shows the PRISMA flow diagram (24). We reviewed the full text of 171 publications in detail. Overall, 25 RCTs reported in 26 publications, and four NRS reports (2 case-control studies and 2 cohort studies) met our inclusion criteria (25). One RCT reported information related to primary

prevention of allergic diseases. The other 24 RCTs reported information related to other outcomes of interest such as nutritional status, development of rickets, and adverse events. The four NRS reported information about primary prevention of allergic diseases such as asthma/wheezing, allergic rhinitis, and food allergy. The numbers of studies according to the population exposed to Vitamin D are as follow: seven RCTs in pregnant women (reported in eight publications) (25-32), one RCT in breastfeeding women (33), five RCTs to infants (34-38), and twelve RCTs to children (39-50). For the four NRS, two studies reported information in two populations (pregnant women and infants' population) (51, 52), one in breastfeeding women (53), and one in the infant population (54). All of the studies were included in the quantitative analysis. Two RCT published in 2016 (11, 12) did not meet the inclusion criteria defined for the systematic review since both studies compared high doses versus low doses of vitamin D supplementation in pregnant women. No additional information from the reports included was required; therefore, authors of the reports were not contacted.

Table 1 presents the characteristics of the 29 eligible studies according to the population of interest that received Vitamin D. The studies excluded, and the reasons for the exclusions, are reported in Appendix B.

Risk of bias in studies included

Overall risk of bias was deemed between moderate to very low either in RCT as NRS (Appendix C). For RCTs, the greatest concerns were related to allocation

concealment and in the blinding of outcome assessment. In terms of the NRS, risk of bias was evaluated in five outcomes that were reported in four studies. All five outcomes presented moderate risk of bias. Only one study (53) presented a serious limitation due to selection bias of participants and departures from the intended intervention.

Effect of interventions

Table 2-5 describes the summary of findings for all four questions. The full evidence profiles with more detailed explanations are available in Appendix D. Forest plots of meta-analysis are depicted in Appendix E.

Vitamin D supplementation during pregnancy

Only one RCT (25) addressed the primary prevention of developing allergic diseases in children if Vitamin D supplementation was administered during pregnancy. No effect of Vitamin D was found for any allergy outcome: atopic dermatitis (RR, 0.96; 95% CI, 0.57 to 1.61), asthma and/or wheezing (RR, 1.12; 95% CI, 0.50 to 2.54), allergic rhinitis (RR, 0.76; 95% CI, 0.31 to 1.85), and food allergy (RR, 1.92; 95% CI, 0.57 to 6.50). We also found two NRSs, one cohort (52) and one case-control (51) reported development of asthma/wheezing and food allergy respectively. In the cohort study, Vitamin D was associated with low risk of developing wheezing (OR, 0.65; 95% CI, 0.46 to 0.93). No benefit was observed in the case-control study that met the inclusion criteria in food allergy

(OR, 1.50; 95% CI, 0.78 to 2.88). The overall certainty in the body of evidence was very low owing to the risk of bias and imprecision.

Two RCTs reported nutritional status at 1 year of age (29) and development of rickets (26). Vitamin D supplementation in pregnant women was not associated with a change in the weight in infants at 1 year old compared to no Vitamin D supplementation (mean difference, 100 g; 95% CI, -273.48 to 473.48). Development of rickets was not found in infants independently of whether their mothers were exposed to or not exposed to Vitamin D during the last trimester of pregnancy.

Adverse events were reported in seven reports corresponding to six RCTs (25-29, 31, 32). Vitamin D supplementation in pregnant women was not associated with a lower birth weight in infants compared to no vitamin D (mean difference, 52.78 g; 95% CI, -64.34.80 to 169.90). The certainty in the body of evidence of the risk of adverse events in pregnant mothers ranged between low to very low due to the development of few events in both arms. Other adverse effects as measured by infants being born small for gestational age, gestational age, symptomatic hypocalcaemia, as well as severe adverse effects in a newborn and in the mother were very infrequent and any estimates are very imprecise. There was no difference between the groups (see forest plots in Appendix E).

Vitamin D supplementation by mother during breastfeeding

No RCT addressed the primary prevention of developing allergic diseases in children if the supplementation was administered in breastfeeding mothers. One case-control study (53) addressed this question and reported the risk of developing asthma in children. Vitamin D supplementation in breastfeeding mothers did not appear to influence the development of asthma and/or wheezing in children: (OR, 1.09; 95% CI, 0.84 to 1.40).

In one RCT (33), 60 children of breastfeeding mothers that used or did not use vitamin D supplements did not develop rickets after 6 weeks of treatment. The time of follow-up was 6 weeks and the outcome were evaluated as biochemical evidence of developing rickets.

No information about adverse events was reported by the studies that we retrieved from our search strategy in the breastfeeding population that was exposed to vitamin D. The overall certainty in the body of evidence was very low (see forest plot in Appendix E).

Vitamin D supplementation in infants

No RCT addressed any allergy outcome for the primary prevention of allergic diseases with vitamin D supplementation in infants. Three NRSs reported information for the allergy outcomes: allergic rhinitis (54), asthma/wheezing (52, 54), and food allergy (51).

A cohort study (54) reported the association of vitamin D supplementation with allergic rhinitis and asthma. Regular vitamin D supplementation during the first year of life may have increased the risk of developing allergic rhinitis but the confidence interval was extremely wide (RR, 1.95; 95% CI, 0.69 to 5.54). Authors of this study also combined a group that did not receive vitamin D with a group that used it irregularly. If those using vitamin D regularly were compared with those who either did not use it or used it irregularly the RR between the groups is 1.31 (95% CI, 1.15 to 1.49); therefore, regular use increased the risk of allergic rhinitis. In the same study, the point estimate suggested that regular vitamin D supplementation during the first year of life increased the risk of developing asthma/wheezing, but the extremely wide confidence interval showed the data was essentially uninformative (RR, 2.93; 95% CI, 0.19 to 45.35). In a comparison of those using vitamin D regularly with those who either did not use it or used it irregularly, the RR would be 1.33 (95% CI, 1.00 to 1.79). The certainty in the body of evidence for these two outcomes was very low. The other cohort study (52) suggested no impact of Vitamin D supplementation in childhood on asthma/wheezing (OR, 1.00; 95% CI, 0.81 to 1.23).

In the case-control study (51), supplementation of vitamin D in infants reduced the risk of developing food allergy during the 1st year of life (RR, 0.49; 95% CI, 0.27 to 0.88). However, the confidence in this estimate is also very low owing to indirectness of the evidence and risk of bias.

Other outcomes such as development of rickets, nutritional status and adverse events were reported in 5 RCTs (34-38). Estimates of nutrition status and the risk of rickets or adverse effects were imprecise owing to small numbers of events and participants in the studies; they did not differ between the groups receiving or not receiving vitamin D supplements.

Adverse events were infrequent, and the estimates imprecise (RR, 0.85; 95% CI, 0.23 to 3.14). Only one study (37) reported one serious adverse event (RR, 0.17; 95% CI, 0.01 to 3.70), sudden infant death syndrome, in the group of infants that was not receiving vitamin D supplementation. According to the author, this event was not related to the study. The certainty in the body of evidence was very low due to serious risk of bias and imprecision (see forest plot in Appendix E).

Vitamin D supplementation in children

No RCT or NRS addressed the efficacy or association of primary prevention of allergic diseases in children after vitamin D supplementation.

Three RCTs (40, 45, 49) evaluated the efficacy of vitamin D supplementation on nutritional status finding that this supplementation did not benefit nutritional status, measured as weight, on children who received this supplementation (mean difference, 0.87 kg; 95% CI, 0.1 to 1.64). The certainty in the body of evidence was low because of the risk of bias and imprecision.

Few adverse events were reported in eight RCTs (RR, 0.78; 95% CI, 0.26 to

2.32). One study (41) reported two serious adverse events (cyanosis owing to croup, Rota-positive gastroenteritis), one in the vitamin D supplementation group and the other one in the non-supplementation group (RR, 0.74; 95% CI, 0.05 to 11.40). None of these serious adverse events were related to the study. The confidence in the estimates was moderate to low owing risk of bias and imprecision (see forest plot in Appendix E).

Sensitivity analysis

Since the number of studies was small in each particular population for any individual outcome, a sensitivity analysis that considered only low risk of bias studies was not carried out.

DISCUSSION

Our findings, from RCT and NRS, show very low certainty in the body of evidence for the primary prevention of allergic diseases when vitamin D supplementation was administered in four different populations: pregnant women, breastfeeding women, infants, and children. In pregnant woman, we found only one RCT that met our inclusion criteria. In this study, vitamin D did not provide benefit in primary prevention of allergic diseases, independently of whether vitamin D was administered daily [atopic dermatitis (RR, 0.91; 95% CI, 0.50 to 1.66), asthma and/or wheezing (RR, 0.70; 95% CI, 0.35 to 1.40), allergic rhinitis (RR, 0.89; 95% CI, 0.34 to 2.36), and food allergy (RR, 2.38; 95% CI, 0.67 to 8.46)] or in single

dose [atopic dermatitis (RR, 1.02; 95% CI, 0.56 to 1.85), asthma and/or wheezing (RR, 1.03; 95% CI, 0.56 to 1.91), allergic rhinitis (RR, 0.61; 95% CI, 0.19 to 1.94), and food allergy (RR, 1.36; 95% CI, 0.32 to 5.78)].

We did not observe benefit of vitamin D supplementation for primary prevention of allergic diseases in other subgroups of the population such as breastfeeding women, infants, and children. Indeed, supplementation of vitamin D was associated with an increase of risk of developing allergic rhinitis in infant populations, although as in all other evidence this was of very low quality.

We also found no effect of vitamin D supplementation on other outcomes, including adverse events. In particular, our review found very few adverse events in all four populations. For most of these other outcomes, the certainty in the body of evidence was between low and very low.

Our systematic review has several strengths. We established comprehensive and explicit eligibility criteria, conducted a comprehensive search, assessed eligibility and risk of bias, extracted data in duplicate, and applied the GRADE criteria to determine the certainty in the body of evidence. Moreover, we included both RCTs and NRSs to broaden the results and applicability our findings

Limitations of our systematic review are inherent to the evidence retrieved from our search. Only one RCT, and four NRSs, addressed information related to primary prevention of allergic diseases after exposition to vitamin D in four

different populations. There are two reasons that explain these findings. First, we had interest only in the primary prevention of allergic diseases. Second, our interest was focused on the primary prevention of allergic diseases when vitamin D was supplemented, not if serum levels of vitamin D were measured. Thus, we excluded a number of studies that addressed the effect of vitamin D in secondary prevention of allergic diseases, and also, studies that reported the efficacy of vitamin D on primary prevention of allergic diseases that were based only on serum levels of this vitamin D. There is also a limitation in the certainty in the body of evidence of the studies available on this topic, resulting in very low to moderate confidence in estimates for our outcomes. Serious considerations about the susceptibility of bias and imprecisions of the pool estimates (Table 2-5) were noted in most of studies. Thus, these studies are unable to adequately address whether vitamin D supplementation on specific population could be beneficial in primary prevention of allergic diseases.

Other systematic reviews have explored the association between vitamin D intake and primary prevention of allergic diseases. In 2012, a systematic review (2) identified seven observational studies (54-60) which determined a reduction in the risk of wheezing in early childhood (OR, 0.56; 95% CI, 0.42 to 0.73) when vitamin D was taken during pregnancy. These studies used food frequency questionnaires or diet history questionnaires to measure exposure to vitamin D. Another study (61) described the association between maternal 25(OH)-vitamin D concentrations, and atopic eczema and asthma at age 9 months and at age 9

years. Non-statistically significant associations were found between a high concentration of maternal 25(OH)-vitamin D and atopic eczema in their children at 9 months (OR, 1.62; 95% CI, 0.67 to 3.89) and at 9 years old (OR, 1.89; 95% CI, 0.51 to 6.99). Also, a high concentration of maternal 25(OH)-vitamin D was associated with an increase in the risk of developing asthma in children at 9 years old compared with those who were exposed to low concentrations of maternal 25(OH)-vitamin D (OR, 5.40; 95% CI, 1.09 to 26.65). The authors did not report information about vitamin D supplementation and its association with atopic eczema or asthma.

Two other RCTs, published in 2016 (11, 12), explored the effect of prenatal high doses versus low doses of vitamin D supplementation in pregnant women for primary prevention of persistent wheezing or asthma in children. In a RCT conducted by Chawes et al (11), healthy pregnant women who were receiving the usual dosage of vitamin D3 as part of usual pregnancy care (400 IU/day) received Vitamin D3 (2400 IU/day; n = 315) or placebo tablets (n = 308) between pregnancy week 24 and week 1 postpartum. Persistent wheeze and eczema were evaluated through age 3, and asthma was diagnosed at age 3. Vitamin D3 did not show an effect on reducing the risk of persistent wheeze (HR, 0.76; 95% CI, 0.52 to 1.12), asthma (OR, 0.82; 95% CI, 0.50 to 1.36), or eczema (HR, 0.90; 95% CI, 0.65 to 1.26). Litonjua et al (12) compared higher doses of vitamin D3 supplementation (4400 IU/day; n= 440) versus placebo plus usual dose vitamin D3 supplementation (400 IU/day; n= 436). The participants were pregnant

women with history of asthma, eczema, or allergic rhinitis, or pregnant women where the biological father of the child had a history of asthma, eczema, or allergic rhinitis. The pregnant women were randomized at gestational stages between 10 to 18 weeks. The outcomes evaluated in this RCT were asthma/recurrent wheeze, and eczema with rash in typical distribution developed during the first three years of age. Vitamin D did not show an effect on reducing the risk of asthma/recurrent wheezing (HR, 0.8; 95% CI, 0.6 to 1.0) or eczema (HR, 0.9; 95% CI, 0.7 to 1.2) in children by 3 years of age. Thus, in these two RCTs, high doses of vitamin D offered to healthy pregnant women or those with history of asthma, eczema, or allergic rhinitis (including pregnant women where the biological father of the child had a history of asthma, eczema, or allergic rhinitis) did not reduce the risk of asthma, recurrent wheezing, or eczema in their children through age 3.

In terms of practice implications, the data in this review together with the two recent RCTs comparing high with low doses of vitamin D supplementation should not make patients, doctors, and public health authorities confident that vitamin D can primarily prevent allergic diseases.

Suggestions for future research

Further research must be done in order to increase the certainty in the body of evidence, including the precision and directness of the findings. Additional RCTs may help evaluating the effects of using vitamin D supplementation in

breastfeeding mothers and during pregnancy, as well as in infants who did not receive vitamin D prenatally and/or during breastfeeding. Different ways of administration and additional doses of vitamin D could be tested and analyzed in these studies. Finally, new RCTs may address if pregnant women, breastfeeding mothers, or infants would benefit from vitamin D supplementation for the primary prevention of allergic diseases.

CONCLUSION

Although different mechanisms have been proposed for how vitamin D can modulate certain immune functions, limited information is available addressing primary prevention of allergic diseases. Our data suggests that supplementation of vitamin D in different populations does not have an effect in primary prevention of allergic diseases. Future trial studies may be carried out to evaluate the impact of vitamin D in primary prevention of allergic diseases in order to address the remaining uncertainty.

REFERENCES

1. Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clinical and experimental immunology. 2010;160(1):1-9.

2. Nurmatov U, Devereux G, Sheikh A. Nutrients and foods for the primary prevention of asthma and allergy: systematic review and meta-analysis. The Journal of allergy and clinical immunology. 2011;127(3):724-33 e1-30.

3. Bacharier LB. Vitamin D status at birth: an important and potentially modifiable determinant of atopic disease in childhood? The Journal of allergy and clinical immunology. 2014;133(1):154-5.

4. Weiss ST, Litonjua AA. Vitamin D, the gut microbiome, and the hygiene hypothesis. How does asthma begin? American journal of respiratory and critical care medicine. 2015;191(5):492-3.

5. Theodoratou E, Tzoulaki I, Zgaga L, Ioannidis JP. Vitamin D and multiple health outcomes: umbrella review of systematic reviews and meta-analyses of observational studies and randomised trials. Bmj. 2014;348:g2035.

6. Arikoglu T, Kuyucu S, Karaismailoglu E, Batmaz SB, Balci S. The association of vitamin D, cathelicidin, and vitamin D binding protein with acute asthma attacks in children. Allergy and asthma proceedings : the official journal of regional and state allergy societies. 2015;36(4):51-8.

7. Rudders SA, Camargo CA, Jr. Sunlight, vitamin D and food allergy. Current opinion in allergy and clinical immunology. 2015;15(4):350-7.

8. Chiu CY, Huang SY, Peng YC, Tsai MH, Hua MC, Yao TC, et al. Maternal vitamin D levels are inversely related to allergic sensitization and atopic diseases in early childhood. Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology. 2015;26(4):337-43.

9. Maalmi H, Berraies A, Tangour E, Ammar J, Abid H, Hamzaoui K, et al. The impact of vitamin D deficiency on immune T cells in asthmatic children: a case-control study. Journal of asthma and allergy. 2012;5:11-9.

10. Hollams EM, Hart PH, Holt BJ, Serralha M, Parsons F, de Klerk NH, et al. Vitamin D and atopy and asthma phenotypes in children: a longitudinal cohort study. Eur Respir J. 2011;38(6):1320-7.

11. Chawes BL, Bonnelykke K, Stokholm J, Vissing NH, Bjarnadottir E, Schoos A-MM, et al. Effect of Vitamin D3 Supplementation During Pregnancy on Risk of Persistent Wheeze in the Offspring: A Randomized Clinical Trial. Jama. 2016;315(4):353-61.

12. Litonjua AA, Carey VJ, Laranjo N, Harshfield BJ, McElrath TF, O'Connor GT, et al. Effect of Prenatal Supplementation With Vitamin D on Asthma or Recurrent Wheezing in Offspring by Age 3 Years: The VDAART Randomized Clinical Trial. Jama. 2016;315(4):362-70.

13. Yepes-Nuñez JJ, Fiocchi A, Pawankar R, Cuello-Garcia C, Zhang Y, Morgano GP, et al. World Allergy Organization-McMaster University Guidelines for Allergic Disease Prevention (GLAD-P): Vitamin D. The World Allergy Organization journal.Forthcoming 2016.

14. Yepes-Nunez JJ, Fiocchi A, Pawankar R, Cuello-Garcia CA, Zhang Y, Morgano GP, et al. World Allergy Organization-McMaster University Guidelines for Allergic Disease Prevention (GLAD-P): Vitamin D. The World Allergy Organization journal. 2016;9:17.

15. Higgins JPT, Green S, (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1 [updated March 2011]. The Cochrane Collaboration, 2011. Available from: http://www.cochrane-handbook.org.

16. Sterne JAC, Higgins JPT, Reeves BC on behalf of the development group for ACROBAT-NRSI. A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0, 24 September 2014. Available from http://www.riskofbias.info [01 December 2014].

17. Review Manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

18. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj. 2008;336(7650):924-6.

19. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). Journal of clinical epidemiology. 2011;64(4):407-15.

20. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. Journal of clinical epidemiology. 2011;64(12):1283-93.

21. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. Journal of clinical epidemiology. 2011;64(12):1294-302.

22. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. Journal of clinical epidemiology. 2011;64(12):1303-10.

23. Schunemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Research synthesis methods. 2013;4(1):49-62.

24. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Bmj. 2009;339:b2535.

25. Goldring ST, Griffiths CJ, Martineau AR, Robinson S, Yu C, Poulton S, et al. Prenatal vitamin d supplementation and child respiratory health: a randomised controlled trial. PloS one. 2013;8(6):e66627.

26. Brooke OG, Brown IR, Bone CD, Carter ND, Cleeve HJ, Maxwell JD, et al. Vitamin D supplements in pregnant Asian women: effects on calcium status and fetal growth. British medical journal. 1980;280(6216):751-4.

27. Hossain N, Kanani FH, Ramzan S, Kausar R, Ayaz S, Khanani R, et al. Obstetric and neonatal outcomes of maternal vitamin D supplementation: results of an open-label, randomized controlled trial of antenatal vitamin D

supplementation in Pakistani women. The Journal of clinical endocrinology and metabolism. 2014;99(7):2448-55.

28. Roth DE, Al Mahmud A, Raqib R, Akhtar E, Perumal N, Pezzack B, et al. Randomized placebo-controlled trial of high-dose prenatal third-trimester vitamin D3 supplementation in Bangladesh: the AViDD trial. Nutrition journal. 2013;12:47.

29. Roth DE, Perumal N, Al Mahmud A, Baqui AH. Maternal vitamin D3 supplementation during the third trimester of pregnancy: effects on infant growth in a longitudinal follow-up study in Bangladesh. The Journal of pediatrics. 2013;163(6):1605-11 e3.

30. Mallet E, Gugi B, Brunelle P, Henocq A, Basuyau JP, Lemeur H. Vitamin D supplementation in pregnancy: a controlled trial of two methods. Obstetrics and gynecology. 1986;68(3):300-4.

31. Yu CK, Sykes L, Sethi M, Teoh TG, Robinson S. Vitamin D deficiency and supplementation during pregnancy. Clinical endocrinology. 2009;70(5):685-90.

32. Sablok A, Thariani K, Batra A, Bharti R, Aggarwal AR, Kabi BC, et al. Supplementation of Vitamin D in pregnancy and its correlation with feto-maternal outcome. Clinical endocrinology. 2015; 83(4):536-41.

33. Rothberg AD, Pettifor JM, Cohen DF, Sonnendecker EW, Ross FP. Maternal-infant vitamin D relationships during breast-feeding. The Journal of pediatrics. 1982;101(4):500-3.

34. Alonso A, Rodriguez J, Carvajal I, Prieto MA, Rodriguez RM, Perez AM, et al. Prophylactic vitamin D in healthy infants: assessing the need. Metabolism: clinical and experimental. 2011;60(12):1719-25.

35. Chan GM, Roberts CC, Folland D, Jackson R. Growth and bone mineralization of normal breast-fed infants and the effects of lactation on maternal bone mineral status. The American journal of clinical nutrition. 1982;36(3):438-43.

36. Greer FR, Searcy JE, Levin RS, Steichen JJ, Steichen-Asche PS, Tsang RC. Bone mineral content and serum 25-hydroxyvitamin D concentrations in breast-fed infants with and without supplemental vitamin D: one-year follow-up. The Journal of pediatrics. 1982;100(6):919-22.

37. Ponnapakkam T, Bradford E, Gensure R. A treatment trial of vitamin D supplementation in breast-fed infants: universal supplementation is not necessary for rickets prevention in Southern Louisiana. Clinical pediatrics. 2010;49(11):1053-60.

38. Madar AA, Klepp KI, Meyer HE. Effect of free vitamin D(2) drops on serum 25-hydroxyvitamin D in infants with immigrant origin: a cluster randomized controlled trial. European journal of clinical nutrition. 2009;63(4):478-84.

39. Abrams SA, Hawthorne KM, Chen Z. Supplementation with 1000 IU vitamin D/d leads to parathyroid hormone suppression, but not increased fractional calcium absorption, in 4-8-y-old children: a double-blind randomized controlled trial. The American journal of clinical nutrition. 2013;97(1):217-23.

40. Guillemant J, Le HT, Maria A, Allemandou A, Peres G, Guillemant S. Wintertime vitamin D deficiency in male adolescents: effect on parathyroid function and response to vitamin D3 supplements. Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA. 2001;12(10):875-9.

41. Hower J, Knoll A, Ritzenthaler KL, Steiner C, Berwind R. Vitamin D fortification of growing up milk prevents decrease of serum 25-hydroxyvitamin D concentrations during winter: a clinical intervention study in Germany. European journal of pediatrics. 2013;172(12):1597-605.

42. Khadilkar AV, Sayyad MG, Sanwalka NJ, Bhandari DR, Naik S, Khadilkar VV, et al. Vitamin D supplementation and bone mass accrual in underprivileged adolescent Indian girls. Asia Pacific journal of clinical nutrition. 2010;19(4):465-72.

43. Lewis RD, Laing EM, Hill Gallant KM, Hall DB, McCabe GP, Hausman DB, et al. A randomized trial of vitamin D(3) supplementation in children: dose-response effects on vitamin D metabolites and calcium absorption. The Journal of clinical endocrinology and metabolism. 2013;98(12):4816-25.

44. Maalouf J, Nabulsi M, Vieth R, Kimball S, El-Rassi R, Mahfoud Z, et al. Short- and long-term safety of weekly high-dose vitamin D3 supplementation in school children. The Journal of clinical endocrinology and metabolism. 2008;93(7):2693-701.

45. Molgaard C, Larnkjaer A, Cashman KD, Lamberg-Allardt C, Jakobsen J, Michaelsen KF. Does vitamin D supplementation of healthy Danish Caucasian girls affect bone turnover and bone mineralization? Bone. 2010;46(2):432-9.

46. Rich-Edwards JW, Ganmaa D, Kleinman K, Sumberzul N, Holick MF, Lkhagvasuren T, et al. Randomized trial of fortified milk and supplements to raise 25-hydroxyvitamin D concentrations in schoolchildren in Mongolia. The American journal of clinical nutrition. 2011;94(2):578-84.

47. Rajakumar K, Moore CG, Yabes J, Olabopo F, Haralam MA, Comer D, et al. Effect of Vitamin D3 Supplementation in Black and in White Children: A Randomized, Placebo-Controlled Trial. The Journal of clinical endocrinology and metabolism. 2015; 100(8):3183-92.

48. Urashima M, Segawa T, Okazaki M, Kurihara M, Wada Y, Ida H. Randomized trial of vitamin D supplementation to prevent seasonal influenza A in schoolchildren. The American journal of clinical nutrition. 2010;91(5):1255-60.

49. Viljakainen HT, Natri AM, Karkkainen M, Huttunen MM, Palssa A, Jakobsen J, et al. A positive dose-response effect of vitamin D supplementation on site-specific bone mineral augmentation in adolescent girls: a double-blinded randomized placebo-controlled 1-year intervention. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2006;21(6):836-44.

50. Ward KA, Das G, Roberts SA, Berry JL, Adams JE, Rawer R, et al. A randomized, controlled trial of vitamin D supplementation upon musculoskeletal

health in postmenarchal females. The Journal of clinical endocrinology and metabolism. 2010;95(10):4643-51.

51. Allen KJ, Koplin JJ, Ponsonby AL, Gurrin LC, Wake M, Vuillermin P, et al. Vitamin D insufficiency is associated with challenge-proven food allergy in infants. The Journal of allergy and clinical immunology. 2013;131(4):1109-16, 16 e1-6.

52. Anderson LN, Chen Y, Omand JA, Birken CS, Parkin PC, To T, et al. Vitamin D exposure during pregnancy, but not early childhood, is associated with risk of childhood wheezing. Journal of developmental origins of health and disease. 2015;6(4):308-16.

53. Bener A, Ehlayel MS, Tulic MK, Hamid Q. Vitamin D deficiency as a strong predictor of asthma in children. International archives of allergy and immunology. 2012;157(2):168-75.

54. Hypponen E, Sovio U, Wjst M, Patel S, Pekkanen J, Hartikainen AL, et al. Infant vitamin d supplementation and allergic conditions in adulthood: northern Finland birth cohort 1966. Annals of the New York Academy of Sciences. 2004;1037:84-95.

55. Back O, Blomquist HK, Hernell O, Stenberg B. Does vitamin D intake during infancy promote the development of atopic allergy? Acta dermato-venereologica. 2009;89(1):28-32.

56. Camargo CA, Jr., Rifas-Shiman SL, Litonjua AA, Rich-Edwards JW, Weiss ST, Gold DR, et al. Maternal intake of vitamin D during pregnancy and risk of

recurrent wheeze in children at 3 y of age. The American journal of clinical nutrition. 2007;85(3):788-95.

57. Devereux G, Litonjua AA, Turner SW, Craig LC, McNeill G, Martindale S, et al. Maternal vitamin D intake during pregnancy and early childhood wheezing. The American journal of clinical nutrition. 2007;85(3):853-9.

58. Erkkola M, Kaila M, Nwaru BI, Kronberg-Kippila C, Ahonen S, Nevalainen J, et al. Maternal vitamin D intake during pregnancy is inversely associated with asthma and allergic rhinitis in 5-year-old children. Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology. 2009;39(6):875-82.

59. Miyake Y, Sasaki S, Tanaka K, Hirota Y. Dairy food, calcium and vitamin D intake in pregnancy, and wheeze and eczema in infants. Eur Respir J. 2010;35(6):1228-34.

60. Nja F, Nystad W, Lodrup Carlsen KC, Hetlevik O, Carlsen KH. Effects of early intake of fruit or vegetables in relation to later asthma and allergic sensitization in school-age children. Acta paediatrica. 2005;94(2):147-54.

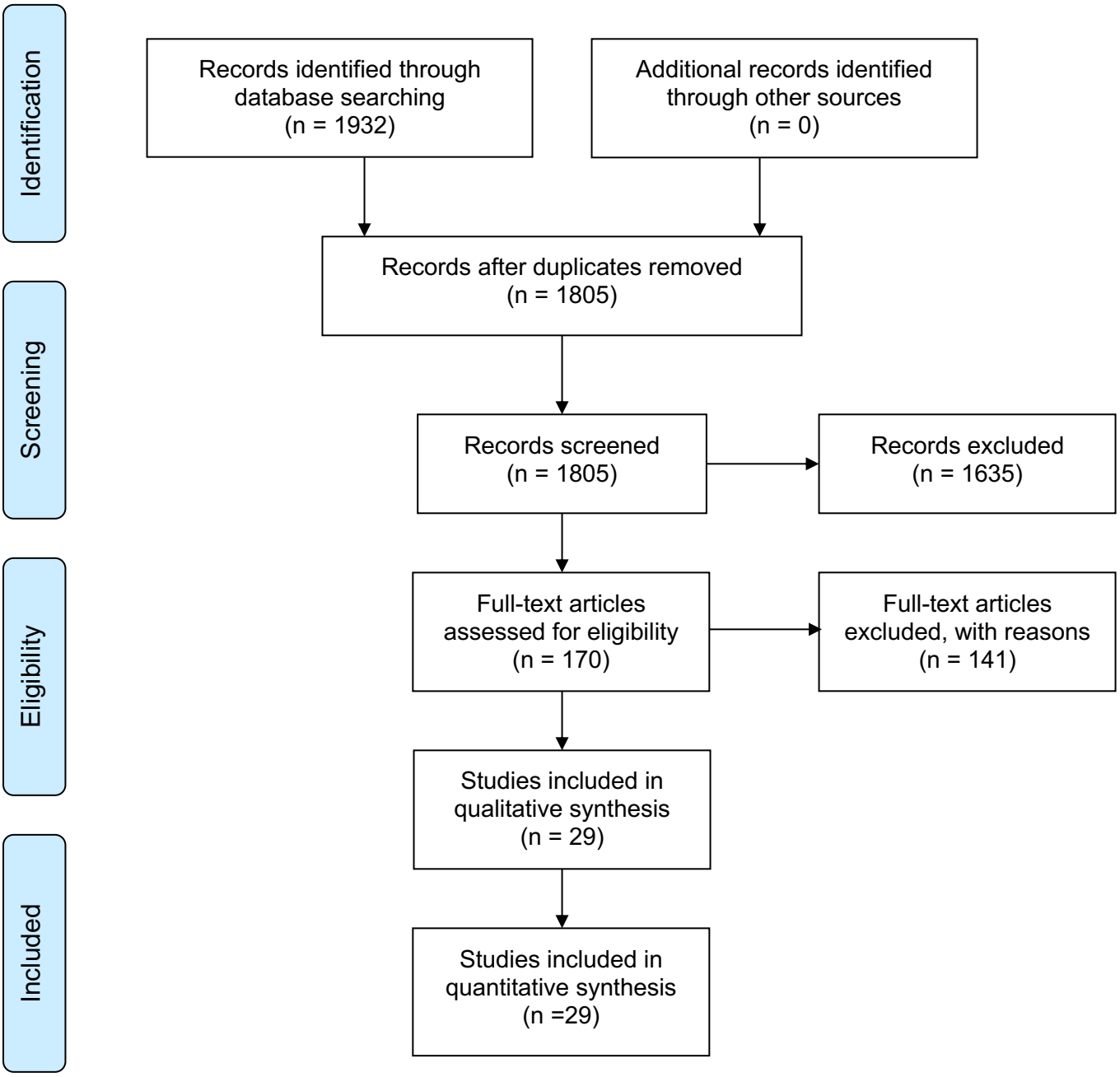
61. Gale CR, Robinson SM, Harvey NC, Javaid MK, Jiang B, Martyn CN, et al. Maternal vitamin D status during pregnancy and child outcomes. European journal of clinical nutrition. 2008;62(1):68-77.

62. Baiz N, Dargent-Molina P, Wark JD, Souberbielle JC, Annesi-Maesano I, Group EM-CCS. Cord serum 25-hydroxyvitamin D and risk of early childhood

transient wheezing and atopic dermatitis. The Journal of allergy and clinical immunology. 2014;133(1):147-53.

FIGURE

Figure 1. PRISMA flow chart of study selection



TABLES

Table 1. Characteristics of included studies

1. PREGNANT WOMEN

Study, Year	n	Population who received vitamin D	Type of vitamin D administered	Route of administration and forms	Doses of vitamin D administered	Time of administration of vitamin D	Time of follow-up for outcome(s)
Randomized control trials							
Brooke, 1980 (29)	126*	**Pregnant women** with 28 gestation weeks	Ergocalciferol	Oral (unreported)	1000 IU/day	Until delivery	At birth
Goldring, 2013 (25)	180*	**Pregnant women** at 27 weeks of gestation	Vitamin D3 (ergocalciferol)	Oral (tablets)	800 IU/day or 200000 UI (single dose)**	2 months (until delivery)	At 36 months
Hossain, 2014 (30)	200*	**Pregnant women** at 20 weeks of gestation	Vitamin D3 (Ergocalciferol)	Oral (liquid formulation)	4000 IU/day	Until delivery	At birth
Mallet, 1986 (52)	57*	**Pregnant women** at seven month	Vitamin D2	Unreported	1000 IU/day or 200000/IU (single dose)	Last trimester of pregnancy	At birth
Roth, 2013-A (31)	160*	**Pregnant women** with gestational age of 26 to < 30 weeks	Vitamin D3 (cholecalciferol)	Oral (liquid formulation)	35000 IU/week	3 months (last trimester of pregnancy)	At birth
Roth, 2013-B (28)	160*	**Pregnant women** with gestational age of 26 to < 30 weeks	Vitamin D3 (cholecalciferol)	Oral (liquid formulation)	35000 IU/week	3 months (last trimester of pregnancy)	One year after delivery
Sablok, 2015 (33)	180*	**Pregnant women** with gestational age at 14–20 weeks	Vitamin D (cholecalciferol)	Oral	Between 60000 IU to 480000 IU	Between second to third trimester of pregnancy	At birth
Yu, 2009 (32)	180*	**Pregnant women** with 27 gestation weeks	Vitamin D (Ergocalciferol or calciferol)	Oral	800 UI/day or 200000 IU (single dose)	Until delivery	At birth
Observational studies							

Allen, 2013 (27)	2252	**Pregnant women**	Maternal use of vitamin D supplement	Unreported	Unreported	During pregnancy	Infants with aged between 11 and 15 months
Anderson 2015 (26)	2478	**Pregnant women**	Maternal use of vitamin D supplement	Unreported	400 or 1000 IU/day	During pregnancy	Infants with aged between 0 to 5 years

2. BREASTFEEDING

Study, Year	n	Population who received vitamin D	Type of vitamin D administered	Route of administration and forms	Doses of vitamin D administered	Time of administration of vitamin D	Time of follow-up for outcome(s)
Randomized control trials							
Rothberg, 1982 (34)	60*	**Breastfeeding mothers**	Vitamin D	Unreported	500 IU or 1000 IU/day	6 weeks	At 6 weeks of age
Observational studies							
Bener, 2012 (24)	966	**Breastfeeding mothers**	Vitamin D supplementation	Unreported	Unreported	During breastfeeding	Children with different age group

3. INFANTS

Study, Year	n	Population who received vitamin D	Type of vitamin D administered	Route of administration and forms	Doses of vitamin D administered	Time of administration of vitamin D	Time of follow-up for outcome(s)
Randomized control trials							
Alonso, 2011 (36)	102*	Full-term 1-month-old healthy **infants**	Vitamin D (cholecalciferol)	Oral (drops)	402 IU/day	12 months	At 1 year of age
Chan, 1982 (40)	51*	Breastfeeding healthy term **infants** and **lactating mother**	Vitamin D	Unreported	400 IU/day	4 months	At 1 year of age
Greer, 1982 (37)	18*	Breastfeeding healthy term **infants**	Vitamin D	Unreported	400 IU/day	Until infants were weaned from breast-feeding	At 6 months of age
Madar, 2009 (38)	66***	6 week-old **infants**	Vitamin D2	Oral (drops)	400 IU/day	6 weeks	At 3 months of age

Ponnapakkam, 2010 (39)	80*	Breast-fed **term infants** †	Vitamin D3	Oral (drops)	200 IU/day	Between 4 to 6 months	At 6 months of age
Observational studies							
Allen, 2013 (27)	2252	**Infants**	Vitamin D supplement	Unreported)	Unreported	During infancy	Infants with aged between 11 and 15 months
Anderson 2015 (26)	2478	**Infants**	Vitamin D supplement	Unreported	400 IU/day	During childhood	Infants with aged between 0 to 5 years
Hypponen, 2004 (35)	7648	**Infants**	Vitamin D supplement	Unreported	<2000 IU to >2000 IU‡	During the first year of life	At age 31 years

4. CHILDREN

Study, Year	n	Population who received vitamin D	Type of vitamin D administered	Route of administration and forms	Doses of vitamin D administered	Time of administration of vitamin D	Time of follow-up for outcome(s)
Randomized control trials							
Abrams, 2013 (53)	64*	**Children** between 4-8.9 years old	Vitamin D3	Oral	1000 IU/day	8 weeks	At the end of the trial (8th week later)
Guillemant 2001 (41)	57*	**Adolescents** between 13 to 16 years old	Vitamin D3	Oral (phial of water-soluble oral solution)	100000 IU/3 times	18 months	At the end of the trial (18 months later)
Hower 2013 (44)	92*	**Children** between 2-6 years old	Vitamin D fortified milk	Oral	400 IU/day	8 months	At the end of the trial (8 months later)
Khadilkar 2010 (54)	50*	**Adolescents** between 14 to 15 years old	Vitamin D2	Oral	300000/4 times	12 months	At the end of the trial (12 months later)
Lewis 2013 (55)	323*	**Children** between 9-13 years old	Vitamin D3	Oral (tablets)	400, 1000, 2000, or 4000 IU/day	12 weeks	At the end of the trial (12 weeks later)
Maalouf 2008 (56)	340*	**Adolescents** between 10 to 17 years old	Vitamin D3	Oral	1400 or 14000 UI/weekly	12 months	At the end of the trial (12 months later)
Mølgaard 2010 (42)	225*	**Children** between 11 to 12 years old	Vitamin D3	Oral	200 or 400 IU/day	12 months	At the end of the trial (12 months later)

Rich Edwards 2011 (57)	579	**Children** between 9 to 11 years old	Vitamin D3	Oral	13700 IU in one week or 300 IU/day	49 days	At the end of the trial
Rajakumar, 2015 (58)	157*	**Children** between 8 to 14 years old	Vitamin D3	Oral (tablets)	1000 IU/daily	6 months	At the end of the trial (6 months later)
Urashima 2010 (59)	430*	**Children** between 6 to 15 years old	Vitamin D3	Oral (tablets)	1200 IU/day	4 months	At the end of the trial (4 months later)
Vijakainen 2006 (43)	228	**Adolescents** between 11 to 12 years old	Vitamin D3	Oral	200 or 400 IU/day	12 months	At the end of the trial (12 months later)
Ward 2010 (60)	73*	**Adolescents** between 12 to 14 years old	Vitamin D 2 (ergocalciferol)	Oral	150000 IU/four times	12 months	At the end of the trial (12 months later)

*Patients randomized. **Results presented in meta-analysis correspond to pooled data. Single oral dose was administered as cholecalciferol (oral bolus). ***Randomization was carried out at the health clinic level. † Two group of infants received vitamin D: one group started the vitamin D supplementation at birth and another group started at 2 months of age. ‡ The daily dose of vitamin D was calculated on the basis of concentration of vitamin D in the product used and the reported dosage (categorized as <50 μg, 50 μg, or >50 μg).

Table 2. Summary of Findings table for Vitamin D versus no vitamin D supplementation in PREGNANT women for the primary prevention of allergy in their children and other important outcomes

Vitamin D versus no vitamin D supplementation in PREGNANT women for the primary prevention of allergy in their children and other important outcomes					
Outcomes	**№ of participants (studies)**	**Certainty in the evidence (GRADE)**	**Relative effect (95% CI)**	**Anticipated absolute effects**	
				Risk with no Vitamin D	**Risk difference with Vitamin D**
Eczema follow up: 3 years	151 (1 RCT)	⊕◯◯◯ VERY LOW [1,2,3]	**RR 0.96** (0.57 to 1.61)	**Study population**	
				306 per 1000	**12 fewer per 1000** (132 fewer to 187 more)
				Moderate	
				10 per 1000	**0 fewer per 1000** [5] (4 fewer to 6 more)
Allergic rhinitis follow up: 3 years	150 (1 RCT)	⊕◯◯◯ VERY LOW [1,2,3]	**RR 0.76** (0.31 to 1.85)	**Study population**	
				143 per 1000	**34 fewer per 1000** (99 fewer to 121 more)
				Moderate	
				146 per 1000	**35 fewer per 1000** [6] (101 fewer to 124 more)
				Study population	

				140 per 1000	**17 more per 1000** (70 fewer to 216 more)
Asthma/wheezing follow up: 3 years	158 (1 RCT)	⊕◯◯◯ VERY LOW [1,2,3]	**RR 1.12** (0.50 to 2.54)	**Moderate**	
				94 per 1000	**11 more per 1000** [7] (47 fewer to 145 more)
Asthma/wheezing [8]	2478 (1 observational study)	⊕◯◯◯ VERY LOW [3]	**OR 0.65** [9] (0.46 to 0.92)	**Moderate**	
				94 per 1000	**31 fewer per 1000** [7] (48 fewer to 7 fewer)
Food allergy follow up: 3 years	151 (1 RCT)	⊕◯◯◯ VERY LOW [1,2,3]	**RR 1.92** (0.57 to 6.50)	**Study population**	
				61 per 1000	**56 more per 1000** (26 fewer to 337 more)
				Moderate	
				33 per 1000	**30 more per 1000** [10] (14 fewer to 182 more)
Food allergy	27 cases 213 controls (1 observational study)	⊕◯◯◯ VERY LOW [3]	**OR 1.50** [9] (0.78 to 2.88)	**Low**	
				33 per 1000	**16 more per 1000** [10] (7 fewer to 56 more)
Rickets	126 (1 RCT)	⊕⊕◯◯ LOW [3]	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Nutritional status: birth weight (gr)	821 (6 RCTs)	⊕⊕⊕◯ MODERATE [4]	-	The mean nutritional status birth weight was **0**	MD **52.78 more** (64.34 fewer to 169.9 more)
			Not estimable	**Study population**	

Any adverse events (children)	234 (2 RCTs)	⊕⊕◯◯ LOW [3]		0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Any adverse events (62) assessed with: Preeclampsia	520 (3 RCTs)	⊕◯◯◯ VERY LOW [3,11]	**RR 1.46** (0.31 to 6.78)	**Study population**	
				117 per 1000	**54 more per 1000** (80 fewer to 674 more)
Serious adverse events (children)	426 (3 RCTs)	⊕⊕◯◯ LOW [3]	**RR 0.79** (0.33 to 1.90)	**Study population**	
				47 per 1000	**10 fewer per 1000** (31 fewer to 42 more)
Serious adverse events (62)	321 (2 RCTs)	⊕⊕◯◯ LOW [3]	**RR 1.55** (0.65 to 3.69)	**Study population**	
				49 per 1000	**27 more per 1000** (17 fewer to 132 more)

*__The risk in the intervention group__ (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; **RR:** Risk ratio; **MD:** Mean difference; **OR:** Odds ratio

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

1. Serious risk of bias. One study was rated as high risk of bias due to lack of concealment of allocation, and lack of blinding of participants and caregivers
2. Indirectness. One study assessed the impact of the intervention in 3-year-old children. However, the estimate of the treatment effect could be different in other populations
3. Very serious imprecision. Wide confidence interval due to a small number of events
4. Very serious imprecision. Wide confidence interval crosses 0
5. Difference in anticipated absolute effect was calculated based on the range of prevalence of eczema in International Study for Asthma and Allergy in Childhood (ISAAC) Phase Three (1%)
6. Difference in anticipated absolute effect was calculated based on the total global prevalence of rhinoconjunctivitis symptoms for the 13- to 14-year old children according with Phase Three ISAAC (14.6%)

7. Difference in anticipated absolute effect was calculated based on the total global prevalence of asthma symptoms in the 6-7 year old children according with Phase Three ISAAC in 9.4%
8. Wheezing was measured using the ISAAC questionnaire
9. Odds ratio adjusted by sociodemographic and environmental factors, health history, and individual activities
10. Difference in anticipated absolute effect was calculated based on the prevalence of food allergy in the population as a result of food challenge tests to any food according with meta-analysis of Rona et al. 2007
11. Serious risk of bias. Studies that carried large weight for the overall effect estimate rated as high risk of bias due to selective outcome reporting

Table 3. Summary of Findings table for Vitamin D versus no vitamin D supplementation in BREASTFEEDING women for the primary prevention of allergy in their children and other important outcomes

Vitamin D versus no vitamin D supplementation in BREASTFEEDING women for the primary prevention of allergy in their children and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no Vitamin D	Risk difference with Vitamin D
Eczema	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Allergic rhinitis	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Asthma/wheezing (follow-up: up to 16 years)	483 cases 483 controls (1 observational study)	⨁◯◯◯ VERY LOW [1,2,3]	**OR 1.09** (0.84 to 1.40)	**Moderate**	
				94 per 1000	**8 more per 1000** [4] (14 fewer to 33 more)
Food allergy	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)

Vitamin D versus no vitamin D supplementation in BREASTFEEDING women for the primary prevention of allergy in their children and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no Vitamin D	Risk difference with Vitamin D
Rickets follow up: 6 weeks	28 (1 RCT)	⨁◯◯◯ VERY LOW [5,6]	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Nutritional status	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Any adverse events	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Severe adverse events	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)

*__The risk in the intervention group__ (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; **OR:** Odds ratio

Vitamin D versus no vitamin D supplementation in BREASTFEEDING women for the primary prevention of allergy in their children and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no Vitamin D	Risk difference with Vitamin D

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

1. Serious risk of bias. Controls were sampled from a population that is unlikely to be representative of the population from which cases were selected, and analysis was not adjusted for confounding factors
2. Indirectness. Over 80% of the children in the study had mild to severe vitamin D deficiency
3. Serious imprecision. Wide confidence interval
4. Difference in anticipated absolute effect was calculated based on the total global prevalence of asthma symptoms in the 6-7 year old children according with Phase Three ISAAC in 9.4%
5. Serious risk of bias. One study was rated as unclear risk of bias due to insufficient information about random sequence generation, allocation concealment, and incomplete outcome data
6. Very serious imprecision with no events in a small study

Table 4. Summary of Findings table for Vitamin D versus no vitamin D supplementation in INFANTS for the primary prevention of allergy and other important outcomes

Vitamin D versus no vitamin D supplementation in INFANTS for the primary prevention of allergy and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no vitamin D	Risk difference with Vitamin D
Eczema	(0 studies)	-	not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Allergic rhinitis [1] follow up: 31 years	6768 (1 observational study)	⊕◯◯◯ VERY LOW [2,3]	**RR 1.95** (0.69 to 5.54)	**Study population**	
				150 per 1000	**143 more per 1000** (47 fewer to 681 more)
				Moderate	
				146 per 1000	**139 more per 1000** [4] (45 fewer to 663 more)
Asthma/wheezing [1] follow up: 31 years	6768 (1 observational study)	⊕◯◯◯ VERY LOW [2,3]	**RR 3.07** (0.19 to 50.88)	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
				Moderate	
				94 per 1000	**195 more per 1000** [5] (76 fewer to 4689 more)

Asthma/wheezing [6] follow up: 72 months	2478 (1 observational study)	⊕◯◯◯ VERY LOW [3]	**OR 1.00** (0.81 to 1.23)	**Moderate**	
				94 per 1000	**0 fewer per 1000** [5] (16 fewer to 19 more)
Food allergy follow up: 1 years	481 (1 observational study)	⊕◯◯◯ VERY LOW [7,8,9]	**OR 0.49** (0.27 to 0.88)	**Study population**	
				145 per 1000	**68 fewer per 1000** (101 fewer to 15 fewer)
				Moderate	
				33 per 1000	**17 fewer per 1000** [10] (24 fewer to 4 fewer)
Rickets follow up: range 2 months to 12 months	177 (4 RCTs)	⊕◯◯◯ VERY LOW [11,12,3]	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Nutritional status (body weight in grams) follow up: range 6 months to 12 months	80 (1 RCT)	⊕⊕◯◯ LOW [13,14]	-	The mean nutritional status (body weight in grams) was **0**	**0** (0 to 0)
Any adverse events	109 (2 RCTs)	⊕⊕◯◯ LOW [15,3]	**RR 0.85** (0.23 to 3.14)	**Study population**	
				78 per 1000	**12 fewer per 1000** (60 fewer to 168 more)
Serious adverse events	25 (1 RCT)	⊕⊕◯◯ LOW [15,3]	**RR 0.17** (0.01 to 3.70)	**Study population**	
				125 per 1000	**104 fewer per 1000** (124 fewer to 338 more)

*__The risk in the intervention group__ (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; **RR:** Risk ratio; **OR:** Odds ratio

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

1. Information corresponds to case control study, Hypponen 2004
2. Serious risk of bias. Criteria for diagnosis of asthma and allergic rhinitis were not provided, and results were not adjusted for confounding factors
3. Serious imprecision. Wide confidence interval
4. Difference in anticipated absolute effect was calculated based on the total global prevalence of rhinoconjunctivitis symptoms for the 13- to 14-year old children according with Phase Three ISAAC (14.6%)
5. Difference in anticipated absolute effect was calculated based on the total global prevalence of asthma symptoms in the 6-7 year old children according with Phase Three ISAAC (9.4%)
6. Information corresponds to cohort study, Anderson 2015
7. Serious risk of bias. Results were not adjusted for confounding factors
8. Serious indirectness. The outcome was measured after 1 year of life, which is likely a short frame time to reliably measure development of food allergy
9. Serious imprecision with only 51 events in total
10. Difference in anticipated absolute effect was calculated based on the prevalence of food allergy in the population as a result of food challenge tests to any food according with meta-analysis of Rona et al. 2007 (3.3%)
11. Serious risk of bias. Studies that carried large weight for the overall effect estimate rated as high risk of bias due to lack of blinding and incomplete outcome data
12. Serious indirectness. Patients included in the study were follow-up for a short time to assess the outcome
13. Serious risk of bias. One study was rated as high risk of bias due to lack of random sequence generation
14. Serious imprecision. Mean weights were not significant different between patients who received vitamin D supplementation versus no vitamin D supplementation
15. Serious risk of bias. Studies that carried large weight for the overall effect estimate rated as high risk of bias due to incomplete outcome data

Table 5. Summary of Findings table for Vitamin D versus no vitamin D supplementation in CHILDREN for the primary prevention of allergy and other important outcomes

Vitamin D versus no vitamin D supplementation in CHILDREN for the primary prevention of allergy and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no Vitamin D	Risk difference with Vitamin D
Eczema	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Allergic rhinitis	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Asthma/wheezing	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Food allergy	(0 studies)	-	Not estimable	**Study population**	
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Rickets	(0 studies)	-	Not estimable	**Study population**	

Vitamin D versus no vitamin D supplementation in CHILDREN for the primary prevention of allergy and other important outcomes

Outcomes	№ of participants (studies)	Certainty in the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no Vitamin D	Risk difference with Vitamin D
				0 per 1000	**0 fewer per 1000** (0 fewer to 0 fewer)
Nutritional status	495 (3 RCTs)	⨁⨁◯◯ LOW [1,2]	-	The mean nutritional status was 0 kg	MD **0.87 more** (0.1 more to 1.64 more)
Any adverse events	1895 (8 RCTs)	⨁⨁⨁◯ MODERATE [2]	**RR 0.78** (0.26 to 2.32)	**Study population**	
				8 per 1000	**2 fewer per 1000** (6 fewer to 11 fewer)
Serious adverse events	80 (1 RCTs)	⨁⨁◯◯ LOW [2]	**RR 0.74** (0.05 to 11.40)	**Study population**	
				29 per 1000	**8 fewer per 1000** (28 fewer to 306 more)

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; **SMD:** Standardised mean difference; **OR:** Odds ratio

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

1. Serious risk of bias. Studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to insufficient information about random sequence generation, allocation concealment, blinding of outcome assessment, and incomplete outcome data.
2. Serious imprecision. Wide confidence interval
3. Serious risk of bias. One studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to insufficient information about allocation concealment, a blinding of outcome assessment

APPENDIXES

Appendix A. Search strategies

See the search strategies in the following link:

https://www.dropbox.com/s/gesfvtaixelge09/Appendix%20A.doc?dl=0

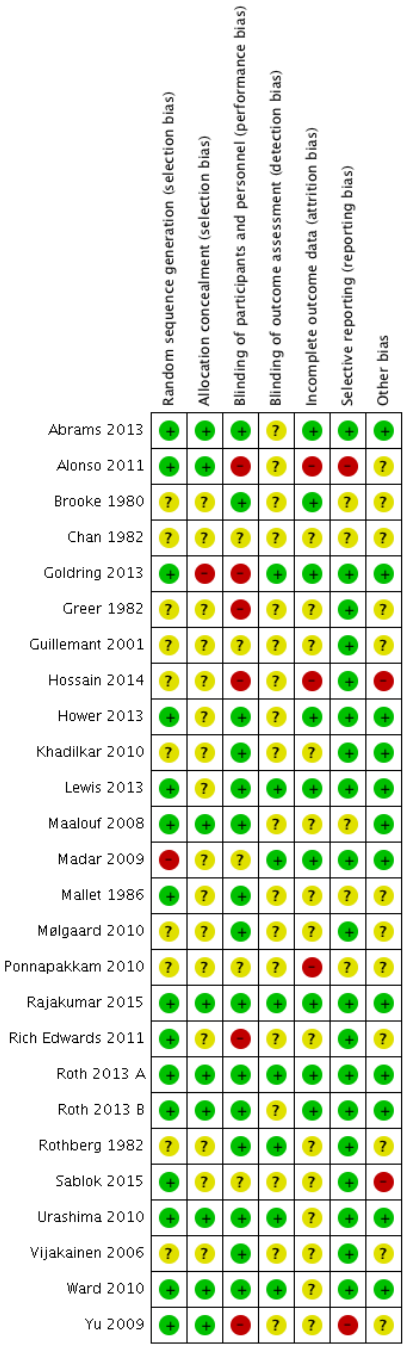
Appendix B. Characteristics of excluded studies

See the list of excluded studies in the following link:

https://www.dropbox.com/s/l9gejv6ntf0gk7r/Appendix%20B.doc?dl=0

Appendix C. Summary of Risk of Bias (RoB) in RCTs and NRS.

- **RoB assessment in RCTs**



- **RoB assessment in NRS**

Allen 2013 (27)

Type of study: Case-control
Participants: Pregnancy women and infants
Experimental intervention: Vitamin D supplementation
Control intervention: No Vitamin D supplementation

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in measurement of interventions***	***Bias due to departures from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	
Food allergy	Benefit	Moderate	Moderate	Moderate	Moderate	Low	Moderate	Moderate	**MODERATE**

Anderson 2015 (26)

Type of study: Cohort
Participants: Pregnancy women and infants
Experimental intervention: Vitamin D supplementation
Control intervention: No Vitamin D supplementation

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in measurement of interventions***	***Bias due to departures from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	

Asthma/wheezing	Benefit	Low	Low	Moderate	Moderate	Low	Moderate	Moderate	**MODERATE**

Bener 2012 (24)

Type of study: Case-control
Participants: Breastfeeding women
Experimental intervention: Vitamin D supplementation
Control intervention: No Vitamin D supplementation

Outcome	**Benefit or harm of intervention**	**Domains**							**Overall Risk of Bias**
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in measurement of interventions***	***Bias due to departures from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	
Asthma	Benefit	Moderate	Serious [1]	Moderate	Serious [2]	Low	Moderate	Moderate	**MODERATE**

1. Controls were sampled from a population that is unlikely to be representative of the population that gave rise to the cases; 2. Switches in treatment, co-interventions, or problems with implementation fidelity are apparent and are not adjusted for in the analyses.

Hypponnen 2014 (35)

Type of study: Case-control
Participants: Infants
Experimental intervention: Vitamin D supplementation
Control intervention: Irregular or No Vitamin D supplementation

Outcome		**Domains**	

	Benefit or harm of intervention	***Bias due to confounding***	***Bias in selection of participants into the study***	***Bias in measurement of interventions***	***Bias due to departures from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	**Overall Risk of Bias**
Asthma	Benefit	Moderate	Low	Moderate	Moderate	Moderate	Moderate	Moderate	**MODERATE**
Allergic rhinitis	Benefit	Moderate	Low	Moderate	Moderate	Moderate	Moderate	Moderate	**MODERATE**

Appendix D. Evidence profiles

All evidence profiles can be found in the following link:

https://www.dropbox.com/s/x6l3g05dze5ur6i/Appendix%20D.docx?dl=0

Appendix E. Forest plots of meta-analysis for each research question

All forest plots can be found in the following link:

https://www.dropbox.com/s/4395vtkoofydmg1/Appendix%20E.doc?dl=0

CHAPTER 5. PHARMACOLOGICAL THROMBOPROPHYLAXIS IN PATIENTS UNDERGOING NEUROSURGICAL INTERVENTIONS FOR PREVENTING VENOUS THROMBOEMBOLISM: A SYSTEMATIC REVIEW OF RANDOMIZED AND NON-RANDOMIZED STUDIES.

PREFACE TO CHAPTER 5

Chapter 5: *Pharmacological thromboprophylaxis in patients undergoing neurosurgical interventions for preventing venous thromboembolism. A systematic review of randomized and non-randomized studies* has been shared among the co-authors for review. The final manuscript will be submitted to Blood Journal.

Chapter 5: Pharmacological thromboprophylaxis in patients undergoing neurosurgical interventions for preventing venous thromboembolism. A systematic review of randomized and non-randomized studies

Authors

Juan José Yepes-Nuñez [1,2], Anita Rajasekhar[3], Maryam Rahman[4], Philipp Dahm [5], David Anderson [6], Luis Enrique Colunga Lozano [1], Stephanie Ross [7], Meha Bhatt [1], Kelly Estrada Orozco [8], Federico Popoff [9], Matthew Ventresca [1], Angela Barbara [1], Sara Balduzzi, [1] Housne Ara [1], Arnav Agarwal [1], Wojtek Wiercioch[1], Robby Nieuwlaat [1], Gian Paolo Morgano [1], Holger J. Schünemann [1,10].

Affiliations

1. Department of Health Research Methods, Evidence, and Impact (HEI), McMaster University, Hamilton, Ontario, Canada
2. Faculty of Medicine, Universidad de Los Andes, Bogotá, Colombia
3. Division of Hematology/Oncology, Department of Medicine, University of Florida, USA

4. Lillian S. Wells Department of Neurosurgery, University of Florida, USA
5. University of Minnesota, Minneapolis VA Health Care System, Urology Section 112D, Minnesota, USA
6. Division of Hematology, Department of Medicine, Department of Community Health and Epidemiology, Department of Pathology, USA
7. Canadian Agency for Drugs and Technologies in Health, Toronto, Canada
8. Clinical Research Institute, National University of Colombia, Bogotá, Colombia
9. Buenos Aires, Argentina
10. Michael DeGroote Cochrane Canada Centre & Department of Medicine, McMaster University, Hamilton, Ontario, Canada

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact (HEI), Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON. L8S 4K1 Canada.

holger.schunemann@mcmaster.ca

Conflict of interest: The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

ABSTRACT

Background: The impact of pharmacological prophylaxis for venous thromboembolism in patients undergoing neurosurgical intervention remains uncertain. **Objective:** To review the efficacy and safety of pharmacological thromboprophylaxis when compared to no pharmacological thromboprophylaxis in neurosurgical patients. **Methods:** Three databases were searched through April 2018, including randomized (RCT) and nonrandomized studies (NRS). Eight reviewers independently extracted data and assessed the certainty of evidence using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. **Results:** Among the 14446 articles identifies, seven RCTs and three NRS proved eligible. Two RCTs suggest that pharmacological thromboprophylaxis may decrease the risk of developing screen-detected proximal deep venous thromboembolism (DVT) (RR 0.50, 95% CI 0.30 to 0.84; low certainty). RCTs left the effect of pharmacological thromboprophylaxis on mortality (RR 1.27; 95% CI 0.57 to 2.86; low certainty), symptomatic pulmonary embolism (PE) (RR 0.84, 95% CI 0.03 to 27.42; very low certainty), screen-detected distal DVT, or reoperation (RR 0.43, 95% CI 0.06 to 2.84, very low certainty) uncertain. Non-randomized studies also revealed uncertainty for mortality (RR 0.72, 95% CI 0.46 to 1.13; very low certainty), and PE (RR 0.18, 95% CI 0.01 to 3.76; very low certainty). Consistent evidence from RCTs and NRS suggests that pharmacological thromboprophylaxis may increase the risk of major bleeding [RCT: RR 1.57, 95% CI 0.70 to 3.50 (low certainty);

and NRS: RR 1.45, 95% CI 0.30 to 7.12 (very low certainty)]. **Conclusion:** In patients undergoing neurosurgical procedures, low certainty evidence suggests pharmacological thromboprophylaxis confer benefit for prevention of screen-detected proximal DVT with large uncertainty regarding impact on other patient-important outcomes.

Keywords: venous thromboembolism; deep venous thrombosis; pulmonary embolism; neurosurgery; thromboprophylaxis; systematic review; GRADE.

Funding/Support: The American Society Hematology founded this project.

HIGHLIGHTS

- To review the impact of pharmacological thromboprophylaxis when compared to no pharmacological thromboprophylaxis in neurosurgical patients on both benefit and harms outcomes summarizing evidence from randomized control trials (RCT) and non-randomized studies (NRS).
- Limited evidence from RCTs and NRS informed the impact of pharmacological thromboprophylaxis in patients underwent neurosurgical interventions.
- Assessment of Risk of Bias (RoB) in RCTs and NRS through the Cochrane RoB tool and ROBINS-I tool offered a rigorous evaluation of the certainty of both resources of evidence as well as a successful integration of it with the GRADE framework for interventions.
- Low certainty of evidence from studies that compared pharmacological thromboprophylaxis with no pharmacological thromboprophylaxis suggests prevention of screen-detected proximal deep venous thromboembolism in patients undergoing neurosurgical interventions.

INTRODUCTION

Venous thromboembolism (VTE) is a common life-threatening complication in patients undergoing neurosurgical procedures. The presence of VTE significantly complicates the delivery of care due to bleeding risks associated with treatment. The risk of VTE in this population varies considerably depending on the patient comorbidities, the neurosurgical procedure performed, method of VTE diagnosis (screening versus symptomatic VTE) and use of different thromboprophylaxis regimens. Although formal VTE risk stratification models do not exist for neurosurgical patients, several VTE risk factors have been identified including active cancer, advanced age, longer duration of surgery, delayed ambulation, inherited thrombophilias, hospital length of stay and paresis (1). On a biochemical level, neurosurgical patients are uniquely susceptible to thrombosis due to release of fibrinopeptide A, fibrinogen, factor VIII, thromboplastin from brain tumors, traumatic spinal and brain injuries, or manipulation of central nervous system tissue during surgery (2).

The incidence of either symptomatic or screen detected deep venous thrombosis (DVT) without prophylaxis in the neurosurgical population has been reported to be as high as 34% (3) in some studies, with a mean of 16% across previous studies (4, 5). The rate of postoperative PE in neurosurgical patients is estimated at 0-4%, with case-fatality ranging from 9-50% (6). The majority of DVT develop within the first week after surgery (7), suggesting that timely postoperative

thromboprophylaxis must be initiated to prevent this potentially devastating complication.

Options for thromboprophylaxis include pharmacologic (e.g. low-dose unfractionated heparin, low-molecular weight heparin) or mechanical measures (e.g. intermittent pneumatic compression devices, graduated compression stockings, or inferior vena cava filters). Randomized controlled trials have evaluated multiple approaches to VTE thromboprophylaxis in neurosurgical patients, each of which requires balancing the risks of bleeding and the benefits in reducing thrombosis. For instance, although mechanical prophylaxis does not increase the risk of bleeding, its effectiveness in reducing VTE is limited (5). In contrast, pharmacological thromboprophylaxis is likely to reduce VTE, but increases the risk of bleeding – of particular concern is intracranial and intraspinal bleeding (4).

The wide variability of thromboprophylaxis protocols in neurosurgical patients is likely due to the paucity of data from well-designed clinical trials and lack of power to clarify safety and efficacy outcomes. Additionally, the ideal initial timing and duration of thromboprophylaxis is a topic of debate amongst neurosurgeons. Prior systematic reviews evaluating neurosurgical VTE prophylaxis have led to conflicting conclusions (4, 5, 8-11). Despite the existing evidence, uncertainty regarding the impact of pharmacological thromboprophylaxis in neurosurgical patients on both benefit and harms outcomes remains.

This systematic review informed the recommendations of the American Society of Hematology-McMaster guidelines for pharmacological thromboprophylaxis in neurosurgical patients. We analyzed comparative studies that examined the effects of either unfractionated heparin (UFH) or low molecular weight heparin (LMWH) or low doses of warfarin vs. non-pharmacological thromboprophylaxis on population and patient-important outcomes in adults who underwent neurosurgical procedures.

METHODS

Inclusion and exclusion criteria

Type of studies

We included randomized control trials and nonrandomized studies (NRS), specifically cohort and case-control studies. We excluded conference abstract reports.

Type of participants

Eligible studies included adult patients undergoing neurosurgical procedures. We excluded studies in which all patients required a neurosurgical intervention following acute trauma.

Type of interventions

We included only studies that reported either unfractionated heparin or low molecular weight heparin or low doses of warfarin vs. placebo, no prophylaxis, or mechanical interventions for VTE prophylaxis (i.e. compression stocking, or intermittent pneumatic compression devices).

Type of outcomes

The outcomes were determined by the ASH guideline panel members following a detailed guideline development process (12). The following outcomes were deemed critical to the decision of whether or not to pharmacological anticoagulation for venous thromboembolism (VTE) prophylaxis:

- Mortality
- Symptomatic pulmonary embolism (PE)
- Screening detected proximal deep venous thrombosis (DVT)
- Screening detected distal DVT
- Major bleeding
- Reoperation

Appendix A summarizes our definitions for pulmonary embolism, DVT, and major bleeding outcomes.

Search methods

We searched MEDLINE, EMBASE, and The Cochrane Central Register of Controlled Trials (CENTRAL) from inception to April 2018 without restrictions on language of publication. Appendix B presents search strategies for RCTs and NRS. We checked reference lists of reviewed articles and contacted clinical experts for additional references.

Study selection and data extraction

Ten independent evaluators, working in pairs (LC, SR, FP, MB, KE, MV, AB, SB, HA, AA). screened the titles and abstracts obtained through the electronic searches. If a study was deemed potentially relevant, we obtained the full-text, and the same ten evaluators, again working in pairs, made final eligibility decisions. If reviewers could not resolve disagreement through discuss, a third reviewer adjudicated (JYN). We contacted the authors if additional information or clarification was needed.

Eight independent evaluators (LC, SR, FP, MB, MV, AB, HA), working in pairs, read all reports of eligible studies in detail and summarized the pertinent details in a standard data extraction sheet (type of study; methodology; characteristics of participants, results, and outcome measurements; as well as an evaluation of the risk of bias or study limitations). Reviewers discussed any disagreements and consulted a third reviewer if necessary (JYN).

Certainty in the body of evidence (confidence in the estimates of effects or certainty of evidence)

Assessment of risk of bias in included studies

Two review authors (JYN, MB) independently assessed risk of bias for each randomized control trial and nonrandomized studies using the criteria outlined in the Cochrane Handbook for Systematic Reviews of Intervention (13), and version 1.0 of the Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions ROBINS-I (14).

Data analysis

Measure of treatment effect

We calculated relative risk (RR) and 95% confidence intervals (95% CI) for dichotomous outcomes. For NRS, we calculated the measure of the relative effect based on information reported by authors. We selected baseline risks (control group risk) to calculate absolute effects measures from a systematic review that estimated the incidence of thromboembolic disease in postoperative spinal patients (9). We estimated absolute effect estimates to facilitate the decision-making process for the effects of pharmacological thromboprophylaxis in individual neurosurgical patients by American Society of Hematology guideline group members (15).

Dealing with missing data

We contacted authors where details about study design or descriptive statistics for outcomes were not presented in original papers. We did not impute data.

Assessment of statistical heterogeneity and inconsistency

We assessed inconsistency between studies by visual inspection of forest plots, in particular extent of overlap of confidence intervals (CI), the Q statistic (with a p value ≤ 0.05 as a suggestion of important statistical heterogeneity), and the I^2 value. We planned to explore reasons for inconsistency by prespecified differences in type of intervention.

Assessment of reporting biases

We planned, if ten or more studies were available for a particular outcome, to create a funnel plot to assess publication bias by visual inspection.

Data synthesis

We assessed the treatment effect using the mean difference for continuous outcomes, and the RR for dichotomous outcomes for individual studies, and used random-effects models to meta-analyse study data. We presented all measures with 95% CI. We carried out all statistical analyses using Review Manager 5.3 (16).

Subgroup analysis

We considered one main group for subgroup analysis based on the type of neurosurgical intervention (craniotomy vs. spinal surgery).

Sensitivity analysis

We planned analyses to determine the effect of including or excluding the studies with high risk of bias on treatment effect estimates.

GRADE assessment of the overall certainty in the body of evidence by outcome

We used the Grading of Recommendation, Assessment, Development, and Evaluation (GRADE) methodology to rate the certainty in the body of evidence for each outcome as high, moderate, low, or very low (17). The assessment included judgments addressing risk of bias, imprecision, inconsistency, indirectness, and publication bias. We also rated the certainty in the body of evidence using a GRADE approach for observational studies (18). We created Evidence Profile and Summary of Findings Tables for each population using GRADE's electronic tool GRADEpro GDT (www.gradepro.org). To assess the usefulness of including NRS, we applied the GRADE guidance using ROBINS-I as a part of GRADE's certainty rating process (19). A senior methodologist checked all GRADE tables and ratings of the certainty in the body of evidence (HJS).

RESULTS

The search of the MEDLINE, EMBASE, CENTRAL yielded 10538 unique records after duplicates removed. Figure 1 shows the PRISMA flow diagram (20). We reviewed the full text of 111 publications in detail.

Overall, 7 RCTs (12, 13, 21-25), and 3 NRS (retrospective cohort studies) (16, 26, 27) met our inclusion criteria. Five RCTs reported the effect of pharmacological thromboprophylaxis vs. no pharmacological intervention on development of mortality (12, 21-23, 25); 3 RCTs reported development of symptomatic PE (13, 21, 24), 2 RCTs addressed symptomatic DVT through development of screen-detected proximal DVT (12, 21) and 1 through screen-detected distal DVT (21), 7 RCTs reported risk of major bleeding (12, 13, 21, 22, 24, 25), and 2 RCTs reported risk of reoperation (22, 25). The three NRS reported information about mortality (26, 27), pulmonary embolism (26, 27), and major bleeding (16, 26, 27).

Three RCTs (3, 28, 29), reported the incidence of DVT regardless of whether it was symptomatic proximal or distal. Since these 3 RCTs did not report findings for the outcomes included in the guideline, we did not incorporate this body of evidence in our meta-analysis.

One NRS (7) reported qualitatively no difference in the rate of hemorrhagic complications and PE between neurosurgical patients who received

pharmacological thromboprophylaxis with UFH compared with control patients who did not receive pharmacological thromboprophylaxis. Since this information was not reported quantitatively, we could not combine these data in our meta-analysis.

Another NRS (30) reported PE in symptomatic and asymptomatic patients with subarachnoid hemorrhage and ventriculostomies, and the number of DVTs for both proximal and distal DVT together. Outcomes were measured through helical CT angiogram by protocol and venous ultrasound, respectively. In the group receiving pharmacological thromboprophylaxis, 6 out of 138 patients developed PE compared to none out of 53 patients in the no pharmacological thromboprophylaxis group. Cases of DVT were less frequent in patients who received pharmacological thromboprophylaxis compare with those who did not receive it (28 of 188 patients vs. 4 of 53, respectively). These data were not combined in the meta-analysis because it included symptomatic (non-screen detected) and asymptomatic (screen-detected) neurosurgical patients for both DVT and PE outcomes.

No additional information from the included reports was required; therefore, we did not contact authors. Table 1 presents the characteristics of the ten eligible studies according to the study design. In Appendix C we report the list of studies excluded.

Risk of bias in studies included

Overall risk of bias was deemed low to very low in individual RCTs (Appendix D). For RCTs, the most significant concerns were related to incomplete outcome data and blinding of participants and personnel. Regarding the NRS, we evaluated risk of bias for four outcomes that were reported in five studies (Appendix E). All four outcomes presented very serious risk of bias. We found inappropriate methods to control for all important confounding domains across all outcomes in all NRS.

Effect of interventions

Table 2 describes the summary of findings for all research questions. Appendix F presents the full evidence profile with more detailed explanations. Appendix G presents forest plots of RCTs and NRS.

Mortality

Five RCTs (12, 21-23, 25) reported the effect of pharmacological thromboprophylaxis on mortality. No effect of pharmacological prophylaxis was found on mortality (RR, 1.27; 95% CI 0.57 to 2.86). We also found two NRS (27, 29) that reported mortality in patients with movement disorders who underwent deep brain stimulation (DBS) surgery, and postoperative patients admitted to surgical intensive care unit. In one NRS (26), investigators observed no deaths in either patients exposed to unfractionated heparin (UFH) or in patients non-

exposed to pharmacological thromboprophylaxis. In the other NRS (27), pharmacological prophylaxis with UFH did not reduce mortality in neurosurgical patients (RR, 0.72; 95% CI 0.46 to 1.13). The overall certainty in the body of evidence ranged from low to very low due to the risk of bias, inconsistency, and imprecision.

Symptomatic pulmonary embolism (PE)

Three RCTs (13, 21, 24) reported symptomatic PE. The evidence was essentially uninformative with respect to relative effects on PE with extremely wide confidence intervals (RR, 0.84; 95% CI 0.03 to 27.42). One NRS (26) reported the development of PE in 2 out of 121 neurosurgical patients who were not exposed to UFH. None of the patients who received UFH developed PE; again, this is uninformative with respect to relative effects (RR, 0.18; 95% CI 0.01 to 3.76). In the other NRS (27), no cases of PE were observed in 522 neurosurgical patients. In absolute effect terms, the difference between groups was 4 fewer events of PE per thousand patients with CI from 5 fewer in 1000 to 13 more in 1000 patients. The overall certainty in the body of evidence was very low owing to the risk of bias, inconsistency, and imprecision.

Screening-detected proximal DVT

Two RCTs reported (12, 21) screening-detected proximal DVT. Pharmacological thromboprophylaxis was associated with a reduction of proximal DVT (RR, 0.50;

95% CI 0.30 to 0.84). No information about screening-detected proximal DVT was reported in NRS. The certainty of the body of evidence was low due to risk of bias and indirectness because the outcome important to patients is symptomatic DVT.

Screening-detected distal DVT

One RCT (21) addressed the detection of distal DVT by screening when pharmacological thromboprophylaxis was administered in patients who underwent elective cranial or spinal surgery. No convincing effect of pharmacological thromboprophylaxis was found for distal DVT (RR, 0.54; 95% CI 0.27 to 1.08). None NRS addressed this outcome. The overall certainty of evidence was very low owing to risk of bias, indirectness (because the outcome of interest is symptomatic DVT), and imprecision.

Data from the six RCTs (12, 13, 23, 25, 28, 29) that reported screening-detected proximal or distal DVT were pooled separately. Pharmacological thromboprophylaxis with UFH or LMWH in neurosurgical patients reduced the risk of developing proximal or distal DVT (RR, 0.67; 95% CI 0.49 to 0.91). Two NRS (16, 30) reported a benefit of pharmacological thromboprophylaxis in reducing any screening-detected proximal or distal DVT events (RR, 0.39; 95% CI 0.20 to 0.78).

Major bleeding

Ten studies, 7 RCTs (12, 13, 21-25) and 3 NRS (16, 26, 27), reported on major bleeding. In the 7 RCTs, pharmacological thromboprophylaxis in patients who underwent neurosurgical procedures may increase major bleeding compared to no pharmacological thromboprophylaxis (RR, 1.57; 95% CI 0.70 to 3.50). In two out (16, 26) of 3 NRS, no cases of major bleeding were reported in exposed and non-exposed neurosurgical patients to pharmacological thromboprophylaxis. One NRS (27) also reported on a very small number of neurosurgical patients who developed major bleeding between those who either exposed or not exposed to pharmacological thromboprophylaxis (RR, 1.45; 95% CI 0.30 to 7.12). In absolute terms, the difference between groups was 3 more events of major bleeding per 1000 patients with CI from 5 fewer in 1000 to 40 more in 1000 patients. The certainty in the body of evidence of the risk of major bleeding ranges from low to very low due to risk of bias and imprecision.

Reoperation

Two RCTs (22, 25) evaluated the risk of reoperation after pharmacological thromboprophylaxis. We were very uncertain in terms of relative effects, with confidence intervals including relative risk reductions of more than 50% and almost three-fold increases (RR, 0.43; 95% CI 0.06 to 2.84). In absolute terms, the difference between groups was 18 fewer events of reoperation per 1000 patients with CI from 29 fewer in 1000 to 57 more in 1000 patients. No NRS

addressed this outcome. The certainty in the body of evidence was very low because of the risk of bias and imprecision.

Subgroup analysis

Appendix H presents the results of the subgroup analysis for the type of neurosurgical intervention in RCTs. We carried out subgroup analysis for the outcomes for which evidence was available: mortality, pulmonary embolism, major bleeding, and reoperation. None of the analyses supported the existence of a subgroup effect. Sparse data from NRS did not allow us to conduct a subgroup analysis.

Sensitivity analysis

The number of studies in each particular population for any individual outcome proved insufficient to conduct sensitivity analyses.

DISCUSSION

Findings from 7 RCTs and 3 NRS provide low to very low-certainty evidence for a reduction in screening detected DVT with pharmacological thromboprophylaxis in neurosurgical patients but it may increase the risk of major bleeding. In these RCTs and NRS, results suggest that pharmacological thromboprophylaxis possibly prevents symptomatic VTE, with uncertain impact on mortality, symptomatic PE, and reoperation.

Specifically, two RCTs provided indirect low-certainty evidence of a benefit of pharmacological thromboprophylaxis for prevention of symptomatic DVT in neurosurgical patients through screening-detected proximal DVT (RR, 0.50; 95% CI 0.30 to 0.84). We did not find a difference in effect in subgroups of population based on the type of surgery.

Our systematic review has several strengths. We established explicit eligibility criteria, conducted a comprehensive search, assessed eligibility and extracted data in duplicate. Our list of outcomes is more comprehensive and more specific to VTE thromboprophylaxis than all previous systematic reviews. We assessed the risk of bias in NRS using the ROBINS-I tool, which allowed us to identified domain areas compromised in the NRS. For the first time, the ROBINS-I assessment was integrated with the GRADE RoB criteria applying a framework recently reported in GRADE guideline (19). We also rated the certainty of the body of evidence using guidance from the GRADE working group. Moreover, we included both RCTs and NRSs that compared pharmacological thromboprophylaxis with a control group to broaden the results and applicability of our findings in both cranial and spinal neurosurgeries.

Limitations

This review also has several limitations inherent to the evidence. We found only seven RCTs, and three NRS, that addressed prevention of VTE after exposure to pharmacological thromboprophylaxis in neurosurgical patients. Two reasons

explain these findings. First, we had focused on specific VTE outcomes for the VTE prevention in neurosurgical patients. Specifically, we included data from studies that reported symptomatic PE and screening-detected proximal or distal DVT. Thus, we excluded studies that addressed pharmacological thromboprophylaxis in patient population with screening-detected pulmonary embolism. We did not find studies that addressed pharmacological thromboprophylaxis for proximal or distal symptomatic DVT in neurosurgical patients. In part because of the resultant sample size limitations, the evidence provides only low to very low confidence in estimates for our outcomes. Serious considerations about the susceptibility to bias and imprecision of the meta-analysis estimates (table 2) were noted in most of the studies.

Findings in other publications

Other systematic reviews have explored the association between pharmacological prophylaxis for VTE in neurosurgical patients. In 2008, Collen et al. (5) identified 30 studies (18 RCTs and 12 prospective cohort studies) that compared pharmacological thromboprophylaxis of UFH, LMWH or low dose warfarin to one another or to mechanical thromboprophylaxis [intermittent compression devices (ICD) or compression stockings (CS)], or placebo. For DVT, ICD and LMWH were associated with reducing DVT events compared to placebo (RR, 0.41; 95% CI 0.21 to 0.78) and CS (RR, 0.60; 95% CI, 0.44 to 0.81 respectively). No other comparisons showed differences in either DVT or PE.

Estimates for the comparisons LMWH vs. UFH, LMWH vs. nonpharmacologic management, and UFH vs. placebo were also not associated with intracranial hemorrhage (ICH), minor bleeding, major bleeding, or death. However, rates of minor bleeding were lower when patients did not receive heparin perioperatively (0.04 per 1,000 patients; 95% CI, 0.00 to 3.7).

For three reasons, Collen's systematic review included more studies than we included. First, Collen's included comparisons than we did not consider. For example, the comparisons of mechanical thromboprophylaxis with placebo or LMWH with UFH. We did not consider mechanical thromboprophylaxis interventions with placebo as mechanical thromboprophylaxis tends to be poorly tolerated and, thus, associated with a high degree of non-compliance (31). We also did not include the comparison of LMWH with UFH because it was not prioritized by the guideline panel.

Second, the prospective cohort studies included by Collen's did not have a non-exposure group to compare against the thromboprophylaxis intervention. In contrast, we included only NRS that reported a comparison or control group for VTE pharmacological thromboprophylaxis.

Third, we conducted our analysis data for outcomes that reported DVT by either screening approaches or clinical validation for proximal and distal DVT and symptomatic PE whereas Collen’s included studies that reported DVT and PE by any objective assessment without prespecifying patient population

characteristics. We chose this approach because the panel felt that it would increase their ability to make judgments. Thus, our review had direct clinical relevance based on guideline panel members' priorities. We also identified RCTs that were not reported in Collen's systematic review.

In 2011, Hamilton et al. (11) conducted a systematic review to evaluate the efficacy and safety of low doses of UFH or LMWH compared to placebo for VTE thromboprophylaxis in patients who underwent cranial neurosurgery. The systematic review found 6 RCTs that compared heparin (UFH or LMWH) vs. placebo or mechanical methods. We identified these 6 RCTs in our systematic survey. Hamilton's systematic review reported findings as symptomatic or asymptomatic VTE. They concluded that heparin prophylaxis reduces the risk of symptomatic or asymptomatic VTE (RR, 0.58; 95% CI, 0.45 to 0.75) without association with an increase for intracranial hemorrhage (ICH) (RR, 1.48; 95% CI, 0.63 to 3.44), extracranial major bleeding (RR, 0.91; 95% CI, 0.38 to 2.17) but with an increase in minor bleeding (RR, 2.28; 95% CI, 1.02 to 5.10). Authors of this review used the Jadad scale to assess the quality of individual RCTs. Similar to Cohen's, the outcome reporting of this systematic review differed from our approach. Moreover, Hamilton et al. did not include spinal neurosurgical procedures in the review. Our work included a broader spectrum of neurosurgical procedures as well as reporting findings for specific components of VTE: proximal and distal DVT, and PE.

Khan et al. in 2017 (4) carried out a systematic review for VTE thromboprophylaxis in patients undergoing cranial or spinal neurosurgeries. They found five RCTs that we also identified in our systematic review. They reported a significant benefit of pharmacological thromboprophylaxis to reduce DVT compared to placebo (OR, 0.51; 95, CI 0.37 to 0.71). Safety outcomes did not show significant increase of major ICH (OR, 1.42; 95%, CI 0.61 to 3.30), major extracranial hemorrhage (OR, 0.98; 95% CI, 0.29 to 3.36), minor bleeding complications (OR, 1.28; 95%, CI 0.50 to3.24), and spinal hemorrhage complications. Again, the authors combined symptomatic or screening detected DVT together, and did not divide proximal or distal DVT.

Study implications

Our systematic review provides an update of pharmacological thromboprophylaxis for patients with cranial and spinal neurosurgery that includes evidence from RCT and NRS. We found low certainty of evidence, that pharmacological thromboprophylaxis reduces symptomatic DVT established through indirect evidence from screening-detected proximal DVT in neurosurgical patients. We did not find a convincing effect of pharmacological thromboprophylaxis on any other VTE outcomes.

Suggestions for future research

Further research must be carried out to increase the certainty in the body of evidence including greater precision of the findings. New RCTs should also increase its certainty not only in aspects of randomization but areas of blinding participants and personal as well as improving outcomes reporting. In particular, RCTs should focus on symptomatic rather than asymptomatic VTE. Additional RCTs may help to evaluate the effects of using pharmacological thromboprophylaxis over other VTE patient-important outcomes with and without mechanical thromboprophylaxis.

CONCLUSION

Different pharmacological and non-pharmacological strategies have been proposed for VTE thromboprophylaxis in neurosurgical patients. The current evidence shows that pharmacological thromboprophylaxis reduces the developing of screening-detected proximal DVT in neurosurgical patients. However, the certainty of the evidence was low because of risk of bias, indirectness and imprecision. Future trial studies may be carried out to evaluate the impact of pharmacological thromboprophylaxis in other beneficial VTE outcomes as well as to determine the harmful effects of these interventions.

REFERENCES

1. Rolston JD, Han SJ, Bloch O, Parsa AT. What clinical factors predict the incidence of deep venous thrombosis and pulmonary embolism in neurosurgical patients? J Neurosurg. 2014;121(4):908-18.

2. Ganau M, Prisco L, Cebula H, Todeschi J, Abid H, Ligarotti G, et al. Risk of Deep vein thrombosis in neurosurgery: State of the art on prophylaxis protocols and best clinical practices. J Clin Neurosci. 2017;45:60-6.

3. Cerrato D, Ariano C, Fiacchino F. Deep vein thrombosis and low-dose heparin prophylaxis in neurosurgical patients. J Neurosurg. 1978;49(3):378-81.

4. Khan NR, Patel PG, Sharpe JP, Lee SL, Sorenson J. Chemical venous thromboembolism prophylaxis in neurosurgical patients: an updated systematic review and meta-analysis. J Neurosurg. 2017:1-10.

5. Collen JF, Jackson JL, Shorr AF, Moores LK. Prevention of venous thromboembolism in neurosurgery: a metaanalysis. Chest. 2008;134(2):237-49.

6. Hamilton MG, Hull RD, Pineo GF. Venous thromboembolism in neurosurgery and neurology patients: a review. Neurosurgery. 1994;34(2):280-96; discussion 96.

7. Khaldi A, Helo N, Schneck MJ, Origitano TC. Venous thromboembolism: deep venous thrombosis and pulmonary embolism in a neurosurgical population. J Neurosurg. 2011;114(1):40-6.

8. Iorio A, Agnelli G. Low-molecular-weight and unfractionated heparin for prevention of venous thromboembolism in neurosurgery: a meta-analysis. Arch Intern Med. 2000;160(15):2327-32.

9. Glotzbecker MP, Bono CM, Wood KB, Harris MB. Thromboembolic disease in spinal surgery: a systematic review. Spine (Phila Pa 1976). 2009;34(3):291-303.

10. Schuster JM, Fischer D, Dettori JR. Is chemical antithrombotic prophylaxis effective in elective thoracolumbar spine surgery? Results of a systematic review. Evid Based Spine Care J. 2010;1(2):40-5.

11. Hamilton MG, Yee WH, Hull RD, Ghali WA. Venous thromboembolism prophylaxis in patients undergoing cranial neurosurgery: a systematic review and meta-analysis. Neurosurgery. 2011;68(3):571-81.

12. Nurmohamed MT, van Riel AM, Henkens CM, Koopman MM, Que GT, d'Azemar P, et al. Low molecular weight heparin and compression stockings in the prevention of venous thromboembolism in neurosurgery. Thromb Haemost. 1996;75(2):233-8.

13. Rokito SE, Schwartz MC, Neuwirth MG. Deep vein thrombosis after major reconstructive spinal surgery. Spine (Phila Pa 1976). 1996;21(7):853-8; discussion 9.

14. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

15. Guyatt G, Oxman A, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66(2):158-72.

16. Dermody M, Alessi-Chinetti J, Iafrati MD, Estes JM. The utility of screening for deep venous thrombosis in asymptomatic, non-ambulatory neurosurgical patients. J Vasc Surg. 2011;53(5):1309-15.

17. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6.

18. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6.

19. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2018.

20. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-12.

21. Agnelli G, Piovella F, Buoncristiani P, Severi P, Pini M, D'Angelo A, et al. Enoxaparin plus compression stockings compared with compression stockings

alone in the prevention of venous thromboembolism after elective neurosurgery. N Engl J Med. 1998;339(2):80-5.

22. Constantini S, Kanner A, Friedman A, Shoshan Y, Israel Z, Ashkenazi E, et al. Safety of perioperative minidose heparin in patients undergoing brain tumor surgery: a prospective, randomized, double-blind study. J Neurosurg. 2001;94(6):918-21.

23. Dickinson LD, Miller LD, Patel CP, Gupta SK. Enoxaparin increases the incidence of postoperative intracranial hemorrhage when initiated preoperatively for deep venous thrombosis prophylaxis in patients with brain tumors. Neurosurgery. 1998;43(5):1074-81.

24. Gruber UF, Rem J, Meisner C, Gratzl O. Prevention of thromboembolic complications with miniheparin-dihydroergotamine in patients undergoing lumbar disc operations. Eur Arch Psychiatry Neurol Sci. 1984;234(3):157-61.

25. Hamidi S, Riazi M. Incidence of venous thromboembolic complications in instrumental spinal surgeries with preoperative chemoprophylaxis. J Korean Neurosurg Soc. 2015;57(2):114-8.

26. Bauman JA, Church E, Halpern CH, Danish SF, Zaghloul KA, Jaggi JL, et al. Subcutaneous heparin for prophylaxis of venous thromboembolism in deep brain stimulation surgery: evidence from a decision analysis. Neurosurgery. 2009;65(2):276-80; discussion 80.

27. Hacker RI, Ritter G, Nelson C, Knobel D, Gupta R, Hopkins K, et al. Subcutaneous heparin does not increase postoperative complications in neurosurgical patients: An institutional experience. J Crit Care. 2012;27(3):250-4.

28. Halim TA, Chhabra HS, Arora M, Kumar S. Pharmacological prophylaxis for deep vein thrombosis in acute spinal cord injury: an Indian perspective. Spinal Cord. 2014;52(7):547-50.

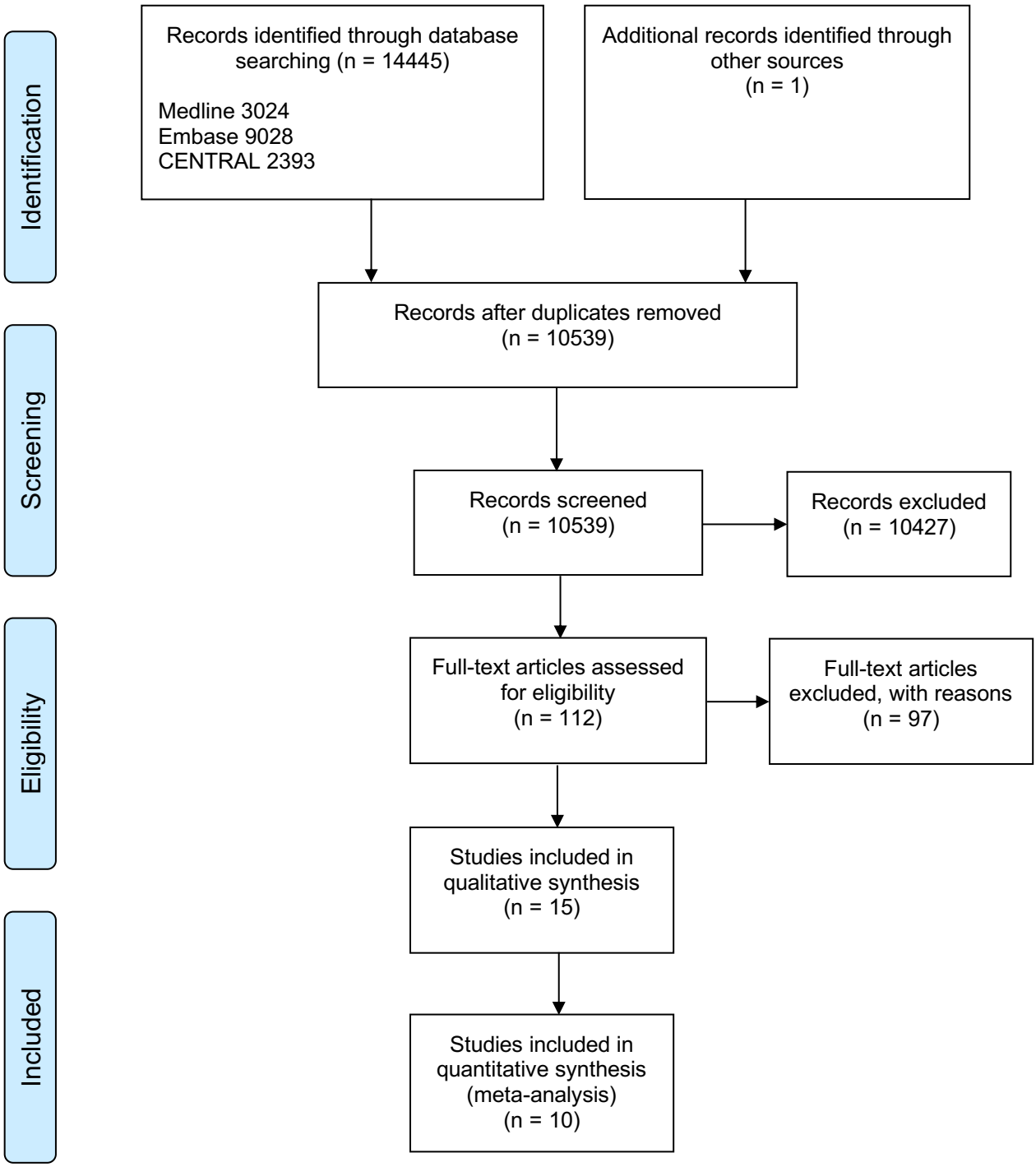
29. Sonaglia F, Agnelli G, Baroni M, Severi P, Quintavalla R, D'Angelo SV. Pre-operative plasma levels of soluble fibrin polymers correlate with the development of deep vein thrombosis after elective neurosurgery. Blood Coagul Fibrinolysis. 1999;10(8):459-63.

30. Zachariah J, Snyder KA, Graffeo CS, Khanal DR, Lanzino G, Wijdicks EF, et al. Risk of Ventriculostomy-Associated Hemorrhage in Patients with Aneurysmal Subarachnoid Hemorrhage Treated with Anticoagulant Thromboprophylaxis. Neurocrit Care. 2016;25(2):224-9.

31. Faraoni D, Comes RF, Geerts W, Wiles MD, Force EVGT. European guidelines on perioperative venous thromboembolism prophylaxis: Neurosurgery. Eur J Anaesthesiol. 2018;35(2):90-5.

FIGURE

Figure 1. PRISMA flow chart of study selection



TABLES

Table 1. Characteristics of studies included

Study, Year	n	Population description	Type of neurosurgery	Pharmacological thromboprophylaxis	Co-interventions	Route, doses, and time of pharmacological thromboprophylaxis	Time to follow-up	DVT diagnostic technique	Outcomes assessed	Included in meta-analysis
Randomized controlled trials										
Agnelli, 1998 (21)	307 *	Patients 18 years of age or older	Elective cranial or spinal surgery	LMWH (Enoxaparin)	Thigh-length TED compression stocking in all patients	Subcutaneous injections, 40 mg per day for 8±1 days	60 days	Ultrasonography and venography in case of positive ultrasonography.	M, PE, proximal DVT, distal DVT, MB	Yes
Cerrato, 1978 (3)	100 *	Patients over 40 years of age	Elective neurosurgical procedures	UFH	No reported	Subcutaneous injections, 5000 units every 8 hours for at least 7 days	No reported	^{125}I-labeled fibrinogen	Any DVT†	No
Constantini, 2001 (22)	103 *	Patients over 40 years of age	Craniotomy for brain tumor removal	UFH	No reported	Subcutaneous injections, 5000 units every 12 hours for 7 days or until full ambulation	No reported	No reported**	M, MB, RO	Yes
Dickinson, 1998 *** (23)	66*	Patients 18 years of age or older with a diagnosis of an intracranial neoplasm	Craniotomy or stereotactic biopsy	LMWH (Enoxaparin)	Thigh-high TED compression stocking in all patients all patients only before randomization	Subcutaneous injections, 30 mg every 12 hours day until hospital discharge from neurosurgery service	No reported	Duplex ultrasonographic	M, MB	Yes
Gruber, 1984 (24)	50*	Adult patients undergoing lumbar disc operations	Herniated lumbar disc operations	Heparin-DHE	No reported	Subcutaneous injections, 2500 IU-0.5 mg every 12 hours for at least 7 days or until hospital discharge	No reported	Phlebogram, plethysmography, Doppler ultrasound or an I^{125} fibrinogen test	PE, MB	Yes
Halim, 2014 (28)	74*	Adult patients with ASCI	Anterior and/or posterior surgical approach	LMWH (Enoxaparin)	Compression stocking in all patients	Subcutaneous injections, 40 mg per day for 8 weeks	2 weeks	Color doppler venous ultrasonography	Any DVT†	No
Hamidi, 2015 (25)	89*	Patients aged 18 to 75 years	Elective instrumental spinal surgery	LMWH (Enoxaparin)	Compression stocking in all patients	Subcutaneous injections, 40 mg per	2 weeks and 8 months after surgery	Compression doppler ultrasonography	M, MB, RO	Yes

						day within 12 hours before the surgery				
Nurmohamed, 1996 (12)	485 *	Patients 18 years of age or older	Craniotomy or spinal column surgery for a tumor or injury	LMWH (Nadroparin calcium)	TED compression stocking in all patients	Subcutaneous injections, 7,500 anti-factor Xa per day for 10 days or until hospital discharge	56 days after surgery	B-mode compression ultrasonography and venogram in case of ultrasonography positive	M, proximal DVT, MB	Yes
Rokito, 1996 *** (13)	110 *	Patients 18 years of age or older	Major reconstructive spinal surgery	Warfarin (Coumadin)	Thigh-high TED compression stocking in all patients	10 mg before surgery and doses were adjusted accordingly to prothrombin time	1 year	Duplex ultrasonography and venography in case of ultrasonography positive	PE, MB	Yes
Sonaglia, 1999 (29)	157 *	Patients 18 years of age or older	Neurosurgery for brain or spinal tumour	LMWH (Enoxaparin)	Compression stocking in all patients	Subcutaneous injections, 40 mg per day for at not less than 7 days	No reported	Venography	Any DVT†	Yes
Non-randomized studies										
Bauman, 2009 (26)	254	Patients with movement disorders (Parkinson disease, essential tremor, dystonia)	Deep brain stimulation (DBS) surgery	UFH	Compression stocking in all patients and pneumatic compression boots postoperativel y	Subcutaneous injections, 50 mg before surgery and 50 mg every 12 hours after surgery. Duration of pharmacological thromboprophylaxis no reported	No reported	Doppler ultrasonography	M, PE, MB	Yes
Dermody, 2011 (16)	174	Neurosurgical patients who underwent screening with weekly VDUS of the bilateral lower extremities	Endovascular coiling or clipping, craniotomy, stereotactic biopsy, spine surgery, trans sphenoidal surgery	UFH or LMWH (Enoxaparin)	Mechanical prophylaxis	UFH: 5000 two or three times daily Enoxaparin: no reported	6 months	Venous duplex ultrasound	MB	Yes
Hacker, 2012 (27)	522	Neurosurgical and head trauma patients	Cervical spinal cord decompression, cervical laminectomy, craniotomy/crani ectomy, decompressive	UFH	Lower extremity compression boots	Subcutaneous injections, 5000 IU every 8 hours until hospital discharge	Until death or until discharged from the hospital	Ultrasound evaluation	M, PE, MB	Yes

			laminectomy, nasal sinuses surgery							
Khaldi, 2011 (7)	2638	Patients who underwent a neurosurgical procedure	Multilevel lumbar surgeries, major spine surgery, head surgery	UFH	Mechanical DVT prophylaxis	Subcutaneous injections, 5000 IU every 12 hours	No reported	Duplex ultrasonography	PE, MB‡	No
Zachariah, 2016 (30)	241	Patients with subarachnoid hemorrhage and external ventricular drain	No reported	UFH, LMWH, or warfarin	Sequential compression devices	No reported	No reported	Venous ultrasound	PE, Any DVT†	No

* Number of patients randomized
** Authors reported DVT only by clinical evidence
*** Three-arm RCT
† Reported as proximal or distal DVT
‡ Findings were reported qualitatively
ASCI: acute spinal cord injury; DHE: dihydroergotamine; DVT: deep venous thrombosis; LMWH: low molecular weight heparin; M: mortality; MB: major bleeding; PE: pulmonary embolism; RO: reoperation; SCD: sequential compression device; TED: thrombosis embolic deterrent; UFH: unfractionated heparin; VDUS: venous duplex ultrasound.

Table 2. Summary of Finding table

Pharmacological prophylaxis compared to no pharmacological prophylaxis in patients undergoing major neurosurgical procedures

Outcomes	№ of participants (studies) Follow-up	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no pharmacological prophylaxis	Risk difference with pharmacological prophylaxis
Mortality - RCTs	1029 (5 RCTs)	⨁⨁◯◯ LOW [a,b]	RR 1.27 (0.57 to 2.86)	35 per 1,000	9 more per 1,000 (15 fewer to 65 more)
Mortality - NRS	674 (2 observational studies)	⨁◯◯◯ VERY LOW [c,d,e]	RR 0.72 (0.46 to 1.13)	115 per 1,000	32 fewer per 1,000 (62 fewer to 15 more)
Symptomatic Pulmonary Embolism - as described by the moderate marker state - RCTs assessed with: Symptomatic PE	434 (3 RCTs)	⨁◯◯◯ VERY LOW [f,g,h]	RR 0.84 (0.03 to 27.42)	Study population	
				14 per 1,000	2 fewer per 1,000 (13 fewer to 359 more)
				Low	
				2 per 1,000 [i]	0 fewer per 1,000 (2 fewer to 53 more)
				Study population	

Pharmacological prophylaxis compared to no pharmacological prophylaxis in patients undergoing major neurosurgical procedures

Outcomes	№ of participants (studies) Follow-up	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no pharmacological prophylaxis	Risk difference with pharmacological prophylaxis
Symptomatic Pulmonary Embolism - as described by the moderate marker state - NRS assessed with: Symptomatic PE	776 (2 observational studies)	⨁◯◯◯ VERY LOW [c,h,j]	RR 0.18 (0.01 to 3.76)	5 per 1,000	4 fewer per 1,000 (5 fewer to 13 more)
				Low	
				2 per 1,000 [i]	2 fewer per 1,000 (2 fewer to 6 more)
Symptomatic Proximal Deep Vein Thrombosis - as described by the moderate marker state - RCTs assessed with: Screening detected Proximal DVT	744 (2 RCTs)	⨁⨁◯◯ LOW [k,l]	RR 0.50 (0.30 to 0.84) [m]	Study population	
				113 per 1,000	56 fewer per 1,000 (79 fewer to 18 fewer) [m]
				Based on study population BLR	
				3 per 1,000 [n]	2 fewer per 1,000 (2 fewer to 1 fewer)
				Low	
				3 per 1,000 [o,p]	2 fewer per 1,000 (2 fewer to 1 fewer)

Pharmacological prophylaxis compared to no pharmacological prophylaxis in patients undergoing major neurosurgical procedures

Outcomes	№ of participants (studies) Follow-up	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no pharmacological prophylaxis	Risk difference with pharmacological prophylaxis
Symptomatic Distal Deep Vein Thrombosis - as described by the severe marker state assessed with: Screening detected Distal DVT	259 (1 RCT)	⨁◯◯◯ VERY LOW [l,q,r]	RR 0.60 (0.33 to 1.08) [m]	Study population	
				194 per 1,000	78 fewer per 1,000 (130 fewer to 16 more) [m]
				Based on study population BLR	
				2 per 1,000 [s]	1 fewer per 1,000 (1 fewer to 0 fewer)
				Low	
				1 per 1,000 [o,p]	0 fewer per 1,000 (0 fewer to 0 fewer)
Major Bleeding - RCTs	1156 (7 RCTs)	⨁⨁◯◯ LOW [t,u]	RR 1.57 (0.70 to 3.50)	17 per 1,000	10 more per 1,000 (5 fewer to 43 more)
Major Bleeding - NRS	930 (3 observational studies)	⨁◯◯◯ VERY LOW [v,w]	RR 1.45 (0.30 to 7.12)	7 per 1,000	3 more per 1,000 (5 fewer to 40 more)

Pharmacological prophylaxis compared to no pharmacological prophylaxis in patients undergoing major neurosurgical procedures

Outcomes	№ of participants (studies) Follow-up	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no pharmacological prophylaxis	Risk difference with pharmacological prophylaxis
Reoperation - RCTs	192 (2 RCTs)	⨁◯◯◯ VERY LOW [x,y]	RR 0.43 (0.06 to 2.84)	31 per 1,000	18 fewer per 1,000 (29 fewer to 57 more)

*The risk in the intervention group (and its 95% confidence interval) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

CI: Confidence interval; RR: Risk ratio

GRADE Working Group grades of evidence
High certainty: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate certainty: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low certainty: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low certainty: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

APPENDIXES

Appendix A. Marker states for VTE outcomes (available on ms.gradepro.org)

- Pulmonary embolism is usually without severe long-terms consequences. What is reported as symptomatic pulmonary embolism is best compatible with a moderate PE from the patient's point of view, even if described by imaging as massive or sub-massive. We will use the following description of a pulmonary embolism to reflect the average severity from a patient's point of view:
    - *Symptoms:* You will experience shortness of breath, sometimes pain and tightness in your chest.
    - *Time Horizon:* Moderate pulmonary embolism will impair you for weeks to months.
    - *Testing and Treatment:* Testing includes x-rays and CT-scans. Treatment will be administered in the hospital for a few days or at home. It typically includes administration of blood thinners using a small tube inserted into your vein or injections, followed by pills for months to years. You may require oxygen administration to improve your symptoms. To identify the cause of your problem you may require additional testing such as blood work or other x-rays and similar tests.

- Proximal deep venous thrombosis (DVT) is usually moderate in severity. What is reported as symptomatic deep venous thrombosis is best compatible with a moderate case scenario for proximal DVT (again the case scenarios or marker states ensure that panels have a similar understanding of the severity). We will use the following description of a proximal DVT to reflect the average severity:
    - *Symptoms:* You experience some swelling, pain, warmth, heaviness or redness in your entire leg.
    - *Time Horizon:* Moderate DVT will persist for months but improve over that time.
    - *Testing and Treatment:* Treatment may be administered in the hospital or at home. Treatment typically includes administration of blood thinners using a small tube inserted in your vein, injections or pills. Long-lasting treatment with blood thinners is often required.
- Distal DVT is assumed to be severe in order to be considered critical for decision-making. Only symptomatic severe distal DVT is critical for decision-making. We will use the following description of a distal DVT to reflect the average severity:
    - *Symptoms:* You experience severe swelling, pain, warmth, heaviness or redness in your leg below the knee (lower leg).
    - *Time Horizon:* Severe lower leg DVT will persist for months and will slowly improve.

- *Testing and Treatment:* Treatment may be administered in the hospital or at home. Treatment typically includes administration of blood thinners using a small tube inserted in your vein, injections or pills. Long-lasting treatment with blood thinners is often required.

- Major Bleeding
  - Fatal bleeding: Bleeding that is symptomatic and occurs in a critical area or organ, such as intracranial, intraspinal, intraocular, retroperitoneal, pericardial, in a non-operated joint, or intramuscular with compartment syndrome.
  - Extra-surgical site bleeding causing a fall in hemoglobin level of 20 g/L (1.24 mmol/L) or more, or leading to transfusion of 2 or more units of whole blood or red cells
  - Surgical site bleeding that requires a second intervention open, arthroscopic, endovascular
  - Hemarthrosis of sufficient size as to interfere with rehabilitation by delaying mobilization or delayed wound healing, resulting in prolonged hospitalization or a deep wound infection
  - Surgical site bleeding that is unexpected and prolonged and/or sufficiently large to cause hemodynamic instability, with an associate fall in hemoglobin level of at least 20 g/L (1.24 mmol/L or transfusion, indicated by the bleeding, of at least 2 units of whole blood or red cells.

- *Symptoms:* You lose a lot of blood (e.g. vomit blood, blood with your stools, blood from a wound) or you have an internal bleeding.
- *Time Horizon:* Bleeding does not stop, and you have to receive specific urgent care.
- *Testing and Treatment:* You may require a CT scan, a flexible tube via your mouth or anus to investigate your bowel, and blood work, and you may be admitted to hospital to receive blood transfusion or surgery.

Appendix B. Search strategies for randomized and non-randomized studies

See the search strategies in the following link:

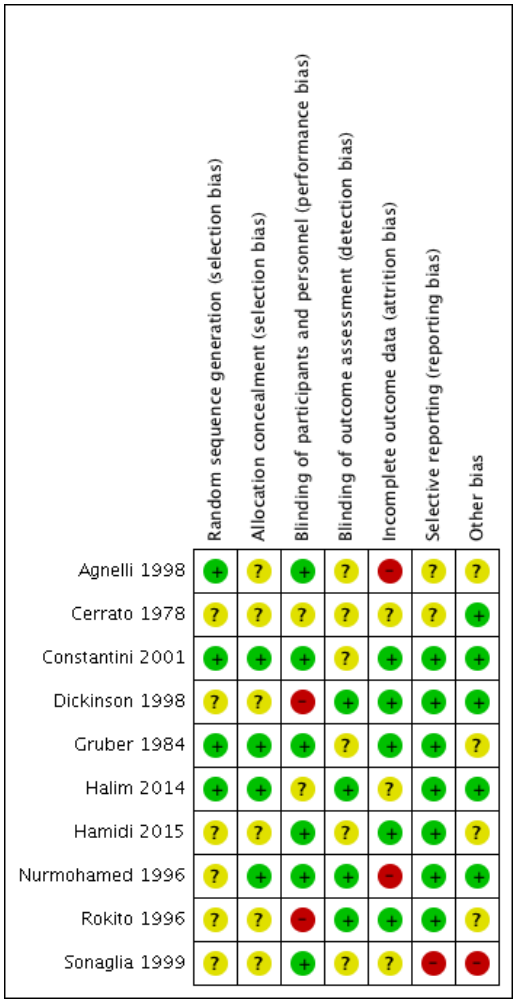
https://www.dropbox.com/s/rlc0tryngtisc4e/Appendix_B.docx?dl=0

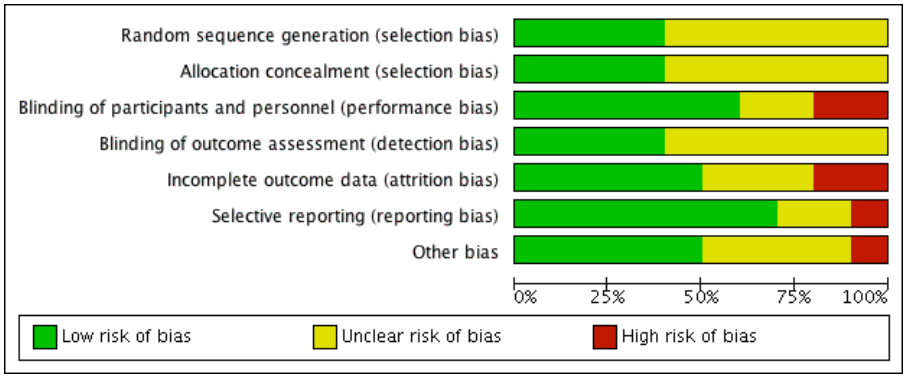
Appendix C. List of studies excluded

See the list of excluded studies in the following link:

https://www.dropbox.com/s/kpvdgbygmucng1d/Appendix_C.docx?dl=0

Appendix D. Summary of Risk of Bias in *randomized control trials* for RCT.





Appendix E. Summary of Risk of Bias in *non-randomized studies*.

Bauman 2009 (26)

Type of study: retrospective cohort study
Participants: patients with movement disorders (Parkinson disease, essential tremor, dystonia)
Experimental intervention: pharmacological thromboprophylaxis
Control intervention: no pharmacological thromboprophylaxis

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	*Bias in selection of participants into the study*	*Bias in classification of interventions*	*Bias due to deviations from intended interventions*	*Bias due to missing data*	*Bias in measurement of outcomes*	*Bias in selection of the reported result*	
Mortality	Benefit	Critical [1]	Moderate [2]	Moderate [3]	Low	Low	Low	Low	**CRITICAL**
Pulmonary embolism	Benefit	Critical [1]	Moderate [2]	Moderate [3]	Low	Low	Low	Low	**CRITICAL**
Major bleeding	Harm	Critical [1]	Moderate [2]	Moderate [3]	Low	Low	Low	Low	**CRITICAL**

Explanatory Footnotes

1. Bias due to confounding since authors did not include adjustments by potential confounders in the analysis and there was not an appropriate analysis method that controlled for all the important confounding domains.
2. Bias due to selection of participants into the study due to participants are not followed from the start of the intervention. Participants were selected if they received a unilateral or bilateral deep brain stimulation (DBS).
3. Bias due to classification of interventions since the information used to define intervention groups was not recorded at the start of the intervention.

Dermody 2011(16)

Type of study: retrospective cohort study
Participants: Neurosurgical patients who underwent screening with weekly venous duplex ultrasound
Experimental intervention: pharmacological thromboprophylaxis
Control intervention: no pharmacological thromboprophylaxis

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in classification of interventions***	***Bias due to deviations from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	
Major bleeding	Harm	Critical [1]	Serious [2]	Serious [3]	Low	Serious [4]	Low	Low	**CRITICAL**

Explanatory Footnotes

1. Bias due to confounding since authors did not include adjustments by potential confounders in the analysis and there was not an appropriate analysis method that controlled for all the important confounding domains.
2. Bias due to selection of participants into the study due to participants are not followed from the start of the intervention. Information was analyzed from patients admitted to a neurosurgical service from October 2007 to January 2010.
3. Bias due to classification of interventions since the information used to define intervention groups was not collected at the start of the intervention. Medical records of patients were reviewed to determine patients exposed and non-exposed to pharmacological prophylaxis.
4. Bias due to missing data since information about missing data for main outcome was not described by researchers.

Hacker 2012 (27)

Type of study: retrospective cohort study
Participants: neurosurgical and head trauma patients
Experimental intervention: pharmacological thromboprophylaxis
Control intervention: no pharmacological thromboprophylaxis

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	*Bias in selection of participants into the study*	*Bias in classification of interventions*	*Bias due to deviations from intended interventions*	*Bias due to missing data*	*Bias in measurement of outcomes*	*Bias in selection of the reported result*	
Mortality	Benefit	Critical [1]	Serious [2,3]	Moderate [4]	Low	Moderate [5]	Low	Low	**CRITICAL**
Pulmonary embolism	Benefit	Critical [1]	Serious [2,3]	Moderate [4]	Low	Moderate [5]	Low	Low	**CRITICAL**
Major bleeding	Harm	Critical [1]	Serious [2,3]	Moderate [4]	Low	Moderate [5]	Low	Low	**CRITICAL**

Explanatory Footnotes

1. Bias due to confounding since authors did not include adjustments by potential confounders in the analysis and there was not an appropriate analysis method that controlled for all the important confounding domains (unmeasured confounding).
2. Bias due to selection of participants into the study due to the selection of participants into the study was based on characteristic of participants after the start of intervention. Analysis was performed on prospectively collected data on all postoperative neurosurgical patients admitted over the course of 11 years to the surgical intensive care unit.
3. No statistical techniques for correcting selection bias were used in this study.
4. Bias due to classification of interventions since the information used to define intervention groups was not collected at the start of the intervention.
5. Bias due to missing data since there was not description of any analysis for missing data.

Khaldi 2011 (7)

Type of study: retrospective cohort study
Participants: patients who underwent a neurosurgical procedure
Experimental intervention: pharmacological thromboprophylaxis
Control intervention: no pharmacological thromboprophylaxis

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in classification of interventions***	***Bias due to deviations from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	
Pulmonary embolism	Benefit	Critical [1]	Moderate [2]	Critical [3]	Low	Serious [4]	Low	Low	**CRITICAL**
Distal DVT	Benefit	Critical [1]	Moderate [2]	Critical [3]	Low	Serious [4]	Low	Low	**CRITICAL**

Explanatory Footnotes
1. Bias due to confounding since authors did not include adjustments by potential confounders in the analysis and there was not an appropriate analysis method that controlled for all the important confounding domains (unmeasured confounding).
2. Bias due to selection of participants into the study due to participants are not followed from the start of the intervention. Participants were selected by reviewing records of all neurosurgical patients between January 2006 and December 2008
3. Bias due to classification of interventions since the information used to define intervention groups was not recorded at the start of the intervention.
4. Bias due to missing data since there was not description of any analysis for missing data.

Zachariah 2016 (30)

Type of study: retrospective cohort study
Participants: patients with subarachnoid hemorrhage and external ventricular drain
Experimental intervention: pharmacological thromboprophylaxis
Control intervention: no pharmacological thromboprophylaxis

Outcome	Benefit or harm of intervention	Domains							Overall Risk of Bias
		Bias due to confounding	***Bias in selection of participants into the study***	***Bias in classification of interventions***	***Bias due to deviations from intended interventions***	***Bias due to missing data***	***Bias in measurement of outcomes***	***Bias in selection of the reported result***	
Pulmonary embolism	Benefit	Critical [1]	Moderate [2]	Critical [3]	Low	Serious [4]	Low	Low	**CRITICAL**

Explanatory Footnotes

1. Bias due to confounding since authors did not include adjustments by potential confounders in the analysis and there was not an appropriate analysis method that controlled for all the important confounding domains (unmeasured confounding).
2. Bias due to selection of participants into the study due to participants are not followed from the start of the intervention. Authors described information of 241 consecutive patients with subarachnoid hemorrhage and external ventricular drain that attended Mayo Clinic from 2001 to 2014.
3. Bias due to classification of interventions since the information used to define intervention groups was not collected at the start of the intervention. Information of how patients received or not thromboprophylaxis was not mentioned.
4. Information about missing data for main outcome was not described.

Appendix F. Evidence profile

Author(s): Juan José Yepes-Nuñez, Luis Enrique Colunga Lozano, Stephanie Ross, Federico Popoff, Meha Bhatt, Kelly Estrada Orozco, Angela Barbara, Sara Balduzzi, Housne Ara, Arnav Agarwal, Wojtek Wiercioch, Gian Paolo Morgano, Holger J. Schünemann.
Date: April 2018
Question: Pharmacological prophylaxis compared to no pharmacological prophylaxis in patients undergoing major neurosurgical procedures
Setting: inpatient
Bibliography: Agnelli (1), Constantini (2), Dickinson (3), Gruber (4), Hamidi (5), Nurmohamed (6), Rokito (7), Bauman (8), Dermody (9), Hacker (10).

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	pharmacological prophylaxis	no pharmacological prophylaxis	Relative (95% CI)	Absolute (95% CI)		
Mortality - RCTs												
5	randomised trials	serious [a]	not serious	not serious	serious [b]	none	27/512 (5.3%)	18/517 (3.5%)	**RR 1.27** (0.57 to 2.86)	**9 more per 1,000** (from 15 fewer to 65 more)	⨁⨁◯◯ LOW	CRITICAL
Mortality - NRS												
2	observational studies	very serious [c]	serious [d]	not serious	serious [e]	none	28/344 (8.1%)	38/330 (11.5%)	**RR 0.72** (0.46 to 1.13)	**32 fewer per 1,000** (from 15 more to 62 fewer)	⨁◯◯◯ VERY LOW	CRITICAL
Symptomatic Pulmonary Embolism - as described by the moderate marker state - RCTs (assessed with: Symptomatic PE)												
3	randomised trials	serious [f]	serious [g]	not serious	very serious [h]	none	2/213 (0.9%)	3/221 (1.4%)	**RR 0.84** (0.03 to 27.42)	**2 fewer per 1,000** (from 13 fewer to 359 more)	⨁◯◯◯ VERY LOW	CRITICAL
								0.2% [i]		**0 fewer per 1,000** (from 2 fewer to 53 more)		
Symptomatic Pulmonary Embolism - as described by the moderate marker state - NRS (assessed with: Symptomatic PE)												
2	observational studies	very serious [c]	serious [j]	not serious	serious [h]	none	0/346 (0.0%)	2/430 (0.5%)	**RR 0.18** (0.01 to 3.76)	**4 fewer per 1,000** (from 5 fewer to 13 more)	⨁◯◯◯ VERY LOW	CRITICAL

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	pharmacological prophylaxis	no pharmacological prophylaxis	Relative (95% CI)	Absolute (95% CI)		
								0.2% [i]		**2 fewer per 1,000** (from 2 fewer to 6 more)		
Symptomatic Proximal Deep Vein Thrombosis - as described by the moderate marker state - RCTs (assessed with: Screening detected Proximal DVT)												
2	randomised trials	serious [k]	not serious	serious [l]	not serious	none	21/371 (5.7%)	42/373 (11.3%)	**RR 0.50** (0.30 to 0.84) [p]	**56 fewer per 1,000** (from 18 fewer to 79 fewer)	⨁⨁◯◯ LOW	CRITICAL
								0.3% [m]		**2 fewer per 1,000** (from 1 fewer to 2 fewer)		
								0.3% [n,o]		**2 fewer per 1,000** (from 1 fewer to 2 fewer)		
Symptomatic Distal Deep Vein Thrombosis - as described by the severe marker state (assessed with: Screening detected Distal DVT)												
1	randomised trials	serious [q]	not serious	serious [l]	serious [r]	none	15/130 (11.5%)	25/129 (19.4%)	**RR 0.60** (0.33 to 1.08) [p]	**78 fewer per 1,000** (from 16 more to 130 fewer)	⨁◯◯◯ VERY LOW	CRITICAL
								0.2% [s]		**1 fewer per 1,000** (from 0 fewer to 1 fewer)		
								0.1% [n,o]		**0 fewer per 1,000** (from 0 fewer to 0 fewer)		
Major Bleeding - RCTs												

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	pharmacological prophylaxis	no pharmacological prophylaxis	Relative (95% CI)	Absolute (95% CI)		
7	randomised trials	serious [t]	not serious	not serious	serious [u]	none	17/572 (3.0%)	10/584 (1.7%)	**RR 1.57** (0.70 to 3.50)	**10 more per 1,000** (from 5 fewer to 43 more)	⨁⨁◯◯ LOW	CRITICAL
Major Bleeding - NRS												
3	observational studies	very serious [v]	not serious	not serious	very serious [w]	none	3/471 (0.6%)	3/459 (0.7%)	**RR 1.45** (0.30 to 7.12)	**3 more per 1,000** (from 5 fewer to 40 more)	⨁◯◯◯ VERY LOW	CRITICAL
Reoperation - RCTs												
2	randomised trials	serious [x]	not serious	not serious	very serious [y]	none	1/95 (1.1%)	3/97 (3.1%)	**RR 0.43** (0.06 to 2.84)	**18 fewer per 1,000** (from 29 fewer to 57 more)	⨁◯◯◯ VERY LOW	CRITICAL

CI: Confidence interval; **RR:** Risk ratio

Explanations

a. Studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to lack of information about the sequence generation process in 3 out of 5 studies and lack of concealment in [2 out of 5 studies.
b. Very serious imprecision. Wide confidence interval with only 45 events in total
c. Serious risk of bias. Studies did not analyze findings adjusting for confounding factors
d. Serious inconsistency. Unexplained inconsistency, with point estimates widely different and confidence intervals not overlapping (P-value chi square= 0.12; I2= 43%)
e. Serious imprecision. 95% CI is consistent with the possibility for important benefit and large harm exceeding a minimal important difference, including only 66 events in total.
f. Studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to lack of blinding of outcome assessment in 2 out of 4 studies.
g. Serious inconsistency. Moderate heterogeneity between studies: I^2 = 64% (P=0.10)
h. Very serious imprecision. Wide confidence interval with only 5 events in total
i. A systematic review of 25 NRS published by Glotzbecker 2009 on elective spinal surgeries (cervical spine, lumbar laminectomy, lumbar spinal fusion, spinal trauma, spinal tumors) reported an incidence of symptomatic PE of 0.2% (34/15204)
j. Serious inconsistency. Moderate inconsistency, with point estimates widely different and confidence intervals not overlapping (P-value chi-square= 0.18; I2= 41%).
k. Studies that carried large weight for the overall effect estimate rated as high risk of bias due to lack of information about the incomplete outcome data
l. Serious indirectness. Patients were identified through screening ultrasound. None of the patients developed symptomatic venous thromboembolism before venography.
m. The baseline risk consists of the control group event rate (11.3%) from studies included in the meta-analysis. Baseline risk estimates for symptomatic proximal DVT (2.26%) has been calculated applying the assumptions that 20% of any proximal DVTs are symptomatic proximal DVTs.
n. Rates of proximal and distal symptomatic DVT in patients receiving no prophylaxis and undergoing elective spinal surgeries (cervical spine, lumbar laminectomy, lumbar spinal fusion, spinal trauma, spinal tumors) were reported in Glotzbecker 2009 1.6% (46/2956) for DVTs and 0.2% (34/15204) for PEs
o. We applied the assumption that approximately 20% of symptomatic DVTs are proximal, 80% distal and 5% of the latter severe
p. If any DVT detected by screening was considered a surrogate, then six randomized controlled trials (RCT) and two non-randomized studies (NRS) measured it; there were a total of 137 events (53 in prophylaxis group and 84 in no prophylaxis group) among 927 patients for the RCTs, and 72 events (32 in prophylaxis group and 40 in no prophylaxis group) among 415 patients for the NRS. For the RCTs, the RR would be 0.65 (95% CI: 0.47 to 0.89), and the risk difference would be 64 fewer per 1,000 (from 21 fewer to 96 fewer) using the control group event rate of 17.7%, or 1 fewer per 1000 (from 1 fewer to 2 fewer) based on the baseline risk of 0.32%. For the NRS, the RR would be 0.48 (95% CI: 0.29 to 0.81), and the risk difference would be 96 fewer per 1,000 (from 35 fewer to 131 fewer) using the control group event rate of 18.4%, or 2 fewer per 1000 (from 1 fewer to 2 fewer) based on the baseline risk of 0.32%.

q. One study that carried large weight for the overall effect estimate rated as high risk of bias due to lack of incomplete outcome data.
r. Very serious imprecision. 95% CI is consistent with the possibility for important benefit and large harm exceeding a minimal important difference, including only 40 events in total
s. The baseline risk consists of the control group event rate (19.4%) from studies included in the meta-analysis. Baseline risk estimates for symptomatic distal DVT (0.194 %) has been calculated applying the assumptions that 20% of any distal DVTs are symptomatic distal DVTs and that only 5% of the symptomatic distal DVTs are assumed to be severe DVTs.
t. Studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to lack of random sequence generation and lack of concealment in 4 out of 7 studies
u. Serious imprecision. 95% CI is consistent with the possibility for important benefit and large harm exceeding a minimal important difference, including only 24 events in total.
v. Serious risk of bias. Studies assessed comorbidities associate with high risk of DVT such obesity, heart failure, obesity, cancer, history of DVT, pregnancy, tobacco use, and history of hypercoagulable disorder. However, authors did not adjust for confounding factors.
w. Very serious imprecision. 95% CI is consistent with the possibility for important benefit and large harm exceeding a minimal important difference, including only 6 events in total.
x. Studies that carried large weight for the overall effect estimate rated as unclear risk of bias due to lack of random sequence generation and allocation concealment] in 1 out of 2 studies and lack of blinding of outcome assessment in 1 out of 2 studies.
y. Very serious imprecision. Wide confidence interval with only 4 events in total.

REFERENCES

1. Agnelli G, Piovella F, Buoncristiani P, Severi P, Pini M, D'Angelo A, et al. Enoxaparin plus compression stockings compared with compression stockings alone in the prevention of venous thromboembolism after elective neurosurgery. N Engl J Med. 1998;339(2):80-5.

2. Constantini S, Kanner A, Friedman A, Shoshan Y, Israel Z, Ashkenazi E, et al. Safety of perioperative minidose heparin in patients undergoing brain tumor surgery: a prospective, randomized, double-blind study. J Neurosurg. 2001;94(6):918-21.

3. Dickinson LD, Miller LD, Patel CP, Gupta SK. Enoxaparin increases the incidence of postoperative intracranial hemorrhage when initiated preoperatively for deep venous thrombosis prophylaxis in patients with brain tumors. Neurosurgery. 1998;43(5):1074-81.

4. Gruber UF, Rem J, Meisner C, Gratzl O. Prevention of thromboembolic complications with miniheparin-dihydroergotamine in patients undergoing lumbar disc operations. Eur Arch Psychiatry Neurol Sci. 1984;234(3):157-61.

5. Hamidi S, Riazi M. Incidence of venous thromboembolic complications in instrumental spinal surgeries with preoperative chemoprophylaxis. J Korean Neurosurg Soc. 2015;57(2):114-8.

6. Nurmohamed MT, van Riel AM, Henkens CM, Koopman MM, Que GT, d'Azemar P, et al. Low molecular weight heparin and compression stockings in the prevention of venous thromboembolism in neurosurgery. Thromb Haemost. 1996;75(2):233-8.

7. Rokito SE, Schwartz MC, Neuwirth MG. Deep vein thrombosis after major reconstructive spinal surgery. Spine (Phila Pa 1976). 1996;21(7):853-8; discussion 9.

8. Bauman JA, Church E, Halpern CH, Danish SF, Zaghloul KA, Jaggi JL, et al. Subcutaneous heparin for prophylaxis of venous thromboembolism in deep brain stimulation surgery: evidence from a decision analysis. Neurosurgery. 2009;65(2):276-80; discussion 80.

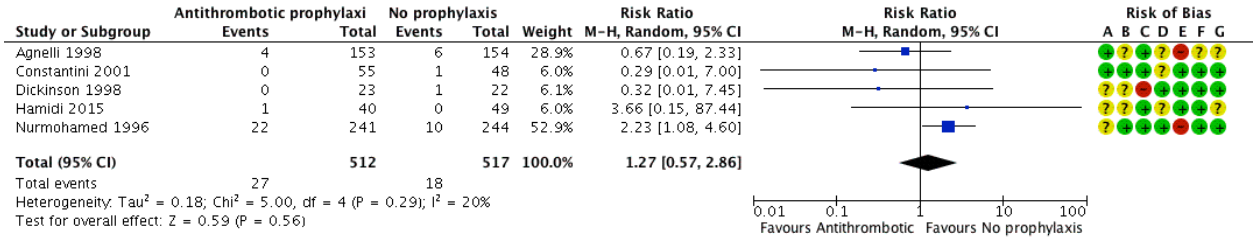
9. Dermody M, Alessi-Chinetti J, Iafrati MD, Estes JM. The utility of screening for deep venous thrombosis in asymptomatic, non-ambulatory neurosurgical patients. J Vasc Surg. 2011;53(5):1309-15.

10. Hacker RI, Ritter G, Nelson C, Knobel D, Gupta R, Hopkins K, et al. Subcutaneous heparin does not increase postoperative complications in neurosurgical patients: An institutional experience. J Crit Care. 2012;27(3):250-4.

11. Glotzbecker MP, Bono CM, Wood KB, Harris MB. Thromboembolic disease in spinal surgery: a systematic review. Spine (Phila Pa 1976). 2009;34(3):291-303.

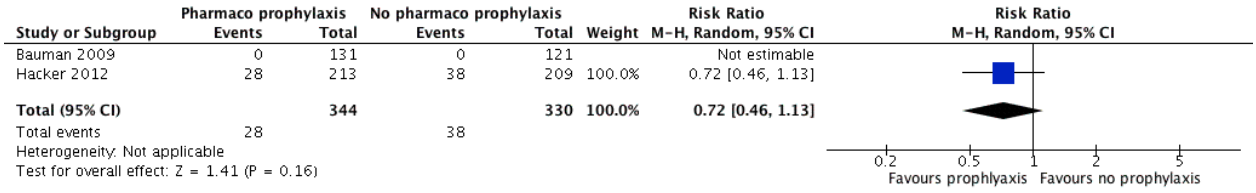
Appendix G. Forest plots of meta-analysis for each outcome for each type of study design

- **Mortality – randomized controlled trials (RCT) –**

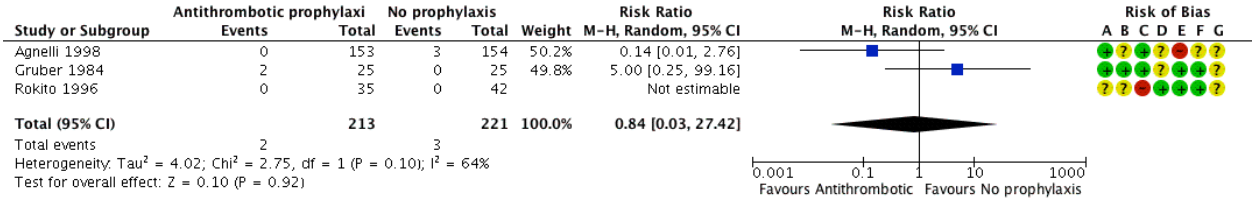


Risk of bias legend
(**A**) Random sequence generation (selection bias)
(**B**) Allocation concealment (selection bias)
(**C**) Blinding of participants and personnel (performance bias)
(**D**) Blinding of outcome assessment (detection bias)
(**E**) Incomplete outcome data (attrition bias)
(**F**) Selective reporting (reporting bias)
(**G**) Other bias

- **Mortality – non-randomized studies (NRS) –**



- **Pulmonary embolism – RCTs –**

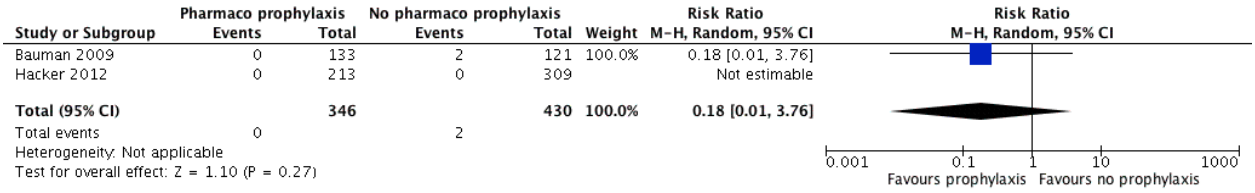


Study or Subgroup	Antithrombotic prophylaxi Events	Total	No prophylaxis Events	Total	Weight	Risk Ratio M-H, Random, 95% CI
Agnelli 1998	0	153	3	154	50.2%	0.14 [0.01, 2.76]
Gruber 1984	2	25	0	25	49.8%	5.00 [0.25, 99.16]
Rokito 1996	0	35	0	42		Not estimable
Total (95% CI)		**213**		**221**	**100.0%**	**0.84 [0.03, 27.42]**
Total events	2		3			

Heterogeneity: Tau² = 4.02; Chi² = 2.75, df = 1 (P = 0.10); I² = 64%
Test for overall effect: Z = 0.10 (P = 0.92)

Risk of bias legend
(A) Random sequence generation (selection bias)
(B) Allocation concealment (selection bias)
(C) Blinding of participants and personnel (performance bias)
(D) Blinding of outcome assessment (detection bias)
(E) Incomplete outcome data (attrition bias)
(F) Selective reporting (reporting bias)
(G) Other bias

- **Pulmonary embolism – NRS –**



Study or Subgroup	Pharmaco prophylaxis Events	Total	No pharmaco prophylaxis Events	Total	Weight	Risk Ratio M-H, Random, 95% CI
Bauman 2009	0	133	2	121	100.0%	0.18 [0.01, 3.76]
Hacker 2012	0	213	0	309		Not estimable
Total (95% CI)		**346**		**430**	**100.0%**	**0.18 [0.01, 3.76]**
Total events	0		2			

Heterogeneity: Not applicable
Test for overall effect: Z = 1.10 (P = 0.27)

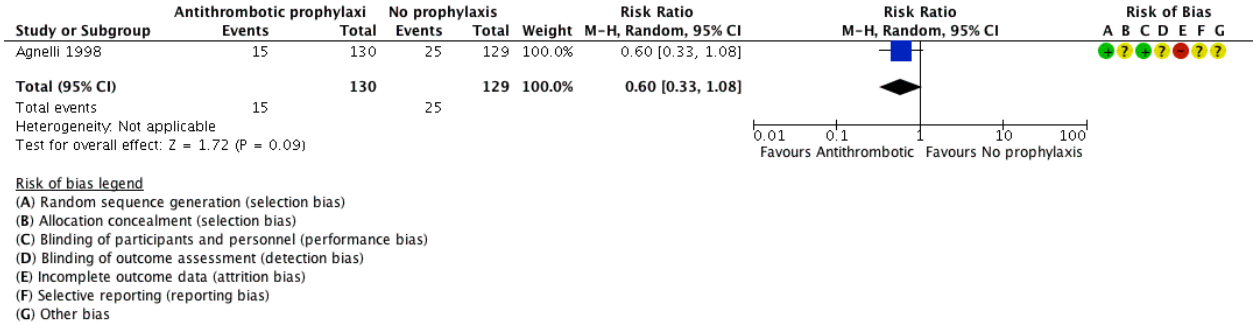
- **Proximal deep venous thrombosis (DVT) – RCTs –**

Study or Subgroup	Antithrombotic prophylaxi Events	Total	No prophylaxis Events	Total	Weight	Risk Ratio M-H, Random, 95% CI
Agnelli 1998	7	130	17	129	35.7%	0.41 [0.18, 0.95]
Nurmohamed 1996	14	241	25	244	64.3%	0.57 [0.30, 1.06]
Total (95% CI)		**371**		**373**	**100.0%**	**0.50 [0.30, 0.84]**
Total events	21		42			

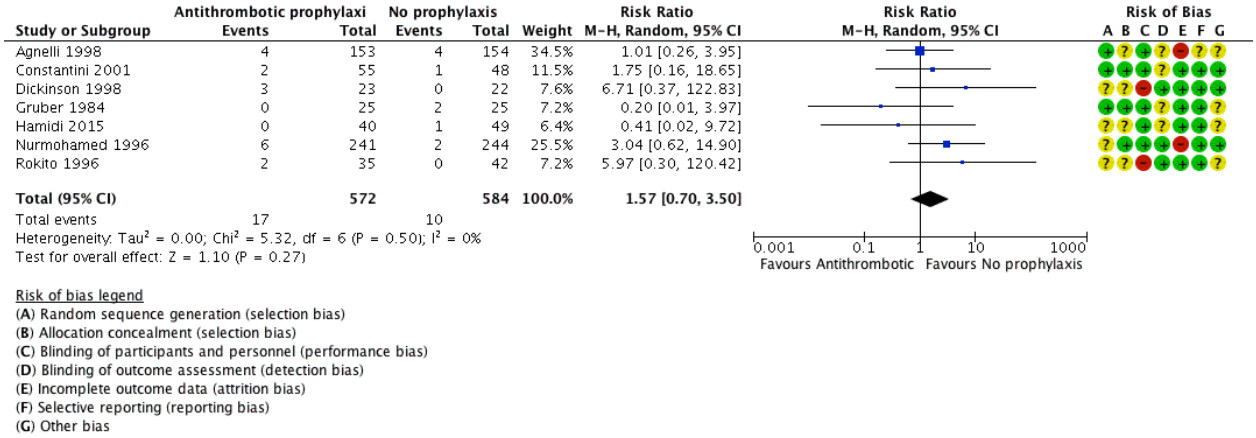
Heterogeneity: Tau² = 0.00; Chi² = 0.37, df = 1 (P = 0.54); I² = 0%
Test for overall effect: Z = 2.66 (P = 0.008)

Risk of bias legend
(A) Random sequence generation (selection bias)
(B) Allocation concealment (selection bias)
(C) Blinding of participants and personnel (performance bias)
(D) Blinding of outcome assessment (detection bias)
(E) Incomplete outcome data (attrition bias)
(F) Selective reporting (reporting bias)
(G) Other bias

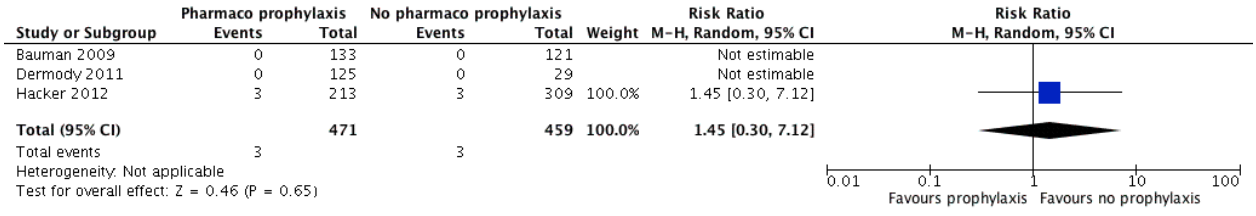
- **Distal DVT – RCTs –**



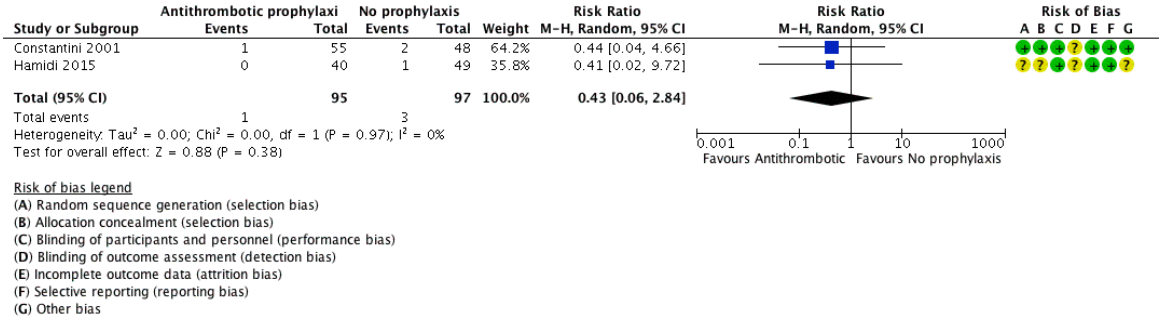
- **Major bleeding – RCTs –**



- **Major bleeding – NRS –**



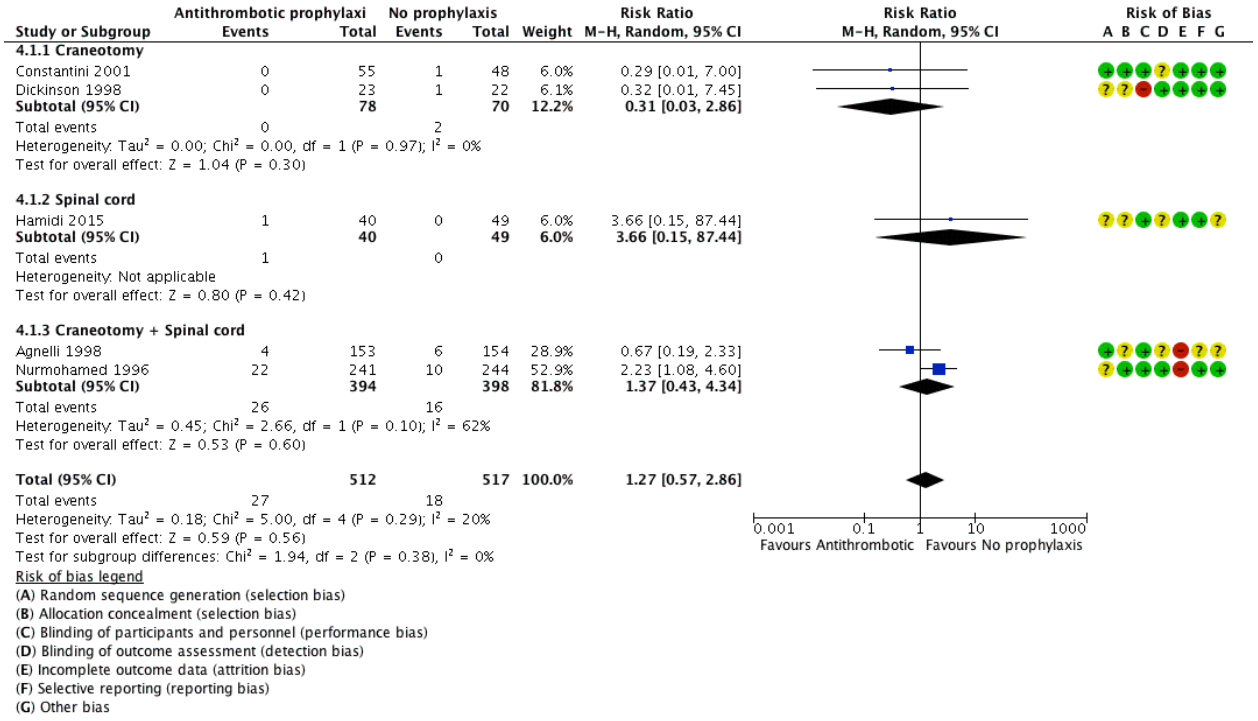
- **Reoperation – RCTs –**



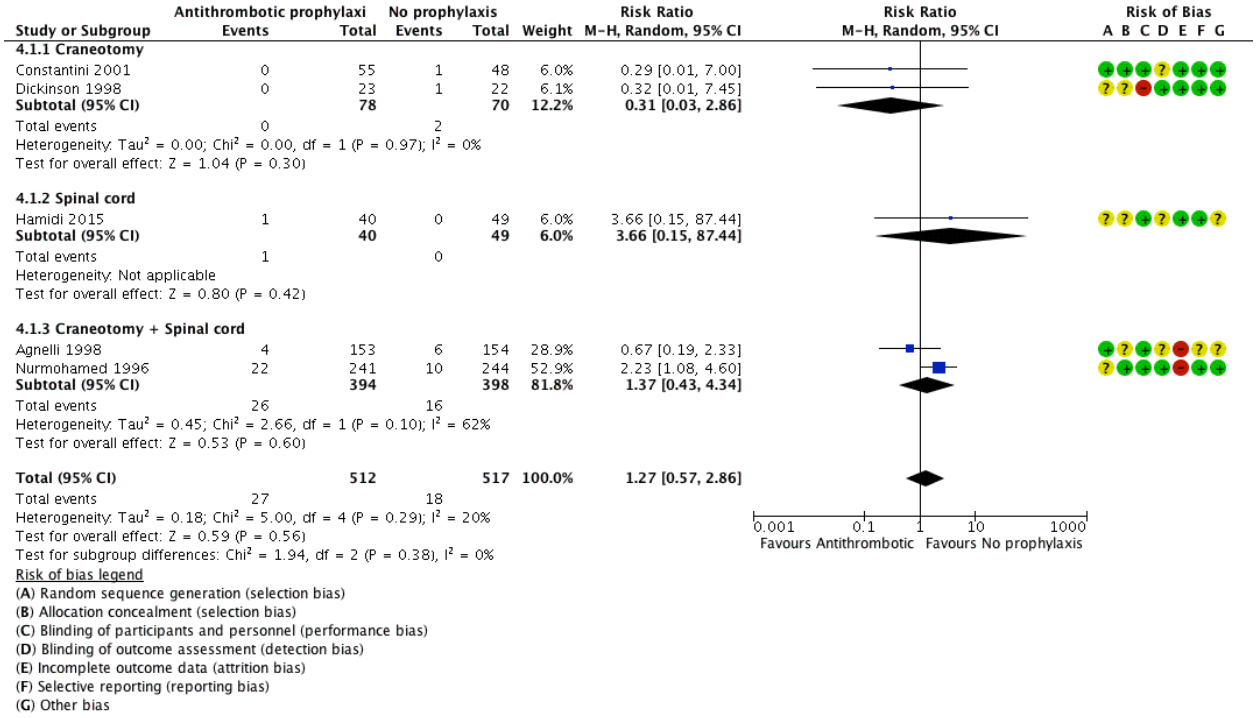
Risk of bias legend
(**A**) Random sequence generation (selection bias)
(**B**) Allocation concealment (selection bias)
(**C**) Blinding of participants and personnel (performance bias)
(**D**) Blinding of outcome assessment (detection bias)
(**E**) Incomplete outcome data (attrition bias)
(**F**) Selective reporting (reporting bias)
(**G**) Other bias

Appendix H. Subgroup analysis by the of neurosurgical intervention in RCTs

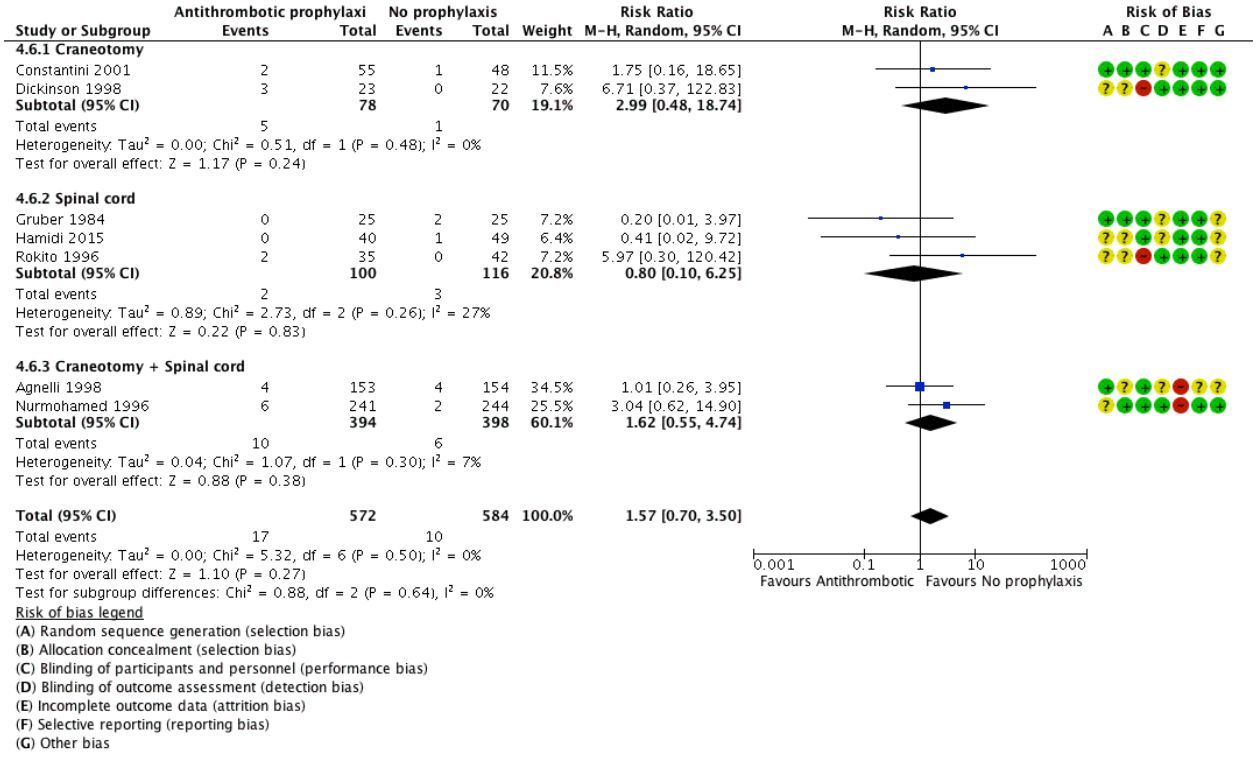
- **Mortality**



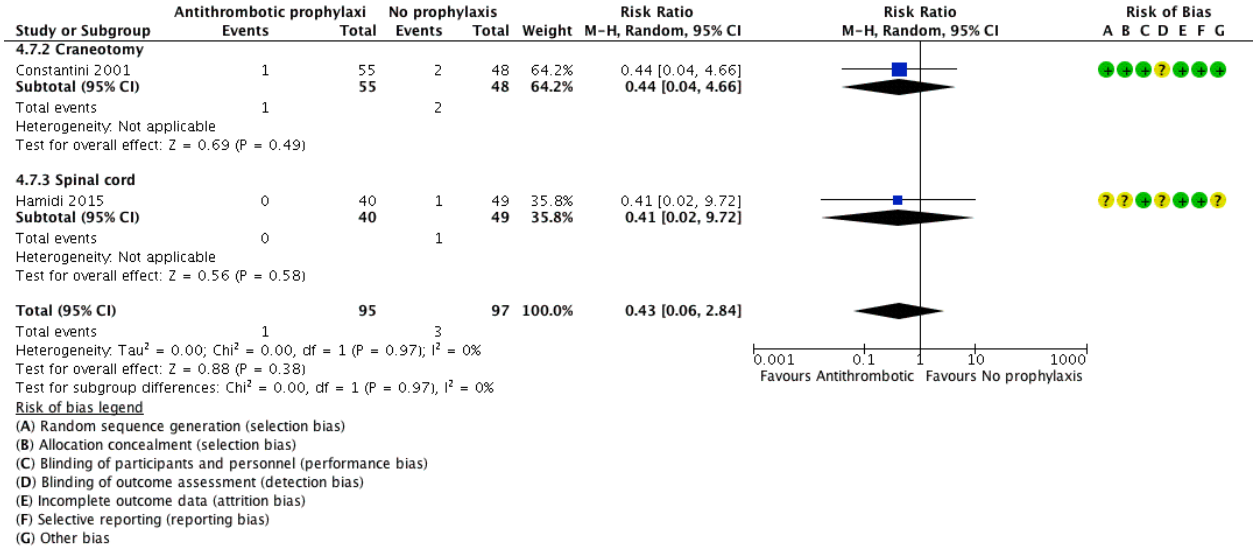
- **Pulmonary embolism**



- **Major bleeding**



- **Reoperation**



CHAPTER 6. CONCLUSION

Chapter 6: Conclusion

In this dissertation, I expose methodological issues in conducting and reporting NMA and SR that include randomized controlled trials (RCT) and non-randomized studies (NRS). We developed two Network meta-analysis (NMA) Summary of Findings (SoF) tables and we inform methodological considerations of conducting and reporting NMA research and systematic reviews (SR) that includes bodies of evidence from RCT and NRS.

We hope the advances reported in this thesis dissertation will be useful for methodologists and health-care professionals to better understand NMA findings and to improve the quality of conducting NMAs and SR of RTC and NRS.

Summary of Findings

This thesis dissertation presents four pillars of information that can be summarized as follows:

a) Network meta-analysis (NMA) studies need to be conducted with higher methodological rigorousness. Issues in the application of statistical methods were identified, mainly in conducting of heterogeneity analysis within individual randomized control trials, and between direct and indirect evidence. Furthermore, analysis to explain heterogeneity was not frequently conducted

in most NMAs. Drawing conclusions using inaccurate information can mislead the applicability of NMA findings

b) Multiple terms for the same statistical approaches are used in NMA reports. These different terms can make it difficult for NMA users to understand the development of an NMA. Harmonization of NMA terminology is needed to increase consistency of NMA reports.

c) Evaluation of body of evidence is applicable for NMA as well as other types of resource of evidence. In 2014, GRADE developed a framework that assesses the certainty in evidence in NMA. However, most NMA publications in 2014 and 2015 did not evaluate the certainty of evidence.

d) New GRADE NMA-SoF tables to display NMA findings are available for NMA authors to report NMA results to NMA users. These new SoF tables will allow NMA users to understand NMA findings in a more approachable manner.

e) Applicability of ROBINS-I tool in individual NRS can be challenged, however, a comprehensive analysis of methodological limitations of NRS is ensured when the tool is implemented. ROBINS-I domains allow better analysis of NRS limitations.

f) We confronted presentations of NRS and RCT findings by outcomes in systematic reviews through forest plots. We found no issues in the presentation of these findings. Conversely, it allowed us to understand differences in the estimates of both resources of evidence. Presentation of these findings through a tabular format could help to understand information

regarding the certainty of evidence, the magnitude of effect for all important outcomes for a specific comparison.

Implications for researchers, guideline developers, policy makers, and clinicians

This work has several implications for different contributors in developing and implementing findings from evidence synthesis summaries.

Implications for researchers, guideline developers, policy makers

Network meta-analysis (NMA) is a relatively novel resource of evidence synthesis in one meta-analysis when three or more different interventions are available and direct and indirect estimates are combined (1, 2). Multiple statistical methods are needed to properly conduct an NMA. Our analysis of statistical methods and format presentation in 276 NMA publications identified multiple weaknesses in conducting NMAs. Researchers, guideline developers, and policy makers can benefit from our findings as they can improve the development of NMA by applying appropriate statistical analysis methods and an evaluation of the certainty of evidence in the body of evidence in their NMA findings. A standard format that describes step by step how to conduct an NMA could help researchers to appropriately develop NMAs. Consistent application of an instrument as the PRISMA statement for NMA (3), development of NMA protocols, and application of a GRADE framework for NMA (4, 5) will also help

researchers, guideline developers, and policy makers to improve reliability in their NMAs.

Non-randomized studies (NRS) are a resource of evidence synthesis for systematic reviews when interventions cannot be randomized or when there are long term and rare outcomes (6). Evaluation of bias in individual RCTs and NRS is needed when both resources of evidence are included in a meta-analysis to estimate the true intervention effect. In our work we confronted the Cochrane risk of bias tool (7) and the ROBINS-I (8) tool to assess the risk of bias in RCTs and NRS respectively. Benefits for researchers, guideline developers, and policy makers are shown in our work as they can use our evaluation as an example of integrating, without combining, both resources of evidence in a single meta-analysis. Our examples showed that it did not compromise the reliability in the estimations of the effect for health-care pairwise comparisons. Moreover, we confronted data presentation for both resources of evidence by exploring the effect estimates in forest plots. We found no differences in analysing the effect estimates, and their magnitude, of RCT and NRS separately. When researchers, guideline developers, and policy makers are synthetizing evidence from NRS and RCT, we suggest looking at the effect estimates separately and exploring potential resources of heterogeneity that can be found in both bodies of evidence. Owing to absence of randomization, NRS are more susceptible to involve potential bias than RCTs and the separate analysis could help to understand differences between both resources. We also included the GRADE approach that

integrates the evaluation of RoB in NRS with the ROBINS-I tool (9). Our work will allow other researchers to continue efforts in integrating NRS in evidence synthesis studies by exploring pros and cons in formatting evaluation of NRS with ROBINS-I tool.

Implications for clinicians and other users of evidence synthesis

Interpretation of NMA findings can be challenging for NMA users as NMA involves multiple and sophisticated statistical analysis (10), and there is not an standardize tabular format to summarize their findings. We developed an NMA-SoF table that can help to understand NMA findings to users such as clinicians. We built our NMA-SoF table based on information retrieved from the health literature as well as with feedback from methodologist [clinical epidemiologist (including clinical practice guideline and systematic review developers) and biostatistics] and clinicians. Input from clinicians was fundamental to design an NMA SoF table useful in health-care practice. Likewise, input in the NMA-SoF table from practice guideline and systematic review developers helped to design and NMA-SoF table useful from health-care decision making purpose in the context of clinical practice guidelines, and systematic reviews with meta-analysis.

Clinicians and other users of evidence synthesis can also benefit from the findings reported in our two systematic reviews. Our systematic reviews reported the assessment of risk of bias for all individual primary studies of both bodies of evidence, RCTs and NRS. We also presented separately estimates and their

uncertainty for both RCTs and NRS for each outcome in forest plots. When we confronted both resources of evidence, we ensured a critical evaluation of individual primary studies as well as a clear distinction the magnitude of estimates of the evidence, improving the presentation data and findings of SR that included RCTs and NRS. We hope the clarity in presenting the findings as we did it, can serve as an example for other research work in the same research area.

Strengths and challenges of this work

This thesis dissertation has multiple strengths. We reviewed many NMA publications, and we identified numerous flaws in these reports. Most NMAs applied insufficient statistical methods to conduct NMA analysis. Moreover, although multiple NMA publications assessed the risk of bias for individual primary RCTs using different tools, only few NMAs assessed the certainty of evidence using the GRADE framework. With our findings, we are convinced that additional work needs to be implemented to improve the applicability of GRADE framework (4, 5) to draw more reliable NMA conclusions.

Our NMA systematic survey also allowed us to understand the components that should be part of an NMA-SoF table. We also built our NMA-SoF table based on quantitative and qualitative findings and we developed an NMA-SoF table that is suitable to present NMA findings in an understandable manner as we included preferences and feedback from diverse NMA users. We hope our format can be

useful for researchers and clinicians that conduct NMAs as the current format SoF format (11-13) has been useful for displaying findings of pair-wise comparison meta-analysis.

We included in our two systematic reviews the body of evidence from RCTs and NRS. The strength of this work falls in the applicability of the ROBINS-I tool in the assessment of risk of bias in individual NRS as well as its integration with the GRADE framework (9). We found no conceptual issues in the application of the RoB tool and its integration with the GRADE framework. We also presented the effect estimates and their magnitude for NRS and RCTs separately. Data interpretation and drawing conclusions for each body of evidence separately were comprehensible and reasonable for methodologists and clinicians who were part of the systematic reviews developer team.

We faced two challenges with the development of the NMA work. First, in the NMA systematic survey, our learning curve to gain familiarity with different NMA terms was slow. The process of reviewing NMA publications took more time than we expected. Unfortunately, NMA research uses different terms for the same concepts which makes it problematic to understand the process of conducting an NMA. Second, in building our NMA-SoF table, the development of the qualitative study allowed us to gather input for multiple users. Although the final NMA-SoF table was appealing for most users, we needed to integrate multiple and varied concepts and preferences from methodologists and clinicians.

Two important challenges came up with the application of the ROBINS-I tool. We spent considerable time gaining familiarity with the tool as well as its implementation. Although there is an available ROBINS-I tool manual to understand the implementation of the tool, ROBINS-I has considerable number items in each domain making it challenging to understand its applicability. Its implementation on individual NRS was challenging for most users.

Further research directions

Drawing conclusions of NMA can be challenging for NMA authors considering the high amount of the body of evidence that an NMA could summarize. Additional efforts need to be settled for implementing the NMA GRADE framework to assess the certainty of evidence in NMAs. Interpretation of NMA findings based on the effect estimates and the certainty of evidence will avoid misleading NMA inferences (14).

We are aware that other users of the NMA-SoF tables will have different preferences of how to display NMA findings in an NMA-SoF table. Therefore, future research should be focused on developing additional NMA SoF table formats that display other components of NMA data presented in SoF tables. Additional work is also needed in user-testing the NMA-SoF table in a clinical practice guidelines context. Although we designed our SoF table with a focus on summary findings of NMA publications, we have not tested this format in a health-care decision making setting. Having the insights of the NMA-SoF table

from guideline developers and panel members of clinical practice guidelines, will increase our knowledge in improving the current NMA-SoF table format.

Methodological issues from systematic reviews of NRS and RCT will arise to the extent that researchers apply, for instance, the ROBINS-I tool in their reviews. Using this tool was not without challenges during our reviews. While there is guidance to comprehend the structure of the tool, its applicability when it was confronted with individual NRS was not always easy to follow. With a high number of items and domains, in addition to the challenge to covering all aspects of potential bias in NRS, implementation of the ROBINS-I tool still needs to be developed.

Final remarks

This thesis dissertation provided reliable and useful information to improve the data analysis and format presentation of NMA findings in NMA publications. It created new knowledge on how to properly conduct and summarize evidence of NMAs.

We also confronted risk of bias assessment of RCTs and NRS in systematic reviews, and we integrated these evaluations with the GRADE framework successfully. In addition, we improved the effect estimates presentation.

Implementation of findings reported in this thesis dissertation on other similar projects will reveal methodological topics for additional research.

REFERENCES

1. Higgins J, Deeks J, Altman D. Chapter 16: Special topics in statistics. 2011. In: Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011] [Internet]. The Cochrane Collaboration. Available from: www.cochrane-handbook.org.

2. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;331(7521):897-900.

3. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, 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.

4. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630.

5. Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol. 2018;93:36-44.

6. Reeves B, Deeks J, Higgins J, Wells G. Chapter 13: Including non-randomized studies. 2011. In: Cochrane Handbook for Systematic Reviews of

Interventions Version 510 [updated March 2011] [Internet]. The Cochrane Collaboration. Available from: www.cochrane-handbook.org.

7. Higgins J, Altman D, Sterne J. Chapter 8: Assessing risk of bias in included studies. In: JPT H, S G, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011): The Cochrane Collaboration; 2011.

8. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

9. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2018.

10. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. BMJ. 2013;346:f2914.

11. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013;66(2):158-72.

12. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol. 2013;66(2):173-83.

13. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format. J Clin Epidemiol. 2016;74:7-18.

14. Mbuagbaw L, Rochwerg B, Jaeschke R, Heels-Andsell D, Alhazzani W, Thabane L, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. Syst Rev. 2017;6(1):79.