

RANDOMIZED  
AND  
NON-RANDOMIZED STUDIES  
IN HEALTH SYNTHESSES

**THE ROLE OF  
RANDOMIZED AND  
NON-RANDOMIZED STUDIES  
IN KNOWLEDGE SYNTHESIS  
OF HEALTH  
INTERVENTIONS**

By CARLOS ALBERTO CUELLO-GARCIA, M.D.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the  
Requirements for the Degree of Doctor of Philosophy

McMaster University © Copyright by Carlos A. Cuello-Garcia, September 2017

McMaster University DOCTOR OF PHILOSOPHY (2017)

Hamilton, Ontario (Health Research Methodology Program).

TITLE: The Role of Randomized and Non-Randomized Studies in Knowledge  
Synthesis of Health Interventions.

AUTHOR: Carlos A. Cuello-Garcia, M.D. (McMaster University)

SUPERVISOR: Professor Holger J. Schünemann, M.D., Ph.D., M.Sc.

NUMBER OF PAGES: xviii, 190

# LAY ABSTRACT

All recommendations about healthcare interventions (from common medicines to strategies to prevent diseases) should ideally come from an adequate synthesis (e.g., systematic reviews) of the least biased studies. Many researchers and authors of health syntheses consider randomized studies (RS), the ‘gold standard’ to demonstrate if an intervention is truly effective. Unfortunately, they are not always available, feasible, or ethical to conduct. Non-randomized studies (NRS), also called observational studies, can potentially provide complementary evidence for a research question. Unfortunately, they are usually considered of poorer quality because of their intrinsic nature of being prone to bias and confounding. In most circumstances, authors of syntheses discard these types of studies from the outset, without considering their potential for providing evidence that could complement or even replace that from randomized studies.

This work aims to improve this situation by offering methods for evaluating the appropriateness of integrating both RS and NRS, guiding authors and researchers in cases where this is possible, hence increasing the certainty in a body of evidence and help all stakeholders reach decisions.

# ABSTRACT

Randomized studies (RS) are considered the best source of evidence for knowledge syntheses (e.g., systematic reviews, health technology assessments, health guidelines, among others) about healthcare interventions. Historically, non-randomized studies (NRS) have been usually discarded from knowledge syntheses of interventions due to their intrinsic risk of bias and confounding, and they are used only when RS are considered unfeasible or unethical to conduct. With better research methods in observational studies and new tools for the evaluation of risk of bias, NRS are more likely to be a helpful source of information when used as replacement, sequential, or complementary evidence. This, together with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, provide an opportunity for guiding decisions about using RS and NRS in knowledge synthesis and increasing our certainty in a body of evidence.

This work aims to improve research synthesis methods by assessing the role and use of RS and NRS in knowledge syntheses using GRADE. This can help health professionals, researchers, guideline developers, and policy-makers build better and more complete healthcare recommendations.

# ACKNOWLEDGEMENTS

Many people supported me throughout this journey and it would be fruitless to try and flawlessly and fully express in this space all the feelings of fulfillment and gratitude that I have for them.

I would like to start by thanking three mentors that have shown me the value of innovation, leadership, and team work: my supervisor, Holger J. Schünemann, and the members of my thesis committee, Jan Brozek, and Gordon Guyatt. Without their support and guidance, this work would not have been possible.

I want to thank all my friends and peers in the HRM program and the department of Health Research Methods, Evidence, and Impact, for their friendship, support, and smiles: Rebecca, Ray, Juanjo, Gian Paolo, Wojtek, Itzi, Housne, Robby, Madison, Matthew, Ivan, Marcela, Alonso, Ignacio, Romina, Reem, Elie, John, Meha, Laura, Lawrence, Nancy, Bram, Charmaine, Deborah, Jennifer, Kristina, Lorraine, Heather, and all the staff from the department that made me feel at home.

I extend my gratitude to the people in Mexico who continued to support me in many ways. To CONACYT agency, both national and in the state of Nuevo Leon, for their financial support through the scholarship and for keeping an open door of communications. To the Dean of the Tecnológico de Monterrey School of Medicine, Dr. Jorge Valdez, all directors, alumni, and fellow professors who

supported me, and for their efforts to continue enhancing and exchanging knowledge and ideas to improve innovation and biomedical research in Mexico.

To my parents, to whom I owe my life; for teaching me to love what I do. To my siblings, José Manuel, Nancy, and Maria Teresa, from whom I learned to never give up and the value of supporting each other.

I want to thank the best and most efficient “motor” that propelled my Ph.D.: my kids Ana Sofia, Santiago, and Maria Elisa (and baby Lucas Adriano, on his way!).

Finally, to my wife and best friend, Yetiani, who decided to embark with me in the most insane and important decision of our lives (only second to getting married). I am lucky to have her and grateful for helping me write another chapter in the book of our lives.

Carlos Cuello

Hamilton Ontario, September 2017

# TABLE OF CONTENTS

LAY ABSTRACT .....	iii
ABSTRACT .....	iv
ACKNOWLEDGEMENTS .....	v
TABLE OF CONTENTS.....	vii
LIST OF FIGURES.....	xii
LIST OF TABLES .....	xiv
LIST OF APPENDICES.....	xv
LIST OF ABBREVIATIONS .....	xvi
DECLARATION OF ACADEMIC ACHIEVEMENT .....	xvii
CHAPTER 1. INTRODUCTION .....	1
1. Better evidence for better outcomes.....	2
2. Knowledge syntheses about health care interventions.....	3
3. Randomized and non-randomized studies in knowledge syntheses.....	4
4. GRADE and non-randomized studies .....	5
5. Why is this research important? .....	8
6. Goals and scope.....	10
7. Thesis overview .....	12
References .....	13
Figures .....	16



CHAPTER 2. A SCOPING REVIEW AND SURVEY PROVIDES THE RATIONALE, PERCEPTIONS AND PREFERENCES FOR THE INTEGRATION OF RANDOMIZED AND NON-RANDOMIZED STUDIES IN EVIDENCE SYNTHESSES.....	17
Abstract .....	19
1. Introduction.....	21
2. Methods.....	24
2.1 <i>Scoping review</i> .....	24
2.2 <i>Survey</i> .....	25
3. Results .....	27
3.1 <i>Scoping review</i> .....	27
3.2 <i>Survey results</i> .....	28
4. Discussion .....	30
4.1 <i>Summary of findings</i> .....	30
4.2 <i>How our results compare to other research?</i> .....	31
4.3 <i>New developments that impact on the existing perceptions and practice</i> .....	32
4.4 <i>Strengths and limitations</i> .....	32
4.5 <i>Implications for research</i> .....	33
5. Conclusions .....	33
<i>Acknowledgements</i> .....	34
<i>Funding sources</i> .....	34
References .....	35
Tables .....	40
Figures .....	43
Appendices .....	47
<i>Supplementary material 1: Search strategy for scoping review</i> .....	48
<i>Supplementary material 2. Complete survey</i> .....	51

CHAPTER 3. STRATEGIES TO OPTIMIZE USE OF RANDOMIZED AND NON-RANDOMIZED STUDIES IN EVIDENCE SYNTHESSES OF INTERVENTIONS USING GRADE .....	67
Abstract .....	68
1. Introduction.....	70
2. Methods.....	72
3. Differences in GRADE domains between randomized and non-randomized studies.....	76
3.1. Risk of bias.....	76
3.2. Inconsistency between a body of evidence from RS and NRS.....	78
3.3. Indirectness .....	81
3.4. Imprecision .....	81
3.5. Publication bias.....	83
3.6. Large effects, dose-response, and opposing residual confounding .....	83
4. Differences in direction and magnitude of effects between randomized and non-randomized studies .....	85
5. Presenting randomized and non-randomized studies in GRADE tables.....	86
6. Discussion .....	86
7. Conclusions .....	89
Acknowledgements.....	90
Funding sources .....	90
References .....	91
Tables .....	95
Figures .....	107
Appendices .....	120
Supplementary material 1 .....	121
Supplementary material 2. ....	127
Supplementary material 3 .....	134

CHAPTER 4. GRADE GUIDANCE: THE ROLE OF RANDOMIZED AND NON-RANDOMIZED STUDIES IN KNOWLEDGE SYNTHESIS OF HEALTH INTERVENTIONS.....	141
Abstract .....	143
Background .....	144
1. How to consider inclusion of non-randomized studies in knowledge syntheses .....	146
1.1. <i>The importance and role of a protocol and search strategy</i> .....	146
1.2. <i>When to include NRS (eligibility criteria)</i> .....	147
2. Optimal use of randomized and non-randomized studies .....	148
2.1. <i>Possible scenarios when dealing with two bodies of evidence</i> .....	148
2.2. <i>Using non-randomized rather than randomized studies</i> .....	148
2.3. <i>Using either or both types of studies</i> .....	150
3. Presenting in GRADE tables.....	151
3.1. <i>Alternative presentations</i> .....	151
3.2. <i>Combining both type of study designs</i> .....	151
4. The role of ROBINS-I .....	152
5. Summary and next steps .....	153
5.1. <i>Unresolved issues and next steps</i> .....	154
5.2. <i>Summary points</i> .....	155
Acknowledgments and conflict of interest.....	156
References .....	157
Tables .....	160
Figures .....	165
Appendices .....	175
<i>Supplementary material 1</i> .....	176
<i>Supplementary material 2</i> .....	178
<i>Supplementary material 3</i> .....	179
CHAPTER 5. CONCLUSIONS.....	180

Summary of findings .....	181
Implications for researchers, guideline developers, clinicians, patients, and policy-makers .....	182
Strengths and challenges of this work .....	184
Further research directions.....	186
Final remarks.....	188
References .....	189

# LIST OF FIGURES

## Chapter One

Figure 1. The GRADE approach –page 16.

## Chapter Two

Figure 1. Strategies used to integrate RS and NRS by experts in systematic reviews –page 44.

Figure 2. Participants agreement on reasons to include NRS with RS in health syntheses – page 45.

Figure 3. Four possible scenarios to integrate RS and NRS –page 46.

## Chapter Three

Figure 1. Forest plots. RS and NRS. Probiotic supplementation in preterm infants –page 108

Figure 2. Forest plots. RS and NRS. Antithrombin III for thromboprophylaxis –page 110.

Figure 3. Forest plots. RS and NRS. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants –page 112.

Figure 4. Forest plots updated –January 2017. RS and NRS. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants –page 114.

Figure 5. Overview of types of bias in RS and NRS and the actions or situations that protect against bias –page 116.

Figure 6. Inconsistency by type of studies (RS or NRS) –page 118.

## Chapter Four

Figure 1. Flowchart for the process of conducting a systematic review considering the role of RS and NRS –page 166.

Figure 2. Sixteen possible scenarios to encounter when evaluating bodies of evidence of RS and NRS – page 169.

Figure 3. Three possible presentations of both RS and NRS in GRADE evidence profiles – page 171.

Figure 4. Overview of types of bias in RS and NRS and the actions or situations that protect against bias – page 173.

# LIST OF TABLES

## Chapter Two

Table 1. Baseline characteristics of survey respondents –page 41.

Table 2. Other questions and results of the survey –page 42.

## Chapter Three

Table 1. Evidence profile table. Probiotics – page 96.

Table 2. Evidence profile table. Antithrombin III –page 98.

Table 3. Evidence profile table. Vitamin D. –page 100

Table 4. Evidence profile table. Vitamin D update –page 102

Table 5. Summary of findings table with two bodies of evidence –page 105.

## Chapter Four

Table 1. Evidence profile table. Antibiotics vs surgery for appendicitis – page 161

Table 2. Evidence profile table. Probiotics –page 162

Table 3. GRADE judgements and implications for integration of randomized and non-randomized studies –page 163

# LIST OF APPENDICES

## **Chapter Two**

Supplementary material 1. Search strategies for scoping review –page 48.

Supplementary material 2. Complete survey. –page 51.

## **Chapter Three**

Supplementary material 1. Example 1 –page 121.

Supplementary material 2. Example 2 –page 127.

Supplementary material 3. Example 3 –page 134.

## **Chapter Four**

Supplementary material 1. Appendix 1 –page 176

Supplementary material 2. Appendix 2 –page 178

Supplementary material 3. Appendix 3 –page 179



# LIST OF ABBREVIATIONS

CI: Confidence Interval

CoE: Certainty of Evidence

CPG: Clinical Practice Guideline

EP: Evidence Profile

EtD: Evidence-to-Decision

G.I.N.: Guidelines International Network

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HTA: Health Technology Assessment

IQR: interquartile range

NRS: Non-randomized study

PICO: population, intervention, comparison and outcome

RCT: randomized controlled trial

RS: Randomized Study

SoF: Summary of Findings

WHO: World Health 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 conduction of the studies here included. We obtained feedback and advice from Professors Guyatt and Brozek, as well as from members of the GRADE working group and Cochrane methods group for the assessment of bias in non-randomized studies.

This work 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 work: design, conception, and writing of surveys; I designed the search strategy, screening, and data extraction for scoping the review about the history and rationale of using NRS in health syntheses. I performed the qualitative analyses and reviewed comments and feedback from experts during the conduction of focus groups and meetings with the GRADE and Cochrane groups.

I conducted all analyses, designed figures and tables, and organized meetings. I wrote the manuscript with editorial advice and supervision of Professor Schünemann, and from feedback from Professors Brozek and Guyatt. The co-

authors on each paper contributed significantly with important comments and advice for the final manuscripts (details with each manuscript).

For all the three manuscripts composing this “sandwich” thesis, earlier drafts of parts of this research have been presented at international academic conferences as part of the manuscript’s development. The first two papers are under review and submitted to peer-reviewed journals. The third one is the final draft of official GRADE guidance which will be submitted to the GRADE working group for review and approval.

# CHAPTER 1. INTRODUCTION

---

## **1. Better evidence for better outcomes**

Better health outcomes require health professionals and policy-makers reaching the right decisions and providing the best recommendations with the best research evidence available. Failing to base recommendations on the best research evidence risks transmitting incomplete, misguided, or biased information. This applies at all levels of the healthcare system: from the care of individuals to the creation of policies for a whole population.

Better methods to synthesize evidence will help practitioners and policy-makers to keep abreast of the literature related to their topic or problem. One of the main challenges of the 21st century is to guarantee a process that facilitates the transfer of high-quality evidence from research into effective changes in health policy, clinical practice, or products; in other words, how to ensure an adequate knowledge translation process.<sup>1</sup>

Looking for alternate or new methods in evidence synthesis is desirable and necessary, and can help patients and all decision-makers understand the outcomes associated with their treatment choices and the consequences of their decisions.

## 2. Knowledge syntheses about health care interventions

Knowledge synthesis is defined as any systematic review, rapid review, health technology assessment,\* or any other attempt to summarize all pertinent studies on a specific question.<sup>2</sup> In the last decades, knowledge syntheses, in the form of systematic reviews, HTAs, and clinical practice guidelines, have had an important role in shaping health policies, and improving the process of translating evidence into clinical practice.<sup>3</sup> Knowledge syntheses rely on an adequate source of research in the form of individual studies that authors use and distinguish as two main types, based on their design: randomized (RS) and non-randomized studies (NRS).

Many of the recommendations provided by knowledge syntheses are related to health care interventions;† for example, if clinicians should recommend antibiotics to all patients undergoing mechanical ventilation in the intensive care unit; or if nurses should delay oral feeding to all patients with acute pancreatitis; or if a policy related to the procurement of magnesium sulfate for the treatment of eclampsia in low income countries should be put in place after recommending being part of the World Health Organization list of essential drugs.

---

\* We will use these terms indistinctively throughout this dissertation.

† In this dissertation, the focus will be on healthcare interventions, i.e., medications, behavioral interventions, etc. as opposed to exposures, which are evaluated elsewhere.

### 3. Randomized and non-randomized studies in knowledge syntheses

Through history, the effectiveness of health care interventions had not been consistently and scientifically tested until the mid-eighteenth century, when James Lind first examined the effect of citrus fruits as treatment for scurvy by conducting the first properly documented clinical trial.<sup>4</sup> Since then, most studies evaluating interventions were tested in a non-randomized fashion. The first acknowledged transition from alternation to randomization in clinical studies occurred in 1946 to 1948 with what is considered the first published randomized controlled trial,<sup>5</sup> where famous statistician, Sir Austin Bradford Hills, decided to randomly allocate patients with tuberculosis to bed rest plus streptomycin versus bed rest alone.<sup>‡ 6, 7</sup> Subsequently, RS and NRS have been considered entirely different (sometimes, even opposites) methodologies in health research. The RS became the preferred type of individual study to include in knowledge syntheses of interventions because of the widely-known risk of bias and confounding that NRS carry due to lack of randomization to assign participants to the study interventions. According to one report,<sup>8</sup> the empirical observations that NRS significantly differ and would give dissimilar results from RS originated from studies performed during the 1970s and

---

<sup>‡</sup> Interestingly, Bradford-Hills' main motivation was not to avoid confounders but to "better conceal the allocation schedule."

1980s stating that NRS tend to inflate positive results as compared to RS; for example, in the study by Chalmers et al.,<sup>9</sup> 56% of non-randomized studies produced favourable treatment effects, as compared with 30% of blinded, randomized, controlled trials; the study by Sacks<sup>10</sup> determined that while 20% of randomized studies found a benefit of a therapy studied, 79% of non-randomized studies (assessing the same comparisons) concluded the same treatment was successful. Other studies from the same period gave similar results.<sup>11</sup>

As we will review in Chapter 2, these views were challenged with new reports assuming newer, stricter, and more sophisticated methods used in later decades, and up until today, exercising better statistical analyses and using better computerized data sets to perform NRS.<sup>12</sup> These measures may serve to reduce systematic error commonly found in NRS.<sup>8</sup> Furthermore, recent evidence suggests that many additional factors, other than study design and execution, should be considered when determining the possibility of integrating<sup>§</sup> RS with NRS.<sup>13</sup>

## 4. GRADE and non-randomized studies

The GRADE methodology emerged as an instrument for evaluating a body of research evidence, and has now been adopted by many organizations.<sup>14</sup> GRADE is

---

<sup>§</sup> For the remainder of this dissertation, unless stated otherwise, we will use the term “integration” referring to any form of using RS and NRS together, either in the same synthesis, in the same summary of findings (same table but separated in rows), or in the same analysis (pooled into a single estimate).



utilized to evaluate a group of studies, either RS or NRS. The premise of this method is to base their assessments by rating the importance of the outcomes, and based on these, proceed to rate a body of evidence for each outcome of the PICO question. The GRADE approach then rates the certainty of the evidence, also called “quality” or “confidence” (see box 1). GRADE uses eight different domains; five that upgrades the certainty in the evidence, and three that downgrades it (figure 1) to obtain a final overall rating of the evidence. The final ratings of the certainty of the evidence can be classified as ‘high’, ‘moderate’, ‘low’, or ‘very low’. An important notation is the different baseline rating GRADE assigns by default to a body of evidence; if this is from RS, the certainty will start as ‘high’, while if it is from NRS, it will start as ‘low’, as shown in figure 1.

**Box 1.** In the context of a systematic review, the ratings of the 'quality' or 'certainty' of the evidence reflect the extent of our confidence that the estimates of the effect are correct.

In the context of making recommendations, the certainty of the evidence reflects the extent of our confidence that the estimates of an effect are adequate to support a decision or recommendation.

This is important and central to this dissertation; it is a core element related to the reasons NRS are considered as providing less certainty, mainly due to the lack of randomization generating the risk of confounding.

An important instrument that will be mentioned in this work is the new tool for assessing the risk of bias in non-randomized studies of interventions (ROBINS-I).<sup>15</sup>

Many tools have been developed for evaluating the risk of bias in individual studies classified as “observational” or non-randomized. To this day, more than 200 exist.<sup>16</sup>

<sup>17</sup> However, as we will review in chapters 3 and 4, all these tools starts by comparing the study being evaluated to what it would be considered an ‘ideal’ observational study. ROBINS-I, on the other hand, starts by asking: “How would this observational study be performed, if it was possible to do it in a randomized fashion?” and, “how it compares to this ‘ideal’ trial (also called ‘target’ trial)?” In other words, ROBINS-I uses an absolute scale of risk of bias that includes all types of studies –RS and NRS. This implies that the assessment of confounding and selection bias are integral parts of the tool. ROBINS-I has undergone careful development by a large group of experienced investigators. It has been tested and scientists have begun to validate it, and experience will continue to accumulate. It will represent a significant part of the discussion in this work. For a clarification in the terminology used between GRADE and ROBINS-I, see the definitions in box 2.

**Box 2. Analogies in the terminology used between GRADE and ROBINS-I.**

Based on GRADE guidance, we will use the term GRADE “*criteria*” for all criteria in the evidence to decision frameworks of GRADE (within these criteria, the “certainty in the evidence” is one criterion). Certainty of the evidence is assessed based on “certainty *domains*” with individual *items* within each domain. Risk of bias is one domain, therefore, in the context of GRADE, we will use the term *risk of bias items* to describe the seven areas of judgment that ROBINS-I calls domains.

## 5. Why is this research important?

Knowledge syntheses provide vital information for decision-making in many healthcare areas, and it is important to ensure that their methods are appropriate and the information they provide is complete and trustworthy. Although researchers consider RS the first choice to include in syntheses of interventions, they might fail to observe the whole picture if NRS are always excluded.

Assessing the appropriateness of integrating RS and NRS in knowledge syntheses is important and with several advantages. Vital information contained in observational studies can increase our certainty in the effect estimates by complementing the information from RS, replacing it, or used sequentially.<sup>18</sup> This would have a domino effect in the knowledge translation process when stakeholders and decision-makers

will have a better understanding of a more complete body of evidence as tools to improve health outcomes.

Today, most authors conducting a knowledge synthesis about health interventions start by setting their eligibility criteria for the included studies to only RS and stop the process of collecting information if they do not find adequate data applicable to their clinical question, and many do not even consider other sources of evidence.<sup>19</sup> Nonetheless, researchers conducting evidence syntheses will frequently encounter well-known caveats of RS such as unavailability due to impracticability or ethical issues.<sup>19</sup> Given these situations researchers must consider relying on evidence obtained from NRS. Currently, many authors already integrate RS and NRS in knowledge syntheses of interventions,<sup>16, 19</sup> but in many of them, the reasons to include or exclude NRS, or even the methods to achieve their integration with RS, are not discussed in detail. Guidance is desirable and, moreover, necessary.

There have been attempts to guide authors of knowledge syntheses to address the issue of integrating NRS and RS.<sup>16, 17</sup> However, they only address the methodological issues considering individual studies, and do not contemplate the certainty in a body of evidence; furthermore, it is not well explored how the GRADE approach, together with the new tool for assessing the risk of bias in NRS of interventions (ROBINS-I), might support considering these diverse evidence streams.<sup>15</sup>

## 6. Goals and scope

This dissertation aims to:

- a. Provide an overview, including the historical background and rationale for including NRS with RS in systematic reviews of interventions, and obtain the perceptions, preferences, and practices from experts in synthesizing evidence, including what it is currently done, and the reasons behind it. Also, to review.
- b. Analyze the options for the optimal use of RS and NRS in health syntheses by using the GRADE approach in evidence profiles and summary of findings tables, by evaluating different methodological challenges of the integration, and the possible solutions. This will include analyzing the differences between RS and NRS in the GRADE domains (including risk of bias), differences in direction and magnitude of the effects, and how to portray both bodies of evidence in GRADE tables, i.e., summary of findings (SoF) tables and evidence profiles (EP).
- c. Generate a sensible and comprehensive guidance that is feasible and easy to follow for all authors invested in knowledge syntheses of healthcare interventions; either from the perspective of systematic reviews, or that from health guidelines.

To achieve these goals, there are three main works around each of these three main topics. The first part consists of a scoping review of the literature, including background and historical perspective, followed by a survey to experts in the field of knowledge syntheses (from Cochrane, G.I.N., the WHO, and the GRADE working group) to obtain their preferences of when and how they integrate RS and NRS.

The second goal is achieved through qualitative analyses of discussions with experts in the field that took place during different GRADE and Cochrane meetings and conferences over the last three years. During these meetings, an initial guidance was drafted and refined with real scenarios of knowledge syntheses through an iterative process of discussion and feedback. This second paper assesses each challenge encountered in the process of a possible integration of RS and NRS using real life research questions as case studies and discussing the differences between RS and NRS within the GRADE domains and other issues such as the feasibility of the process.

The third, and final part, is the overarching product that is presented to the GRADE working group as official guidance. This includes guidance from the outset, at the protocol stage of a systematic review to the assessment and presentation of both bodies of evidence assessing specific scenarios; for example, what to do if a systematic review author finds a body of evidence of RS with low certainty but a body of evidence of NRS with moderate certainty?

The results of this project are important for methodologists, statisticians, researchers, clinicians, policy-makers, and every professional that is confronted with evidence from both RS with NRS in a health-related knowledge synthesis

## **7. Thesis overview**

This dissertation is about improving methods for synthesizing evidence. Its focus is on increasing the certainty of a body of evidence, when appropriate, with the integration of RS and NRS. This work is divided in three main sections based on the three goals described above. These topics are evaluated and discussed in Chapters 2, 3, and 4 of this thesis, with an overall discussion and conclusion in Chapter 5.

Chapter 2 assesses the first goal, i.e., the perceptions, behaviors, common practice and preferences from experts. Chapter 3 deals with assessment, presentations, and discussion of the different methods that will help authors with the integration of RS and NRS. Chapter 4 is the final proposed guidance of the GRADE working group. Chapter 5 will summarize and provide conclusions of this project and speaks to the challenges and future steps.

## References

1. Lang ES, Wyer PC, Haynes RB. Knowledge translation: closing the evidence-to-practice gap. *Ann Emerg Med.* 2007;49(3):355-363.
2. Kastner M, Tricco AC, Soobiah C, Lillie E, Perrier L, Horsley T, et al. What is the most appropriate knowledge synthesis method to conduct a review? Protocol for a scoping review. *BMC Med Res Methodol.* 2012;12:114.
3. Oxman AD, Lavis JN, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 1: What is evidence-informed policymaking? *Health Res Policy Syst.* 2009;7 Suppl 1:S1.
4. Dunn PM. James Lind (1716-94) of Edinburgh and the treatment of scurvy. *Arch Dis Child Fetal Neonatal Ed.* 1997;76(1):F64-65.
5. D'Arcy Hart P. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *BMJ.* 1999;319(7209):572-573.
6. Chalmers I. Why transition from alternation to randomisation in clinical trials was made. *BMJ.* 1999;319(7221):1372.
7. Bhatt A. Evolution of clinical research: a history before and beyond James Lind. *Perspect Clin Res.* 2010;1(1):6-10.
8. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000;342(25):1878-1886.



9. Chalmers TC, Celano P, Sacks HS, Smith H, Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med*. 1983;309(22):1358-1361.
10. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. *Am J Med*. 1982;72(2):233-240.
11. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med*. 1989;8(4):455-466.
12. Colditz GA. Overview of the Epidemiology Methods and Applications: Strengths and Limitations of Observational Study Designs. *Critical Reviews in Food Science and Nutrition*. 2010;50:10–12.
13. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014;4:MR000034.
14. 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):924-926.
15. 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.
16. Peinemann F, Kleijnen J. Development of an algorithm to provide awareness in choosing study designs for inclusion in systematic reviews of healthcare interventions: a method study. *BMJ Open*. 2015;5(8):e007540.

17. Norris S, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Selecting Observational Studies for Comparing Medical Interventions. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville (MD)2008.
18. 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. *Res Synth Methods*. 2013;4(1):49-62.
19. Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. *J Clin Epidemiol*. 2014;67(6):645-653.

## Figures

Figure 1. The GRADE approach

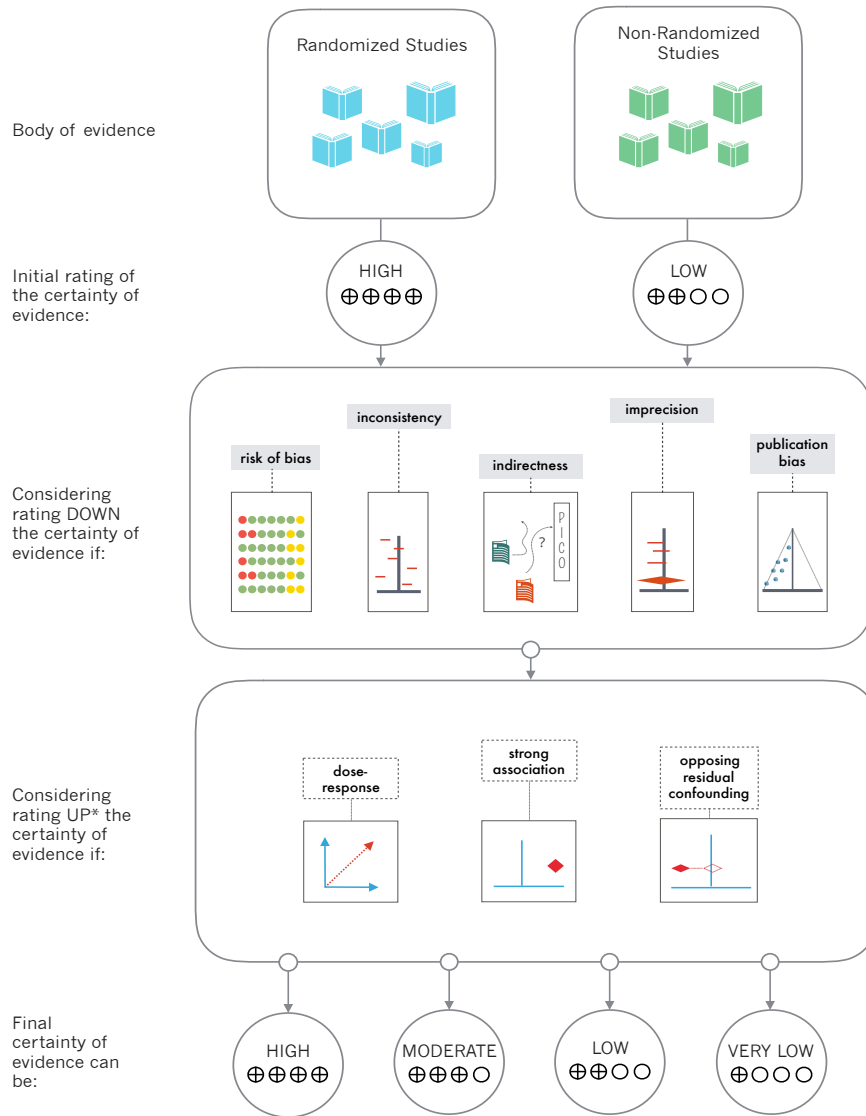


Figure 1. The GRADE approach to rate a body of evidence of randomized or non-randomized studies for a health synthesis. \*GRADE suggest usually rating up only non-randomized studies. See text for description.

# CHAPTER 2. A SCOPING REVIEW AND SURVEY PROVIDES THE RATIONALE, PERCEPTIONS, AND PREFERENCES FOR THE INTEGRATION OF RANDOMIZED AND NON-RANDOMIZED STUDIES IN EVIDENCE SYNTHESSES

## Authors

Carlos A. Cuello-Garcia<sup>a</sup>, Rebecca L. Morgan<sup>a</sup>, Jan Brozek<sup>a</sup>, Nancy Santesso<sup>a</sup>, Jos Verbeek<sup>b</sup>, Kris Thayer<sup>c</sup>, Gordon Guyatt<sup>a</sup>, and Holger J. Schünemann<sup>a</sup>

## Affiliations

- a. Department of Health Research Methods, Evidence, and Impact, McMaster University. Hamilton Ontario, Canada.
- b. Cochrane Work Review Group. Finnish Institute of Occupational Health, Helsinki, Finland
- c. National Center for Environmental Assessment. Environmental Protection Agency. USA

## **Corresponding Author**

Holger J. Schünemann, M.D., Ph.D.

Chair, Department of Health Research Methods, Evidence, and Impact.

McMaster University. Health Sciences Centre Room 2C16.

1280 Main Street West. Hamilton, Ontario Canada. L8N 4K1

**Word count:** 3,191

## Abstract

**OBJECTIVES:** Health care decision makers may need evidence from randomised and non-randomised studies to understand the effects of interventions. Our objective is to review the literature and obtain preferences and perceptions from experts regarding the role of randomized studies (RS) and non-randomized studies (NRS) in systematic reviews of intervention effects.

**STUDY DESIGN AND SETTING:** We conducted a scoping review and surveyed experts in this field. Using the levels of certainty framework developed by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, experts expressed their preferences and decisions about the use of RS and NRS in health syntheses.

**RESULTS:** Of 189 initial respondents, 123 had the expertise required to answer the questionnaire; 116 provided their extent of agreement with approaches to use of NRS with RS. Most respondents would include NRS when RS was unfeasible (83.6%) or unethical (71.5%) and a majority to maximize the body of evidence (66.3%), compare results in NRS and RS (53.5%) and to identify subgroups (51.7%). Sizable minorities would include NRS and RS to address the effect of randomization (29.5%) or because the question being addressed was a public health intervention (36.5%). In summary of findings (SoF) tables, most respondents would include both

bodies of evidence – either separately in the same table or in separate tables – when RS provided moderate, low, or very-low certainty evidence; even when RS provided high certainty evidence, a sizable minority (25%) would still present results from both bodies of evidence. Very few (3.6%) would ever, under realistic circumstances, pool SR and NRS results together.

**CONCLUSIONS:** Most experts would include both RS and NRS in the same review under a wide variety of circumstances, but almost all would present results of two bodies of evidence separately.

#### **KEYWORDS**

Systematic reviews; Randomized trials; Non-randomized trials; GRADE; clinical practice guidelines; Research methodology.

## 1. Introduction

Randomized studies (RS) provide the best source of evidence for systematic reviews of healthcare interventions. <sup>1</sup> Non-randomized studies (NRS) addressing the effects of interventions are defined as “quantitative studies estimating effectiveness (harm or benefit), which did not use randomization to allocate units (individuals or clusters of individuals) to comparison groups” <sup>2</sup>.

From an epistemological point of view, RS and NRS of interventions or exposures have more common traits than dissimilarities. Both aim at providing insight about ‘causation’, i.e., an intervention or exposure that is truly linked to an effect. In rare situations (e.g., epinephrine for anaphylactic shock, or dialysis for terminal renal failure), large effects of interventions will establish causation, making confounding and other biases of less concern and providing certainty in a large effect. In these cases, RS will be unnecessary - and unfeasible and unethical - to conduct. Moreover, for purposes such as assessing risk factors, and obtaining prognostic information and baseline risks, NRS provide the optimal study design. <sup>3,4</sup>

Unfortunately, most health interventions have moderate to small effects (i.e., risk reduction), with biases, imprecision, or inconsistency among studies that will restrict the certainty in a body of evidence. <sup>5-7</sup> Although both RS and NRS try to use adequate sample sizes to minimize imprecision, and methodological rigor to reduce



bias, most experts consider concealed randomization as the best methods that can protect fully against confounding.<sup>2</sup> The extent of the superiority of RS over NRS is, however, a matter of dispute. NRS try to deal with confounding using multivariable analyses (standard regression and propensity), and matching. However, many experts believe that, even after optimal adjusted analysis, residual confounding remains a major issue reducing certainty of evidence, and thus consider RS far superior to observational studies in establishing causation.<sup>8, 9</sup> Others authorities describe NRS as “real world research”, and that newer NRS analysis approaches employing causal inference techniques, such as instrumental variables, marginal structural models, propensity scores, among others, can to a certain extent address concerns of selection bias and confounding,<sup>4, 10, 11</sup> and that in particular circumstances, NRS may provide higher certainty than RS or help with judging the confidence in RS.

Experts from both camps have addressed reasons for using both RS and NRS in the same systematic review, or exclusively using NRS, and suggested approaches for so doing.<sup>5, 12-17</sup> For example, a recent framework,<sup>5</sup> using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) certainty of evidence as guide, suggests that NRS evidence can prove useful in a systematic review addressing the causal effect of an intervention in the circumstances described in Box 1.

**Box 1. Role of NRS in systematic reviews**

**Replacement:** when NRS are used instead of RS for several reasons; i.e., because they are unethical, not available or they are of very poor quality, and thus NRS provide the best available evidence.

**Sequential:** when RS provide the best evidence for some outcomes, but not for others (rare or possibly long-term outcomes).

**Complementary:** when RS leave effects uncertain in patient groups (i.e., children, pregnant women, the elderly) not included in RS.

There is also a strong opinion among policy-makers and stakeholders that the evidence from NRS should be used more efficiently, especially when RS could be absent.<sup>18</sup> However, little is known about the practice and concerns by experts and if using NRS and RS is appropriate. There is debate about how the evidence from RS and NRS should be presented when both are available, traditionally, in summary of findings tables (SoF) and evidence profiles. From its inception, the GRADE approach has provided a framework for how to assess a body of evidence from either RS or NRS, but this evidence is assessed and presented to decision makers separately.

In this work, we utilized a mixed-methods approach to understand expert perspectives on the use of RS and NRS in knowledge synthesis. First, we performed

a scoping review limited to reviews and overviews that describe the differences between RS and NRS of healthcare interventions. Second, we obtained and analyzed the knowledge, attitudes, and perceptions of systematic review authors about regarding the use of RS and NRS by focusing on different levels of certainty of the evidence and how to present in GRADE tables. We exclusively focus on NRS of interventions (as opposed to exposures). This paper is part of a broader project to provide guidance on the appropriateness and methods for using RS and NRS in knowledge syntheses.

## **2. Methods**

### **2.1 Scoping review**

We performed a focused search for overviews and systematic reviews to assess the differences and similarities between RS and NRS, this is, whether RS and NRS have different effect estimates on the same health question. We also looked for essays and narrative reviews for a historical background perspective of the reasons to include NRS in knowledge syntheses. We searched The James Lind Library (from inception to July 2017), Medline (from inception up to July 2017), and the Cochrane Library (from inception up until July 2017) using an updated search strategy based on a recent Cochrane overview.<sup>4</sup> We searched for additional references in included reviews. We used the “similar articles” and “citing articles” features from Medline

to identify additional relevant articles. From these we checked references related to the history of systematic reviews and the study design of the included studies (see Appendix 1 for a complete description of the search strategies). We exclusively focused on NRS of interventions (as opposed to environmental or occupational exposures, i.e., unintentional interventions).

## **2.2. Survey**

We contacted experts in knowledge synthesis, i.e., from Cochrane, the Guideline International Network (GIN), and mailing lists of recognized institutions that regularly produce evidence syntheses (e.g., World Health Organization). We involved members of the GRADE Non-Randomized Risk of Bias Project Group using online meetings and during workshops at the GRADE working group meetings (Cochrane Colloquium in Vienna 2015, and Washington D.C. in May 2016). We surveyed experts from the aforementioned organizations to obtain their understanding, attitudes, and preferences about the use of NRS and RS in knowledge syntheses as well as their inclusion within GRADE summary of findings tables. We pilot-tested a questionnaire among 20 methodologists from McMaster University and the Cochrane Methods group. Based on their feedback, we modified the survey to improve the clarity and relevance of the questions. The final online version consisted of five introductory and demographic questions, five questions addressing their expertise conducting systematic reviews (with and without NRS),

and the rest of 13 questions about their understanding, attitudes, and preferences regarding the use and integration of NRS with RS applying the GRADE approach to four hypothetical scenarios.

One of the latter group of questions asked respondents whether they had included RS and NRS in a systematic review, and if so, the approach they had used (e.g., in a single analysis; in side-by-side analyses). Four questions asked respondents for narratives regarding reasons for including or excluding NRS from systematic reviews of RS. One question asked respondents for their level of agreement (strongly agree to strongly disagree) regarding eight reasons (formulated as eight questions) for including NRS in a systematic review. These reasons were obtained from previous analyses and the Cochrane handbook.<sup>5, 16</sup> We matched these reasons to one of the three categories related to the use of NRS in systematic reviews (box 1). Question 1 was categorized to the use of NRS as “sequential” evidence; question 2 reflected the use of NRS as “replacement”, and question 5 referred to the use of NRS as “complement” evidence in systematic reviews. The rest of questions (questions 3, 4, 6, 7, and 8) were grouped and considered as methods and applicability questions.

Respondents were then presented with four scenarios (e.g., high certainty evidence is available for both RS and NRS) and asked how they would handle the situation (combine and present in a single analysis; present separately in single SoF table; present separately in two SoF tables; use only RS; or use only NRS).

We estimated an approximate finite number of 150 to 200 experts, based on distribution lists from the organizations aforementioned, that would have potential expertise and knowledge on the GRADE approach and systematic reviews; this was considered the target population. Estimating a response rate of 50% and a confidence interval of 95%, a total of 132 respondents were needed. We performed descriptive analyses to summarize participant characteristics. Our estimates are based on available cases. We used the  $\chi^2$  test with contingency tables to compare the proportion of responses between various groups of questions according to different levels of certainty in the evidence among RS and NRS.

We performed analyses using SPSS version 21 (IBM SPSS Statistics for Mac) and Excel for descriptive statistics. The survey was anonymous and informed consent was waived by the ethics committee serving McMaster University. We received support from a Cochrane Methods Innovation grant, the National Toxicology Program within the National Institutes for Health, and the McMaster GRADE centre.

## **3. Results**

### **3.1. Scoping review**

Among 15,645 references from Cochrane and Medline, we found one Cochrane overview assessing the differences between RS and NRS,<sup>4</sup> which includes 14

systematic reviews comparing effect estimates between RS and NRS. No further reviews on this topic were found. We assessed 184 records, 152 articles, and 23 essays in The James Lind Library; from these we found that the earliest documented transition from alternation to randomization for allocating participants in clinical studies could be traced to 1948<sup>19, 20</sup> with the first randomized trial on streptomycin for tuberculosis. Since then, scientists have regarded NRS and RS as distinct (sometimes even opposite extremes) study types for inclusion in systematic reviews.<sup>21</sup> The empirical observations that NRS significantly differ from RS originated from studies performed during the 1970s and 1980s stating that NRS tend to inflate positive results as compared to RS.<sup>22-25</sup> These views were soon challenged when NRS from later decades were assumed to use newer, stricter, and more sophisticated techniques, with better statistical analyses and computerized datasets.<sup>3, 26-34</sup> The recent overview by Anglemyer<sup>4</sup> analyzed 14 systematic reviews, of which 11 (79%) found no significant difference between RS and NRS, one suggested that NRS had larger effects, and two found to have smaller effects.

### **3.2. Survey results**

A total of 189 experts in the field of systematic reviews and clinical guidelines were approached and invited to participate, of whom 138 completed the survey. One hundred and twenty-three of those completing the survey (89%) had the minimum required experience (at least one systematic review conducted in the last five years)

for inclusion in our analysis. Of the 123 respondents, 108 (87%) had conducted at least three systematic reviews and 112 (91%) described using GRADE and summary of findings tables on a regular basis (table 1).

Eighty of the 123 respondents had conducted at least one systematic review that included both RS and NRS (see figure 1). The most frequent approach they used (39 of 80 respondents –48.8%) was to conduct separate meta-analyses (one for each body of evidence). Of the 80 respondents, 14 (17.5%) had, on at least one occasion, pooled RS and NRS in a single meta-analysis.

Of the respondents, 116 reported the extent of their agreement with reasons for including RS and NRS in the same review<sup>5,16</sup> (figure 2) and were classified according to the criteria in box 1. Of the 116 respondents 97 (83.6%) strongly agreed or somewhat agreed to include a body of evidence from NRS if it serves as sequential information, 83 (71.5%) if NRS serves as replacement, and 77 (66.4%) if it serves as complement.

One hundred and twelve participants assessed four possible situations for the presentation of bodies of evidence from both RS and NRS presented as four simulated scenarios (depicted in figure 3). In three of the four scenarios (A, C, and D from figure 3), most respondents (52 [72%], 98 [86.8%], and 88 [78%] respectively) would present data from both RS and NRS separately, either in a single SoF table



or each in its own table. When high certainty evidence was available from RS but lower certainty from NRS, preferences varied markedly: 45 (40.2%) would use only RS, but 64 (57.1%) would present both sets of data separately, either in the same or separate SoF. Responses to this scenario differed significantly from any of the others ( $p < 0.001$ ).

We asked participants if they would ever consider pooling RS with NRS in meta-analyses in any of these scenarios. Of 102 respondents, 51 (50%) would not consider combining, and most others would consider combining only if both bodies of evidence were of high quality (table 2). We analyzed the comments from those willing to pool both bodies of evidence in a meta-analysis; nine experts agreed with pooling as long as both had low risk of bias and performing a sensitivity analysis; also, when a perfect matching of the clinical (PICO) question was found ( $n=4$ ), no heterogeneity ( $n=6$ ), similar direction of effect ( $n=2$ ), or when the RS was poorly conducted study and the NRS would be of low risk of bias ( $n=3$ ). Table 2 presents other responses to the survey questions.

## **4. Discussion**

### **4.1. Summary of findings**

In this study, we evaluated experts' views on the inclusion of RS and NRS in systematic reviews. Of the respondents, 90% used GRADE summary of findings on

a regular basis, and 65% had included RS and NRS in at least one review. Most respondents would include NRS when RS was unfeasible or unethical, and a majority to maximize the body of evidence, compare results in NRS and RS and to identify subgroups (Figure 2). Sizable minorities would include NRS and RS to address the effect of randomization or because the question being addressed was a public health intervention

In summary of findings (SoF) tables, most respondents would include both bodies of evidence – either separately in the same table or in separate tables – when RS provided moderate or low certainty evidence; even when RS provided high certainty evidence, a sizable minority would still present results from both bodies of evidence (Figure 3). Although half of respondents would consider pooling RS with NRS when the latter are of high certainty, this was considered an unlikely scenario and very few would ever, under realistic circumstances, pool RS and NRS (Table 2).

## **4.2. How our results compare to other research?**

Previous research has mainly focused on analyzing the frequency, methodological constraints, and suggestions for use of NRS in systematic reviews of effects of interventions.<sup>5, 12-16</sup> In our study, we aimed to obtain the knowledge, attitudes, and perceptions from experienced systematic review authors and guideline developers about the integration of evidence from NRS with RS in systematic reviews at

different levels. To our knowledge this is the first attempt to systematically obtain such data from experts in research methodology.

### **4.3. New developments that impact on the existing perceptions and practice**

Obtaining higher certainty evidence, at least for some outcomes (such as infrequently occurring serious adverse effects) provides one compelling rationale for including NRS in systematic reviews of interventions either as complement, sequential or replacement for RS<sup>5</sup> While this integration is occurring in practice, improvement in the risk of bias assessment and statistical analysis of observational methods, and enthusiasm for comparative effectiveness addressed in observational studies, has reenergized the debate regarding the relative merits of RS and NRS, and thus considerations regarding their inclusion together in systematic reviews. One such development, ROBINS-I<sup>2</sup>, provides a new instrument for evaluating the risk of bias in NRS of interventions; it is based on the premise that any NRS could be compared to an ideal RS called ‘target trial’. This implies that a NRS with low risk of bias is theoretically equivalent to a well-performed RS. Thus, it allows an assessment of risk of bias in NRS on the same metric as RS. Given this, there is a theoretical increased likelihood of considering RS and NRS to provide similar certainty of evidence, and thus a stronger rationale for their inclusion in the same systematic review.

### **4.4. Strengths and limitations**

We constructed our questionnaire based on a review of prior literature and the GRADE conceptual framework, and pre-tested the questionnaire to ensure clarity and ease of completion. We achieved a high response rate. Our survey is limited in the total sample size given the relatively small number of systematic review authors with expertise in the GRADE approach and in use of NRS in systematic reviews.

#### **4.5. Implications for research**

These results will provide a baseline assessment from which further work by a GRADE project group and the Cochrane GRADE Methods group, supported by a Cochrane Methods Innovation Fund, will determine the best approaches for use of RS and NRS in systematic reviews and Summary of Findings Tables. We will address how the use of NRS affects GRADE domains and the overall certainty of evidence. For example, how should NRS that address concerns about indirectness in RS but are at higher risk of bias be assessed and integrated in the SoF table and affect the overall certainty in the evidence? We aim to continue investigating the best mode of integration by using GRADE.

### **5. Conclusions**

Experts see a wide variety of circumstances in which RS and NRS can provide complementary, sequential or replacement information in systematic reviews, including their presentation in GRADE SoF tables. Future research will evaluate the

specific assessment on each GRADE domain between both bodies of evidence and help with the utilization of new risk of bias tools. The work to improve observational methods and the GRADE approach for evaluating and presenting this evidence provides new and exciting opportunities to move forward from the traditionally historical perspective of keeping these two bodies of evidence separate and distinct.

--

## **Acknowledgements**

We are grateful to McMaster University fellows, graduate students, and professionals who helped piloting the survey and providing feedback.

## **Funding sources**

This work was supported by a Cochrane Methods Innovation grant, the National Toxicology Program within the National Institutes for Health, and the McMaster GRADE centre. The sponsors had no role in the design of the study or interpretation of the results except for through the lead authors of this article.

## References

1. Prasad V, Jorgenson J, Ioannidis JP, Cifu A. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *J Clin Epidemiol*. 2013;66(4):361-366 e364.
2. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *BMJ*. 2016(355):i4919.
3. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342(25):1878-1886.
4. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014;4:MR000034.
5. 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. *Res Synth Methods*. 2013;4(1):49-62.
6. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
7. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med*. 1984;3(4):409-422.

8. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ*. 2007;334(7589):349-351.
9. Prasad V, Berger VW. Hard-Wired Bias: How Even Double-Blind, Randomized Controlled Trials Can Be Skewed From the Start. *Mayo Clin Proc*. 2015;90(9):1171-1175.
10. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J*. 2012;33(15):1893-1901.
11. Kitsios GD, Dahabreh IJ, Callahan S, Paulus JK, Campagna AC, Dargin JM. Can We Trust Observational Studies Using Propensity Scores in the Critical Care Literature? A Systematic Comparison With Randomized Clinical Trials. *Crit Care Med*. 2015;43(9):1870-1879.
12. Norris S, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Selecting Observational Studies for Comparing Medical Interventions. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville (MD)2008.
13. O'Neil M, Berkman N, Hartling L, Chang S, Anderson J, Motu'apuaka M, et al. Observational evidence and strength of evidence domains: case examples. *Syst Rev*. 2014;3:35.
14. Treadwell JR, Singh S, Talati R, McPheeters ML, Reston JT. A framework for best evidence approaches can improve the transparency of systematic reviews. *J Clin Epidemiol*. 2012;65(11):1159-1162.

15. Higgins JP, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Res Synth Methods*. 2013;4(1):12-25.
16. Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. *J Clin Epidemiol*. 2014;67(6):645-653.
17. Peinemann F, Kleijnen J. Development of an algorithm to provide awareness in choosing study designs for inclusion in systematic reviews of healthcare interventions: a method study. *BMJ Open*. 2015;5(8):e007540.
18. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why observational studies should be among the tools used in comparative effectiveness research. *Health Aff (Millwood)*. 2010;29(10):1818-1825.
19. Chalmers I. Why transition from alternation to randomisation in clinical trials was made. *BMJ*. 1999;319(7221):1372.
20. D'Arcy Hart P. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *BMJ*. 1999;319(7209):572-573.
21. Cochran WG, Diaconis P, Donner AP, D.C. H, O'Connor NE, Peterson OL, et al. Experiments in surgical treatments of duodenal ulcer. In: Bunker JP, Barnes BA, Mosteller F, editors. *Costs, risks and benefits of surgery*. Oxford: Oxford University Press.; 1977. p. 176-197.



22. Chalmers TC, Celano P, Sacks HS, Smith H, Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309(22):1358-1361.
23. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med.* 1989;8(4):441-454.
24. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med.* 1989;8(4):455-466.
25. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. *Am J Med.* 1982;72(2):233-240.
26. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess.* 1998;2(13):i-iv, 1-124.
27. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342(25):1887-1892.
28. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess.* 2003;7(27):iii-x, 1-173.
29. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA.* 2001;286(7):821-830.

30. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317(7167):1185-1190.
31. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess*. 2000;4(34):1-154.
32. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schunemann H, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev*. 2011(4):MR000012.
33. Oliver S, Bagnall AM, Thomas J, Shepherd J, Sowden A, White I, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess*. 2010;14(16):1-165, iii.
34. Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol Methods*. 2001;6(4):413-429.

## Tables

**Table 1. Baseline characteristics of survey respondents**

<b>Characteristic</b>	<b>Respondents n=123</b>
<b>Female</b>	55 (44.7)
<b>Age group</b>	
25 to 34	30 (24.2)
35 to 44	35 (28.2)
45 to 54	35 (28.2)
55 to 64	17 (13.7)
65 to 74	7 (5.6)
75 or older	0
<b>Systematic reviews conducted or participated in over the last 5 years*</b>	
1	9 (7.3)
2	6 (4.9)
3	9 (7.3)
4	9 (7.3)
5 or more	90 (73.2)
<b>Role in the systematic reviews conducted or participated in over the last 5 years *</b>	
Main author / coordinator	84 (68.3)
Co-author / research assistant / data screening / extraction	76 (61.8)
Methods and statistical advice	57 (46.3)
Search strategies / librarian	14 (11.4)
As an expert in the field, stakeholder, or consumer providing advice and approving final document	19 (15.4)
Other	4 (3.2)

Values represent the number and in parentheses the percentage.

\* Percentages do not add up to 100 because respondents could state more than one option

**Table 2. Other questions and results of the survey**

<b>Question</b>	<b>Response options</b>	<b>Respondents n (%)</b>
	My background knowledge and experience	89 (76.1)
What guidance do you use when determining evidence to include in a review? (check all that apply) *	The Cochrane Handbook	90 (76.9)
	My organization's guidance	42 (35.9)
	Other	13 (11.1)
Would you consider combining RS and NRS in a single meta-analysis in any of these scenarios from FIGURE 3? (you can choose more than one option) ¶§	SITUATION A	45 (44.1)
	SITUATION B	10 (9.8)
	SITUATION C	12 (11.8)
	SITUATION D	19 (18.6)
	I would not consider pooling under any circumstances	51 (50)

\* Out of 117 who responded

¶ Out of 102 who responded. Percentages do not add to 100 because respondents could choose more than one option.

§ Situation A: Randomized studies (RS) and non-randomized studies (NRS) are high certainty. Situation B:RS are high certainty and NRS are moderate, low, or very-low. Situation C: RS are moderate, low, or very-low certainty while NRS are high certainty. Situation D: both are moderate, low, or very-low certainty.

## Figures

Figure 1.

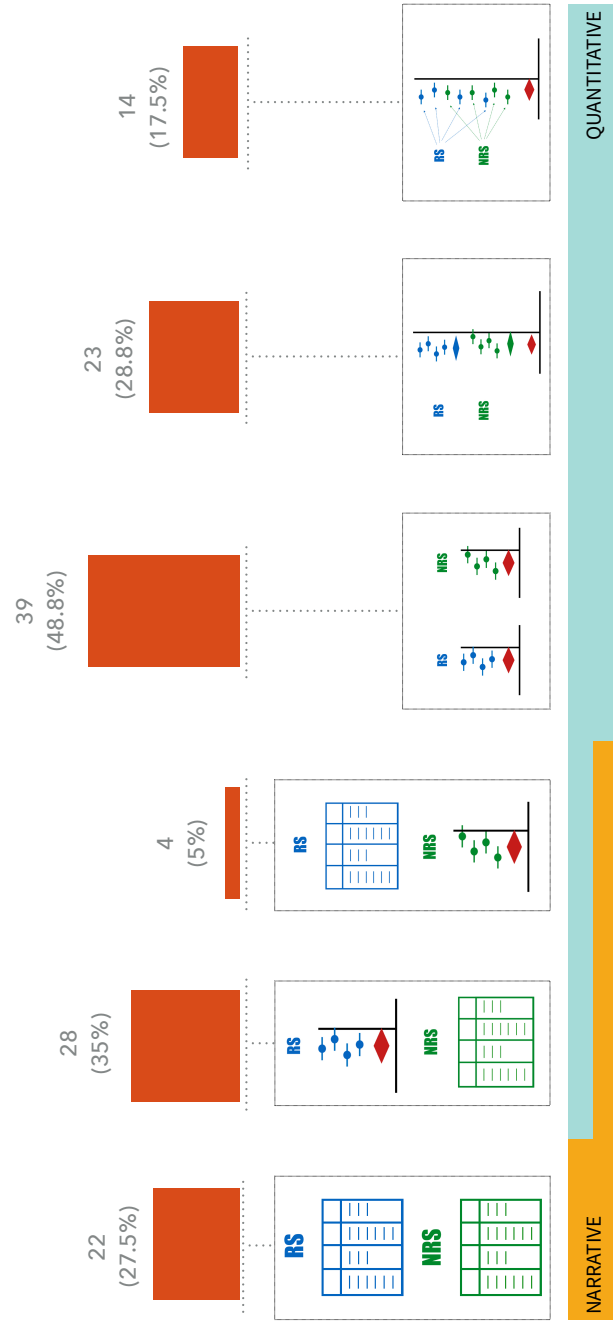


Figure 1. Most frequently used strategies by respondents to integrate randomized studies (RS) and non-randomized studies (NRS) in systematic reviews. Numbers represent the number (and percentages - %) of occasions in which experts have used the strategy (n=80). "Narrative", or "descriptive" synthesis is represented by the drawn 'tables', while quantitative refers when a meta-analysis is performed and is represented by the forest plots of either RS (blue colours) or NRS (red colours). Percentages may not add to 100 because participants could use more than one option.

Figure 2

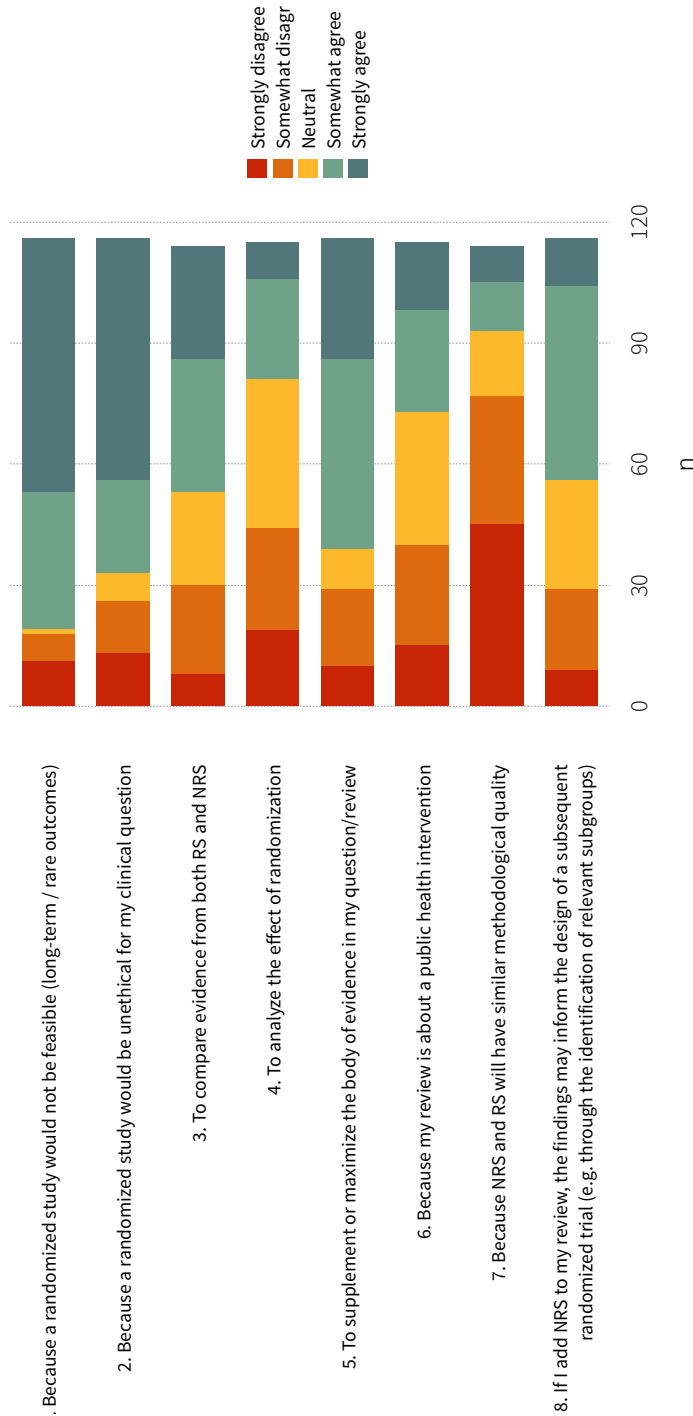


Figure 2. Participants' agreement on the reasons commonly cited for integrating NRS in systematic reviews. Numbers represent the number of participants voting for that option in the Likert scale. N=116 respondents.





## Appendices

## **Supplementary material 1: Search strategy for scoping review**

### ***THE COCHRANE LIBRARY***

#1 "observational":ti,ab

#2 "case-control":ti,ab

#3 "cohort":ti,ab

#4 "before-after":ti,ab

#5 non-random\*:ti,ab

#6 "cross-sectional":ti,ab

#7 non-experiment\*:ti,ab

#8 "interrupted time series":ti,ab

#9 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8

### ***MEDLINE***

#1 compara\*[tiab] OR comparison\*[tiab] OR contrast\*[tiab] OR similar\*[tiab]

OR consistent\*[tiab] OR inconsistent\*[tiab] OR dissimilar\*[tiab] OR

differen\*[tiab] OR concordan\*[tiab] OR discordan\*[tiab] OR heterogene\*[tiab]  
OR “Research Design”[mh]

#2 “Observation”[mh] OR “Cohort Studies”[mh] OR “Longitudinal  
Studies”[mh] OR “Retrospective Studies”[mh] OR “Prospective Studies”[mh] OR  
observational[tiab] OR cohort\*[tiab] OR cross-sectional[tiab] OR  
longitudinal[tiab] OR causal inference\*[tw] OR causality[tw] OR “instrumental  
variable”[tw] OR “structural model”[tw] OR practice-based[tw] OR propensity  
score\*[tw] OR natural experiment\*[tw] OR case-control[tw] OR before-after[tw]  
OR pre-post[tw] OR case-cohort[tw] OR case-crossover[tw] OR serial[tiab] OR  
non-experimental[tiab] OR “nonrandomized”[tiab] OR “nonrandomised”[tiab]  
OR “study designs”[tiab] OR “newcastle ottawa”[tiab] OR (evidence[tiab] AND  
quality[tiab])



#3 Cochrane Database Syst Rev [TA] OR search[tiab] OR meta-analysis[PT]  
OR MEDLINE[tiab] OR PubMed[tiab] OR (systematic\*[tiab] AND review\*[tiab])  
OR review[ti]

### ***THE JAMES LIND LIBRARY***

We used a broad search strategy starting with the terms “observational”, “reviews”,  
and “studies”; we then proceeded with an iterative approach of searching in the

references of included articles, in related essays, records, topics, and articles from the library.

## Supplementary material 2. Complete survey



### Randomised / Non-Randomised studies project

### INTRODUCTION

This survey is part of a larger study, supported in part by Cochrane (formerly the Cochrane Collaboration), and will aid understanding concepts and practices among persons who conduct systematic reviews. Specifically, we are hoping to learn more about the type of evidence that is considered during the review process.

It should take approximately 15 minutes to complete it.



For the purpose of this survey, the terms randomised study (RS) will be used to denote a randomised controlled trial (RCT). Non-randomized studies (NRS) will be the term used to denote observational studies, and will be defined as any research study that do not use randomization to allocate individuals or groups to interventions or a technology; these could include cohort, case-control, before-after studies, cross-sectional, interrupted time series, or a combination of designs.

In some of the questions we would like you to consider how you or your organization currently conduct reviews.

We will be using Cochrane and GRADE terminology such as Summary of Findings (SoF) tables, evidence profiles (EP), and **certainty** of the evidence (also known as *quality* of evidence or *confidence* in the evidence).

All information will be kept anonymous, no individual completing this survey will be identified or judged, and neither will the institution you work in. No personal identifiable data will be collected (unless you agree to contact us). All data will be analysed in aggregate form and cannot be traced back to individual participants.

For questions regarding this survey please contact us at [cuelloca@mcmaster.ca](mailto:cuelloca@mcmaster.ca)  
Thank you for participating!



### Randomised / Non-Randomised studies project

### Agreement to participate

**\* 1. This study has been reviewed by the Hamilton Health Sciences/ Faculty of Health Sciences Research Ethics Board at McMaster University, and it has waived the requirement for individual consent. As previously mentioned, all information is kept anonymous and confidential.**

- I agree to participate
- I do NOT wish to participate



## Randomised / Non-Randomised studies project

Participation in systematic reviews

**\* 2. To participate in this survey, you must have conducted or been engaged in at least one systematic review.**

- Yes, I have participated as an author or co-author of a systematic review
- No, I have never participated in a systematic review



## Randomised / Non-Randomised studies project

Background information

**3. What is your gender?**

- Female
- Male

2

**4. What is your age?**

- 18 to 24
- 25 to 34
- 35 to 44
- 45 to 54
- 55 to 64
- 65 to 74
- 75 or older

**5. On which of the following languages do you feel comfortable performing and completing a systematic review (or any evidence synthesis)? (Please select all that apply.)**

- Arabic
- Dutch
- English
- French
- German
- Italian
- Japanese
- Korean
- Mandarin
- Portuguese
- Polish
- Russian
- Spanish
- Other (please specify)



**Randomised / Non-Randomised studies project**

Preliminary

2



questions

**\* 6. How many systematic reviews have you been involved in during the last 5 years?**

- 1
- 2
- 3
- 4
- 5 or more

**\* 7. What was your role in the last systematic review on which you participated?  
(check all that apply)**

- Main author / coordinator
- Co-author / research assistant / data screening / extraction
- Methods and statistical advice
- Search strategies / librarian
- as an expert in the field, stakeholder, or consumer providing advice and approving final document

Other (please specify)

**\* 8. How familiar are you with Summary of Findings (SoF) tables?**

- Not familiar at all (I've never used them)
- Very little familiar (I've only used them once)
- A bit familiar (I've used them more than once, but I still need help when creating one)
- Somewhat familiar (I've used them on several occasions but still need help on some issues)
- Very familiar (I've used them on several occasions and/or I can help others create them)
- Expert (I am involved in GRADE methods and I can teach others how to create SoF tables)



**Randomised / Non-Randomised  
studies project**

**\* 9. Have you ever conducted a systematic review that included evidence from both randomised and non-randomised studies?**

- No
- Yes, once
- Yes, more than once

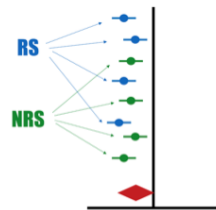
Mac  
**GRADE**  
Centre | McMaster University

Cochrane Methods  
GRADEing

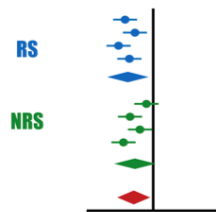
## Randomised / Non-Randomised studies project

**\* 10. If you have been involved in a systematic review that integrated non-randomised studies (NRS) with randomised studies (RS), how was the evidence analysed and presented? (check all that apply)**

- RS and NRS were pooled in a single meta-analysis

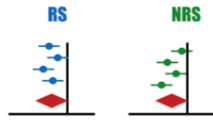


- The two types of studies were separated into two study sub-groups (NRS and RS) with a final overall effect pooled

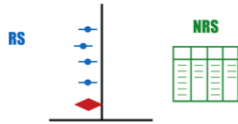


5

- RS and NRS were not pooled together, and evidence was presented in two separate meta-analyses



- RS were pooled, while NRS were not pooled in a final meta-analysis (only qualitative analysis)



- NRS were pooled, while RS were not pooled in a final meta-analysis (only qualitative analysis)



- Both NRS and NRS were not pooled, only qualitative analysis



use this space to comment on any of these options or if you used other methods that are not listed here



## Randomised / Non-Randomised studies project

Previous guidance

**\* 11. What guidance do you use when determining evidence to include in a review?  
(check all that apply)**

- My background knowledge and experience
- The Cochrane Handbook
- My organization's guidance

Other (please specify)



**Randomised / Non-Randomised  
studies project**

to combine or not to  
combine?

On these next 4 questions, think about reasons to include (combine) non-randomised studies (NRS) with randomised studies (RS). Remember that they could be combined at the systematic review level (with or without doing a meta-analysis) and at the meta-analysis level.

**REASONS TO INCLUDE NRS**

**12. In your opinion/experience, what are the reasons NRSshould be included in a SYSTEMATIC REVIEW of randomised studies? (whether you use meta-analysis or not)**

**13. In your opinion/experience, what are the reasons NRScould be integrated in a META-ANALYSIS together with the randomised studies?**

**REASONS NOT TO INCLUDE NRS**

**14. In your opinion/experience, what are the reason NRS *should not* be included in a SYSTEMATIC REVIEW of randomised studies? (whether you use meta-analysis or not)**

**15. In your opinion/experience, what are the reasons NRS *could not* be integrated in a META-ANALYSIS together with the randomised studies?**




## Randomised / Non-Randomised studies project

**16. The following are reasons commonly cited (see, for instance, [Ijaz 2014](#); [Schünemann 2013](#)) for including NRS (alone or together with RS) in a systematic review. Mark your agreement / disagreement with the reason:**

	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree
Because a randomised study would not be feasible for my clinical question (eg., long-term / rare outcomes, or the outcomes were not known to be important when existing major RCTs were conducted)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Because a randomised study would be unethical for my clinical question	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
To compare evidence from both RS and NRS	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

o

	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree
To analyze the effect of randomization	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
To supplement or maximize the body of evidence in my question/review	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Because my review is about a public health intervention	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Because NRS and RS will have similar methodological quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If I add NRS to my review, the findings may inform the design of a subsequent randomized trial (e.g. through the identification of relevant subgroups)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Comments	<input type="text"/>				




## Randomised / Non-Randomised studies project

**HYPOTHETICAL SCENARIOS**

Imagine you are conducting a systematic review, and you found evidence from both randomised studies (RS) and non-randomised studies (NRS) to answer your review question (and let's assume you're willing to include both if possible).

*First*, you assess the risk of bias using the [Cochrane RoB tool](#) for randomised studies.  
*Second*, you assess the [risk of bias of observational studies](#) using, for example, the ROBINS tool –formerly known as [ACROBAT](#), the Newcastle-Ottawa Scale ([NOS](#)), or other.  
*And third*, you use GRADE to evaluate the body of evidence of both RS and NRS (separately) to

o

create Summary of Findings (SoF) tables.

If you do this, you can encounter different scenarios depicted in the figure below. By using the figure examine the possible (hypothetical) scenarios that you might encounter and answer accordingly

**Possible scenarios**

**after GRADE  
assessment of  
the body of  
evidence**

**overall confidence from  
NON-RANDOMISED STUDIES  
is...**

**overall  
confidence  
from  
RANDOMISED  
STUDIES IS...**

	<p>HIGH ⊕⊕⊕⊕</p>	<p>MODERATE ⊕⊕⊕○ LOW ⊕⊕○○ VERY LOW ⊕○○○</p>
<p>HIGH ⊕⊕⊕⊕</p>	<p><b>A</b></p>	<p><b>B</b></p>
<p>MODERATE ⊕⊕⊕○ LOW ⊕⊕○○ VERY LOW ⊕○○○</p>	<p><b>C</b></p>	<p><b>D</b></p>

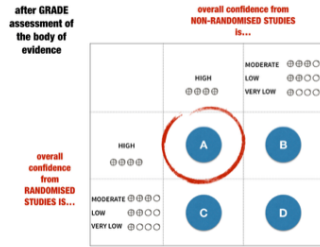
Click on NEXT to go to the scenarios




Randomised / Non-Randomised  
studies project

Scenario  
A

**17. SCENARIO A: after applying GRADE, the body of evidence from RS is high quality but you also find that the body of evidence from NRS is high quality. What would you like to do (regarding the combination of RS and NRs in a SoF table)?**



- If possible, combine all studies in a SoF table with no distinction between RS and NRS
- Put them in a single SoF table, but differentiating between RS and NRS
- I would use 2 different SoF tables, one for RS, other for NRS
- I would only use randomised studies
- I would only use non-randomised studies

Comments

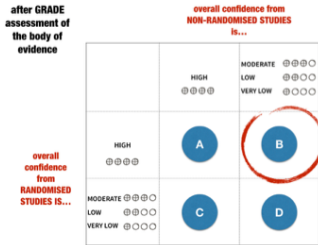
Randomised / Non-Randomised studies project

Scenario B



**18. SCENARIO B: after applying GRADE, the body of evidence from RS is high quality but you find that the body of evidence from NRS is either moderate, low, or very low quality.**

**What would you like to do (regarding the combination of RS and NRS in a SoF table)?**



- If possible, combine all studies in a SoF table with no distinction between RS and NRS
- Put them in a single SoF table, but differentiating between RS and NRS
- I would use 2 different SoF tables, one for RS, other for NRS
- I would only use randomised studies
- I would only use non-randomised studies

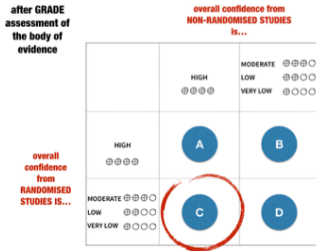
Comments

Randomised / Non-Randomised studies project

Scenario C

**19. SCENARIO C: after applying GRADE, the body of evidence from RS is either moderate, low, or very low quality but you find that the body of evidence from NRS is high quality.**

**What would you like to do (regarding the combination of RS and NRS in a SoF table)?**



- If possible, combine all studies in a SoF table with no distinction between RS and NRS
- Put them in a single SoF table, but differentiating between RS and NRS
- I would use 2 different SoF tables, one for RS, other for NRS
- I would only use randomised studies
- I would only use non-randomised studies

Comments




Randomised / Non-Randomised studies project

Scenario D

**20. SCENARIO D: after applying GRADE, the body of evidence from both RS and NRS is either moderate, low, or very low.**

**What would you like to do (regarding the combination of RS and NRS in a SoF table)?**

		overall confidence from NON-RANDOMISED STUDIES is...		
		HIGH ●●●●●	MODERATE ●●●○	LOW ●●○○
			VERY LOW ●○○○	
overall confidence from RANDOMISED STUDIES IS...	HIGH ●●●●●	A	B	
	MODERATE ●●●○	C	D	
	LOW ●●○○			
	VERY LOW ●○○○			

- If possible, combine all studies in a SoF table with no distinction between RS and NRS
- Put them in a single SoF table, but differentiating between RS and NRS
- I would use 2 different SoF tables, one for RS, other for NRS
- I would only use randomised studies
- I would only use non-randomised studies

Comments




Randomised / Non-Randomised studies project

**21. Would you consider combining RS and NRS in a single meta-analysis in any of these scenarios? (you can choose more than one option)**

		overall confidence from NON-RANDOMISED STUDIES is...		
		HIGH ●●●●●	MODERATE ●●●○	LOW ●●○○
overall confidence from RANDOMISED STUDIES IS...	HIGH ●●●●●	A	B	
	MODERATE ●●●○○	C	D	

- SITUATION A
- SITUATION B
- SITUATION C
- SITUATION D
- I would not consider combining under any circumstances

Comments




## Randomised / Non-Randomised studies project

**\* 22. Are you aware of any publication(s) or written instructions that describe the reason why not to combine RS and NRS in a single meta-analysis?**

- No
- Yes

If YES and possible, can you please specify the reference?



## Randomised / Non-Randomised studies project

**23. If you have any other comments or suggestions, we would like to hear them**



## Randomised / Non-Randomised studies project

The  
End

We thank and appreciate your support, your answers and opinions are very important to us. Please feel free to [contact us](#) if you have questions and/or would like to receive more information or participate in this project.

# CHAPTER 3. STRATEGIES TO OPTIMIZE USE OF RANDOMIZED AND NON- RANDOMIZED STUDIES IN EVIDENCE SYNTHESES OF INTERVENTIONS USING GRADE

## AUTHORS

Carlos A. Cuello<sup>a</sup>, Rebecca L. Morgan<sup>a</sup>, Jan Brozek<sup>a</sup>, Nancy Santesso<sup>a</sup>, Jos Verbeek<sup>b</sup>,  
Kris Thayer<sup>c</sup>, Mohammed T. Ansari<sup>d</sup>, Gordon Guyatt<sup>a</sup>, Holger J. Schünemann<sup>a</sup>

## AUTHORS AFFILIATIONS

- a. Department of Health Research Methods, Evidence, and Impact, McMaster University. Hamilton Ontario, Canada.
- b. Cochrane Work Review Group. Finnish Institute of Occupational Health, Helsinki, Finland
- c. National Center for Environmental Assessment. Environmental Protection Agency. USA
- d. School of Epidemiology and Public Health. University of Ottawa, Ottawa, Ontario. Canada.

## CORRESPONDING AUTHOR

Holger J. Schünemann, M.D., Ph.D.

Chair, Department of Health Research Methods, Evidence, and Impact.

McMaster University. Health Sciences Centre Room 2C16.

1280 Main Street West. Hamilton, Ontario Canada. L8N 4K1

Word count: 4,589

## **Abstract**

**OBJECTIVES:** Randomized studies (RS) provide the most trustworthy sources for the relative effect of health interventions summarized in knowledge syntheses. Non-randomized studies (NRS) can be used as replacement, sequential, or complementary evidence for a body of evidence of RS. In this paper, we present options for the optimal use of RS and NRS in health syntheses by using the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach in evidence profiles and summary of findings tables.

**METHODS:** We used three examples, as case studies, for the optimal use of RS and NRS in addressing health questions. We tested several options and developed solutions for dealing with methodological challenges based on feedback from experts in GRADE methods, Cochrane authors, and health guideline developers (including experts in the new Risk of Bias Tool for Non-Randomized Studies of interventions (ROBINS-I)).

**RESULTS:** Based on the three examples we address possible scenarios for use of RS and NRS in evidence syntheses. We provide descriptions for the solutions found that will be useful to users of GRADE and the ROBINS-I tool, and how ratings on the GRADE domains impact on the findings and the overall assessment of the evidence.

CONCLUSIONS: Considering their differences and similarities based on the GRADE domains, NRS and RS can complement one another in maximizing the value for improving knowledge syntheses and health recommendations.

**Keywords**

GRADE; Systematic Reviews; Clinical Guidelines; Evidence Synthesis; Randomized studies; Non-randomized studies; Research Methodology.



# 1. Introduction

Randomized studies (RS) provide the most trustworthy source of evidence for the relative effects of health care interventions summarized in knowledge syntheses (e.g., systematic reviews, health technology assessments – HTA –, and clinical practice guidelines). If high certainty evidence is not available from RS, non-randomized studies (NRS) of interventions, may provide replacement, sequential, or complementary evidence for of RS. <sup>1</sup>

Roughly half of the clinical questions about effectiveness of healthcare interventions cannot be answered with evidence from RS, <sup>2</sup> mainly due to feasibility issues (e.g., the studies that would answer the question would be unethical to conduct, or the outcome is a rare/long term event). Authors of health syntheses must decide from the outset (in a protocol) whether they aim to search and include NRS, and their presumptions about sources of the best evidence available to inform a recommendation. On many occasions clinicians, decision-makers, and clinical guideline developers will need evidence from NRS. Observational studies enrolling representative populations will, if available, provide the best evidence regarding baseline risk of outcomes of interest. In terms of relative effects, NRS may provide the best evidence when RS does not directly address the question of interest or is altogether unavailable; more than half of comparative effectiveness research (CER) studies includes NRS as source of evidence. <sup>3, 4</sup>

In the remainder of this presentation we will focus exclusively on estimates of relative effects and discuss issues in the context of the GRADE approach to rating certainty of evidence. Optimizing use of NRS with RS in knowledge syntheses in this context poses three main challenges: 1) How one should deal with differences between the two types of studies in one or more of the certainty of evidence domains (i.e., risk of bias, indirectness, imprecision, publication bias, and inconsistencies); 2) how to deal with the possibility of different direction and magnitude of effect estimates between RS and NRS; and 3) how to present, in GRADE evidence profiles (EP) and Summary of Findings (SoF) tables, the evidence from RS and NRS in transparent and understandable formats.

Although several methodological reviews and guidance papers acknowledge the importance of including RS and NRS in systematic reviews,<sup>3, 5-7</sup> there is currently no specific advice on when and how to do that in knowledge syntheses. In particular, it is not well explored how the GRADE approach, together with the new Cochrane tool for assessing the risk of bias in non-randomized studies of interventions (ROBINS-I), might support considering RS and NRS.<sup>8</sup>

In our prior work, we surveyed experts to understand and assess their preferences and attitudes towards the use of RS with NRS,<sup>9</sup> and reviewed how ROBINS-I will impact the GRADE assessment<sup>10</sup> of a body of evidence from NRS. In this study, we begin addressing each of the three challenges above through detailed examples. We

provide scenarios and methods for optimal use of RS and NRS in knowledge syntheses and health guidelines considering GRADE domains and evidence profiles.

## 2. Methods

We searched for examples from clinical questions in which the simultaneous use of RS and NRS of interventions in addressing relative effects in a systematic review was deemed worthy of consideration. For this, we used the results from a recent search strategy performed in a previous scoping review <sup>9</sup> and asked experts in the field of knowledge syntheses for illustrative cases. We chose three final examples guided by feedback from clinical epidemiologists, methodologists, and researchers with experience in systematic reviews, health guidelines, and in using tools for the assessment of risk of bias in NRS (e.g., ROBINS-I).

Through an iterative process of debate and revisions, we examined the appropriateness of options for dealing with RS and NRS in evidence syntheses and developed solutions for dealing with the challenges that arose, i.e., how to handle differences between RS and NRS in one or more of the certainty of evidence GRADE domains (risk of bias, indirectness, imprecision, publication bias, and inconsistencies); how to deal with the possibility of different direction and magnitude of effect estimates; and how to portray the two bodies of evidence from RS and NRS in transparent and understandable GRADE formats. We developed a draft guide,

based on a previous survey<sup>9</sup> and three GRADE and Cochrane expert meetings (Washington DC 2016, Seoul 2016, and Rome 2017) on how and when to integrate NRS with RS in systematic reviews and clinical guidelines. We included previous discussions from the NRS risk of bias GRADE group.<sup>10</sup>

We conducted discussions around the examples obtained, supported by previous methodological guidance from Cochrane,<sup>11</sup> the GRADE handbook,<sup>12</sup> and the ROBINS-I tool,<sup>8</sup> when applicable. We asked experts to discuss the way RS and NRS could be optimally used and presented in evidence profiles and summary of findings tables. The three examples are from health questions that raise relevant challenges and are presented in box 1, box 2, and box 3 respectively (complete case studies are also found in the appendices 1, 2, and 3 respectively). Based on these examples, we describe which judgements used by guideline panelists and systematic review authors would be required. First, we analyze each GRADE domain and the implication of their differences between RS and NRS; second, we analyze how the different direction and magnitude of effects impact on the decision on how to use RS or NRS; and third, we present several options on how to portray both bodies of evidence in summary of findings.

This project is supported by a Cochrane Methods Innovation grant, the National Toxicology Program within the National Institutes for Health (United States), and the McMaster GRADE centre.

**EXAMPLE 1 (see also appendix 1)**

A health guideline for the prevention and treatment of necrotizing enterocolitis (NEC) in the neonatal intensive care unit (NICU) is conducting a systematic review of the effects of supplementing probiotics to premature infants in the NICU. The outcomes assessed are 'NEC' grade II-III, 'overall mortality', and 'sepsis'. The review team found RS for the outcomes 'NEC' and 'mortality' that provide high certainty and decide not to look for NRS for these outcomes. However, for the outcome 'sepsis', the overall certainty from RS is deemed 'low' mostly due to inconsistency and imprecision. The panel decides they would feel more comfortable by looking for NRS, especially when case series and reports have linked the use of probiotics to sepsis in very preterm infants. Authors find seven NRS deemed low certainty due to risk of bias (confounding) but no other concerns (figure 1 and table 1).

**EXAMPLE 2 (see also appendix 2)**

A systematic review team is assessing the question for a health guideline on thromboprophylaxis: "Should antithrombin III (AT-III) versus no AT-III be used in critically ill infants undergoing extra-corporeal membrane oxygenation (ECMO) for the prevention of venous thrombosis?" Their search strategy yields four RS comparing AT-III to placebo, of which only two assess populations in the desired age group, i.e., children above one month of age. The certainty in the evidence from RS is very low due to risk of bias, imprecision, and indirectness. The review team decides to look at NRS; they found eight, of which only two directly assess the population of interest and are included with a certainty of evidence deemed low. (Table 2 and Figure 2)

**EXAMPLE 3 (see also appendix 3)**

In January 2016, a Health Technology Assessment unit is working on the question: “Should vitamin D be supplemented to all pregnant women for the prevention of recurrent wheeze or asthma in their infants?” The review team evaluates the body of evidence from RS first. The certainty is deemed very low due to risk of bias and very serious imprecision (only one RS is found –figure 3a and table 3). The team decides to include six NRS (figure 3b), providing low certainty for this outcome. In consultation with content experts, the authors judged that there is a dose-response effect reported in most studies (inverse relationship between the adjusted ORs and increased dosages or levels of vitamin D) which upgrade the certainty by one level, from low to moderate (table 3). In January 2017, an updated search yields two more RS and are added to the body of evidence (figure 4). Now, three RS provide an effect estimate excluding the null with moderate certainty in the evidence (only downgraded one level due to imprecision). Authors portray both study designs in a single table (table 4).

## **3. Differences in GRADE domains between randomized and non-randomized studies**

### **3.1. Risk of bias**

In the GRADE approach, limitations in the detailed design or conduct of the studies (bias) of both RS and NRS may lead to rating down the certainty of evidence. However, the initial certainty rating in RS starts as high, while that from NRS starts as low unless ROBINS-I is used. However, when ROBINS-I is used, raters are still required to have strong justification to not consider risk of bias due to confounding a very serious concern leading to a rating of low certainty<sup>8, 10</sup>. This initial downgrading in NRS by default of two levels is related to the fact that adequate randomization is the only secure method to protect against confounding and selection bias.<sup>13</sup> In consequence, biases arising from failure to randomize are the main reason authors separate RS from NRS in knowledge syntheses of interventions. However, random allocation does not protect against missing outcome data, measurement of outcomes, and selection of the reported results (figure 5).

As there are more than 200 instruments for assessing the risk of bias of individual studies,<sup>14</sup> GRADE does not require using a specific risk of bias tool for RS or NRS. Existing instruments for the assessment of risk of bias in NRS use as benchmark the “ideal” observational study of a specific design to assess the risk of bias of individual NRS (e.g., the ideal cohort or case-control study). ROBINS-I, rather than using the

ideal observational study, addresses risk of bias using an ideal (or target) randomized trial as a standard.<sup>8</sup>

In example 1 (probiotics for neonates, outcome: sepsis), there are no serious limitations in the design and conduct of the relevant RS (figure 1 and table 1) and thus risk of bias in GRADE is deemed as not serious. On the other hand, the body of evidence from NRS is also assessed (not using ROBINS-I) as not serious if we compare them to an ideal observational cohort study. However, as we have previously pointed out, the inherent risk of residual confounding leads to a starting evidence certainty of low (table 1). With the use of ROBINS-I, however, the same example would take a different route, but would end at the same point (see same example 1 on the appendix, where ROBINS-I is used). In this case, the body of evidence from NRS will not start as low by default, but as high certainty, as with RS, and then authors will judge the risk of bias due to confounding and selection of participants and rate down by two levels after judging the risk of bias from ROBINS-I as very serious. The resulting overall certainty will still be low.

An advantage of this process is tackling concerns and confusion about the issue of double counting the risk of confounding and selection bias (once for absence of randomization and then for additional concerns about confounding and selection bias as part of the evaluation of the observational studies). The labelling of observational studies as being of low risk of bias in evidence profiles has led to



confusion by users of this information (when in reality, it already included lack of randomization as a risk of bias consideration). This repeated consideration of risk of bias often lead to rating down further from low to very low. By using ROBINS-I, authors begin NRS as high certainty but rate down two levels based on the ROBINS-I items related to lack of randomization and concealment process if no other measures against selection or confounding bias are taken (for example, cohort studies with interrupted time series design will have less concern about confounding and the risk of bias domain in GRADE could be considered only serious, or not serious –instead of very serious).

### **3.2. Inconsistency between a body of evidence from RS and NRS**

Inconsistency drives considerations leading to interpreting bodies of evidence from RS and NRS separately, relying on one of the two, or pooling them both. If the body of evidence from RS and NRS indicates inconsistent results, RS and NRS must be considered separately. If one body of evidence is rated as higher certainty, we will rely on that body of evidence. Consideration of individual GRADE domains, particularly, inconsistency, indirectness and imprecision, may bear on the judgment of whether to include RS and NRS together to generate a single pooled estimate.

When raters evaluate inconsistency, they explore *a priori* hypotheses about differences in the populations, interventions, outcomes, or study methods that may

explain the observed heterogeneity. If they fail to find a compelling explanation for serious inconsistency, GRADE suggests rating down the certainty of evidence.<sup>15</sup> Here we will focus on inconsistency due to different study methods.

Once RS and NRS are assessed individually on the inconsistency domain, authors should assess the extent of differences in effect estimates. Some scenarios from figure 6 will give authors more confidence about pooling RS and NRS, for example, in scenario D1, where both RS and NRS have no concerns with inconsistency and the effect estimates are in the same direction, RS and NRS would be considered appropriate to integrate in a single pooled estimate. On the other hand, scenario D2, although each body of evidence has no concerns with inconsistency, they yield important differences in effect – indeed one suggesting benefit, the other harm.

If the bodies of evidence from RS and NRS each show internal inconsistency (figure 6, scenario A) authors should explore in detail according to their *a priori* hypotheses. If, for example, the explanation for inconsistency in both RS and NRS lies in the population (e.g., disease severity or risk), this could be detected in a subgroup analyses, as shown in scenario A2, making RS and NRS now consistent in their results. Authors should feel more comfortable about pooling because the inconsistency is now explained by factors other than the study design *per se*.

Our example 1 (probiotics for neonates, outcome: sepsis; table 1; figure 1) represents a situation where there is inconsistency in the individual RS (95% C.I. from some studies do not overlap, and  $I^2$  and p values suggest heterogeneity; figure 1), but the individual NRS yield similar results to one another (i.e., there is no inconsistency), this is comparable to scenario B from figure 6. In this case, NRS can replace the evidence of RS.

Example 2 (antithrombin III for infants undergoing extra-corporeal membrane oxygenation; table 2), represents a circumstance similar to scenario D2 in figure 6, where RS and NRS are both internally consistent ( $I^2$  of 0%), yet there is an obvious indirectness and imprecision in RS and a difference in effect estimates between RS and NRS (i.e., RS provide a RR 0.71 [95% C.I. 0.36 to 1.39] while NRS provide a OR of 1.54 [95% C.I. 1.35 to 1.76]; figure 2), which would make the pooling of the results inappropriate. In this case, the decision to use one body of evidence over another is not influenced by the inconsistency domain but by other domains such as indirectness and imprecision.

Example 3 (vitamin D supplementation to pregnant mothers to prevent asthma or wheezing in their infants; figure 4 and table 4) represents a circumstance like scenario D1 from figure 6, in which both RS and NRS present no inconsistency, and point estimates results are similar. In this case, inconsistency is not an issue at any level. Authors must decide to portray both bodies of evidence in a single summary of

findings table and decide whether to pool the final effect estimate or present them separated in two rows.

### **3.3. Indirectness**

Indirectness results if the research evidence utilized to answer a question does not directly relate to the population, intervention, comparisons or outcomes of interest.<sup>16, 17</sup> It is not uncommon for knowledge synthesis developers to use only indirect evidence to address a research question when there is a paucity of direct evidence. It is possible for direct evidence from NRS to provide equivalent or higher certainty than indirect evidence from RS.<sup>1</sup>

There may be situations when choosing between utilizing RS, NRS, or both, will be decided mostly based on indirectness. Example 2 (antithrombin III for critically ill infants undergoing ECMO for the prevention of thrombosis; figure 2 and table 2) represents an illustration of a body of evidence from RS that provide evidence rated as very low certainty (due to indirectness, imprecision, and risk of bias), while NRS provide direct evidence with low certainty due to risk of bias (confounding). In this example, it is sensible to use only the body of evidence from NRS because they provide higher certainty (low versus very low), and especially direct evidence for the clinical recommendation.

### **3.4. Imprecision**

Imprecision is determined by the examination of the 95% C.I. with the help of the sample size calculations and consideration of thresholds of appreciable benefit and harms.<sup>18, 19</sup> When imprecision is the only affected domain (i.e., with an equal risk of bias between RS and NRS and other GRADE domains unaffected), it is feasible and appropriate to integrate both types of studies in a single row of a GRADE table and in a single pooled estimate.

Imprecision may influence our decision to use one body of evidence over another. In example 1 (probiotic supplementation for preventing necrotizing enterocolitis in neonates; figure 1 and table 1), authors are concerned about a potentially harmful outcome (sepsis) which may be caused (or even prevented) by probiotics based on previous case reports and case series of probiotics administered to neonates (see also appendix 1). In this case, if authors only look at the body of evidence from RS, downgraded to low certainty due to inconsistency and imprecision, the result might inform their recommendation (to provide or not probiotics) but it will not be considered precise enough to reach a plausible threshold of benefit or harm (the 95% C.I. of the absolute effect goes from 37 fewer to 18 more cases of sepsis per 1,000 treated; table 1). In the body of evidence of NRS, due to residual confounding concerns, GRADE rates down the certainty in the body of evidence to low (with no other concerns in the other GRADE domains), but it provides an effect estimate that makes the decision-maker more confident in recommending probiotics because it

excludes a plausible threshold of harm (absolute effect 95% C.I. now goes from 48 fewer to 0 cases of sepsis per 1,000 treated). Authors of the systematic review would have to choose between utilizing only NRS, or pooling these with the body of evidence of RS.

### **3.5. Publication bias**

Strong suspicion of publication bias leads to rating down a body of evidence<sup>20</sup>. Both RS and NRS are prone to this type of bias,<sup>21, 22</sup> and authors should be attentive to its presence, especially when evidence comes from a small number of studies and/or there is large commercial interest.

Although NRS added to a body of evidence of RS can increase the number of (large) studies and possibly improve the assessment of publication bias by using the common techniques of analysis of patterns of data (i.e., funnel plots), whether a body of evidence with RS and suspicion of publication bias can be improved by adding NRS without suspected publication bias (or vice versa) is still unknown, yet possible, and more empirical evidence is required.

### **3.6. Large effects, dose-response, and opposing residual confounding**

Large effects, dose-response, and opposing residual confounding are the three GRADE domains that can rate up a body of evidence of NRS.<sup>23</sup> Any of these

domains can improve the certainty of evidence from NRS, hence increasing the likelihood of replacing or complementing RS. Because a large effect or dose-response associations can still be accompanied by residual confounding, GRADE recommends authors should proceed with caution and transparency when rating up a body of evidence. Example 3 (tables 3 and 4) represents a possible illustration of a dose-response gradient that would warrant rating up the certainty from NRS from low to moderate, hence facilitating the utilization of NRS as complement or replacement of RS in the systematic review, this is, using only NRS.

Opposing residual confounding applies when unmeasured plausible residual confounding bias would act to reduce the demonstrated effect, or increase the effect if no effect was observed.<sup>23</sup> If opposing plausible residual confounding is suspected, authors can rate up one level the certainty of evidence in NRS and apply other GRADE criteria to evaluate the appropriateness of using NRS as replacement or complement of RS. One important distinction is that the evaluation of opposing residual confounding is optionally included in the new tool for assessing risk of bias in NRS (ROBINS-I),<sup>8</sup> integrating this GRADE domain on each ROBINS-I item as an add-on for signalling questions. Therefore, when using ROBINS-I, authors may evaluate opposing residual confounding in the risk of bias GRADE domain, and not as a stand-alone domain; more testing and empirical observations are, however, needed.<sup>10</sup>

## **4. Differences in direction and magnitude of effects between randomized and non-randomized studies**

The direction, magnitude, and precision of the estimated effects in RS and NRS are intertwined concepts that will have an influence on the decision to include NRS with RS either in a single review, a single summary of findings table, or as a single pooled estimate. Different directions of effects between RS and NRS will make their pooling inappropriate, especially if the results are precise and clinically important. For instance, in example 2 (on the use of antithrombin III; table 2 and figure 2), the estimated effect from NRS reaches a precise result in opposite direction from RS – the effects in the RS are neither large nor precise. In this situation, the body of evidence of NRS can be used as replacement of RS, which are now unnecessary given the very low certainty due to indirectness, imprecision, and risk of bias.

A similar direction in both RS and NRS, on the other hand, provides confidence to the decision-maker that both bodies of evidence are similar to be presented on a single summary of findings (with or without pooling the effect estimates). On example 3 (figure 4), an updated review of the evidence provides RS and NRS with similar effects in terms of direction, with NRS providing precise results. In this case, authors of the systematic review can use NRS as complement for RS and present both in a single summary of findings table (or even into a single pooled estimate).



## **5. Presenting randomized and non-randomized studies in GRADE tables**

If the above cases and discussions lead to inclusion of RS and NRS in a knowledge synthesis, authors must portray both bodies of evidence in GRADE tables (evidence profiles and summary of findings tables). By using RS and NRS in summary of findings tables, authors will have three options to display both. First, they can portray the two bodies of evidence in a single summary of findings and keep RS and NRS separated in two rows. From example 1, we present both bodies of evidence in a single evidence profile (table 1) and how a summary of findings table would look (table 5), with separation between RS and NRS in two different rows with the same outcome. Second, they can use a single row with the two bodies of evidence combined into a single pooled estimate. Third, authors could use two separate tables, one for each body of evidence. Based on a previous survey<sup>9</sup> we found that experts often prefer the first option of keeping RS and NRS separated in two rows in a single summary of findings, followed by including RS and NRS pooled in a single row, and then the option of two summary of findings.

## **6. Discussion**

We reviewed the main challenges of integrating RS and NRS and its appropriateness using three examples representing a subset of situations that depict how NRS could

help decision-makers and systematic reviewers gain insight into bodies of evidence for clinical questions about interventions. The integration of both types of studies requires judgement and methodological expertise, with adequate transparency, considering the whole picture of both bodies evidence in relation to the certainty of evidence on each outcome.

During our interviews and discussions with methodologists and experts, we have perceived a series of concerns and questions that are worth examining. First, there will be situations where choosing RS over NRS will be straightforward, e.g., if RS are classified as high certainty of evidence. With these levels of certainty in RS, there should be no reason to look for NRS.<sup>7, 9</sup> The lower the certainty of evidence from RS, the more authors should be encouraged to look for NRS. In some instances, NRS will be the sole existing evidence, and could be included until an RS is available. It is important to be open about the possibility that NRS with low or moderate certainty of evidence can complement or replace RS classified as very low or low respectively, as long as the process keeps its transparency, a condition that GRADE provides.<sup>1</sup>

Second, a comprehensive protocol for a systematic review is vital to also report the intention of the authors to include NRS if they expect to find low or very low certainty of evidence from RS (or not find them at all). Authors might question whether they should portray both bodies of evidence that they have found, even if

one of them is very low certainty of evidence. Another concern is that selecting one design over the other could open the door for misuse, for instance, by using only NRS for the approval/use of new drugs. Selection of evidence and outcomes is an issue that authors must confront, although it is not exclusive to the integration of RS and NRS, but to the process of conducting health syntheses.<sup>24, 25</sup> In these situations, authors should present the body of evidence that provides the more trustworthy information; yet again, the importance of the protocol and transparency is emphasized.

Third, authors should pay attention to the differences between RS and NRS in the GRADE domains. Among these, risk of bias, imprecision, inconsistency, and indirectness represent challenges, especially when they differ between RS and NRS that have similar certainty of evidence and authors must decide which body of evidence should be used. When such differences are found, pooling RS and NRS into a single effect estimate will be inappropriate and authors should use the body of evidence that provides the highest certainty for decision-making. Also in some cases, one GRADE domain can influence the decision to include or exclude one body of evidence; in example one, for instance, indirectness in RS influences the decision to exclude them from the consideration in making a clinical recommendation.<sup>1, 12, 16</sup>

Fourth, although more observations and examples are needed, ROBINS-I will have an important influence when more syntheses start utilizing this tool. For instance,

the GRADE domain plausible residual confounding, optionally included in the ROBINS-I tool, may be rated in an integrated manner with the GRADE risk of bias domain and be accounted for during the GRADE risk of bias rating.<sup>8, 10</sup>

## **7. Conclusions**

The GRADE approach is a sensible and transparent way of helping authors of health syntheses assessing the appropriateness and providing methods for using RS and NRS to improve the certainty of a body of evidence by including the best available information. Next steps of this endeavour aim at creating detailed guidance for authors of knowledge syntheses to reach a decision on when and how to include both study designs.

## **Acknowledgements**

We are grateful to McMaster University fellows, graduate students, and all professionals and experts in GRADE who provided valuable feedback during the meetings where we discussed this topic.

## **Funding sources**

This work was supported by a Cochrane Methods Innovation grant, the National Toxicology Program within the National Institutes for Health, and the McMaster GRADE centre. The sponsors had no role in the design of the study or interpretation of the results except for through the lead authors of this article.

## References

1. 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. *Res Synth Methods*. 2013;4(1):49-62.
2. Various. What conclusions has Clinical Evidence drawn about what works, what doesn't based on randomised controlled trial evidence? <http://clinicalevidence.bmj.com/x/set/static/cms/efficacy-categorisations.html>. Published 2017. Accessed May 1, 2017.
3. Faber T, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Meta-analyses including non-randomized studies of therapeutic interventions: a methodological review. *BMC Med Res Methodol*. 2016;16:35.
4. Holve E, Pittman P. A first look at the volume and cost of comparative effectiveness research in the United States. [https://www.academyhealth.org/files/FileDownloads/AH\\_Monograph\\_09FINAL7.pdf](https://www.academyhealth.org/files/FileDownloads/AH_Monograph_09FINAL7.pdf). Published 2009. Accessed May, 2017.
5. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014;4:MR000034.
6. Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. *J Clin Epidemiol*. 2014;67(6):645-653.

7. Norris S, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Selecting Observational Studies for Comparing Medical Interventions. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville (MD)2008.
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. Cuello-Garcia C, Schunemann H, Morgan R, Santesso N, Thayer K, Verbeek JH, et al. Integration of randomized and non-randomized studies in systematic reviews of interventions: rationale, perceptions, and preferences [in preparation]. 2017.
10. Schunemann H, Akl EA, Morgan R, Cuello-Garcia C. GRADE Guidelines 19. How new tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence [in publication]. 2017.
11. Higgins JPT, Green S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. The Cochrane Collaboration, 2011. Available from <http://www.cochrane-handbook.org.2011/>.
12. Schunemann H, Brozek J, Guyatt GH, Oxman AD. GRADE handbook for grading quality of evidence and strength of recommendations. In: Schunemann H, Brozek J, Guyatt GH, Oxman AD, editors.: The GRADE working group; 2013.
13. 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). *J Clin Epidemiol*. 2011;64(4):407-415.

14. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovich C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess.* 2003;7(27):iii-x, 1-173.
15. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol.* 2011;64(12):1294-1302.
16. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol.* 2011;64(12):1303-1310.
17. Schunemann HJ. Methodological idiosyncracies, frameworks and challenges of non-pharmaceutical and non-technical treatment interventions. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen.* 2013;107(3):214-220.
18. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol.* 2011;64(12):1283-1293.
19. Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *J Clin Epidemiol.* 2016;75:6-15.
20. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol.* 2011;64(12):1277-1282.



21. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA*. 1992;267(3):374-378.
22. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010;14(8):iii, ix-xi, 1-193.
23. 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):1311-1316.
24. Norris SL, Holmer HK, Ogden LA, Fu R, Abou-Setta AM, Viswanathan MS, et al. *Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness*. Rockville (MD)2012.
25. Norris SL, Moher D, Reeves BC, Shea B, Loke Y, Garner S, et al. Issues relating to selective reporting when including non-randomized studies in systematic reviews on the effects of healthcare interventions. *Res Synth Methods*. 2012;4(1):36-47.

## Tables

Table 1.

Table 1. Probiotic supplementation in preterm infants in the neonatal intensive care unit. Outcomes: All-cause mortality, and culture proven sepsis.

N° of studies	Quality assessment							N° of patients		Effect		Quality	Importance
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Probiotics	no probiotics	Relative (95% CI)	Absolute (95% CI)			
All cause neonatal mortality – Randomized studies													
17	randomised trials	not serious <sup>a</sup>	not serious	not serious	not serious <sup>b</sup>	none	118/2635 (4.5%)	181/2668 (6.8%)	RR 0.70 (0.55 to 0.88)	20 fewer per 1,000 (from 8 fewer to 31 fewer)	⊕⊕⊕⊕ HIGH	CRITICAL	
All cause neonatal mortality – Non-randomized studies													
11	observational studies	not serious <sup>c</sup>	not serious	not serious	not serious	none	358/5126 (7.0%)	372/5642 (6.6%)	RR 0.72 (0.61 to 0.86)	18 fewer per 1,000 (from 9 fewer to 26 fewer)	⊕⊕○○ LOW	CRITICAL	
Sepsis – Randomized studies													
19	randomised trials	not serious	serious <sup>f</sup>	not serious	serious <sup>g</sup>	none	391/2662 (14.7%)	434/2676 (16.2%)	RR 0.92 (0.77 to 1.11)	13 fewer per 1,000 (from 18 more to 37 fewer)	⊕⊕○○ LOW	CRITICAL	
Sepsis – Non-randomized studies													
7	observational studies	not serious	not serious	not serious	not serious	none	570/3979 (14.3%)	538/2914 (18.5%)	RR 0.86 (0.74 to 1.00)	26 fewer per 1,000 (from 0 fewer to 48 fewer)	⊕⊕○○ LOW	CRITICAL	

CI: Confidence interval; RR: Risk ratio

**Explanations**

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process
- b. Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise
- c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.
- d. Includes both randomized and non-randomized studies
- e. The risk of bias in NRS stems from the residual confounding which is almost invariably impossible to discard.
- f. Statistical heterogeneity of 47% on the I square value.
- g. Confidence intervals still include the plausible harm and benefit thresholds

**References for table 1.**

[1] AlFaleh K, Anabrees J. Probiotics for prevention of necrotizing enterocolitis in preterm infants. *Cochrane Database Syst Rev.* 2014;CD005496.

[2] Olsen R, Greisen G, Schroder M, Brok J. Prophylactic Probiotics for Preterm Infants: A Systematic Review and Meta-Analysis of Observational Studies. *Neonatology.* 2016;109:105-12.

Table 2

Table 2. Evidence profile. Randomized and non-randomized studies. Antithrombin III for the prevention of thrombosis in children undergoing ECMO. Outcome: Any Thrombosis.

Nº of studies	Study design	Quality assessment						Nº of patients		Effect		Quality	Importance
		Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	ATIII	Control	Relative (95% CI)	Absolute (95% CI)			
<b>Thrombosis – Randomized Studies [1, 2]</b>													
2	randomized trials	serious <sup>a</sup>	not serious	serious <sup>b</sup>	serious <sup>c</sup>	none	8/33 (24.2%)	25/69 (36.2%)	<b>RR 0.71</b> (0.36 to 1.39)	<b>105 fewer per 1,000</b> (from 141 more to 232 fewer)	⊕○○○ VERY LOW	CRITICAL	
<b>Thrombosis – Non-randomized Studies [3, 4]</b>													
2	observational studies	not serious <sup>d</sup>	not serious	not serious	not serious	none	464/1961 (23.7%)	1048/6704 (15.6%)	<b>OR 1.54</b> (1.35 to 1.76)	<b>66 more per 1,000</b> (from 44 more to 90 more)	⊕⊕○○ LOW	CRITICAL	

a. Both studies not blinded. There are concerns in one study about the random sequence generation and incomplete outcome data.  
 b. One study assesses children undergoing cardiac surgery while other includes children with leukemia and asparaginase treatment.  
 c. Wide confidence intervals that do not exclude a plausible harm or benefit. Also, very low number of events and participants.  
 d. Residual confounding was not possible to exclude even after authors perform a good assessment and adjustment of confounders (in this example ROBINS-i was not used).

95%CI, 95% confidence interval  
 ATIII: Antithrombin III

## References for table 2

- [1] McCrindle BW, Manliot C, Holtby HM, Chan AK, Brandao LR, Rolland M, et al. Abstract 18061: Supplementation to Treat Antithrombin Deficiency Improves Sensitivity to Heparin, Anticoagulation and Decreased Thrombogenicity in Neonates and Infants Undergoing Cardiac Surgery With Cardiopulmonary Bypass. *Circulation*. 2015;132:A18061-A.
- [2] Mitchell L, Andrew M, Hanna K, Abshire T, Halton J, Wu J, et al. Trend to efficacy and safety using antithrombin concentrate in prevention of thrombosis in children receiving l-asparaginase for acute lymphoblastic leukemia. Results of the PAARKA study. *Thromb Haemost*. 2003;90:235-44.
- [3] Wong TE, Delaney M, Gernsheimer T, Matthews DC, Brogan TV, Mazor R, et al. Antithrombin concentrates use in children on extracorporeal membrane oxygenation: a retrospective cohort study. *Pediatr Crit Care Med*. 2015;16:264-9.
- [4] Wong TE, Nguyen T, Shah SS, Brogan TV, Witmer CM. Antithrombin Concentrate Use in Pediatric Extracorporeal Membrane Oxygenation: A Multicenter Cohort Study. *Pediatr Crit Care Med*. 2016;17:1170-8.

Table 3

Table 3. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Outcome: Asthma/recurrent wheezing.

Nº of studies	Quality assessment							Effect		Quality	Importance	
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Nº of patients		Relative (95% CI)			Absolute (95% CI)
<b>Asthma / recurrent wheezing – Randomized studies [1]</b>												
1	randomised trials	serious <sup>a</sup>	not serious	not serious	very serious <sup>b</sup>	none	17/108 (15.7%)	7/50 (14.0%)	RR 1.12 (0.50 to 2.54)	17 more per 1,000 (from 70 fewer to 216 more)	⊕○○○ VERY LOW	CRITICAL
<b>Asthma / recurrent wheezing – Non-randomized studies [2-7]</b>												
6	non-randomized studies	not serious <sup>c</sup>	not serious	not serious	not serious	dose response gradient <sup>d</sup>	total 8,831 <sup>e</sup>	26,553 (risk 14%) <sup>e</sup>	OR 0.76 (0.69 to 0.84)	30 fewer per 1,000 (from 20 fewer to 39 fewer)	⊕⊕⊕○ MODERATE	CRITICAL

- a. There were 22/180 participants who were not analyzed (lost to follow-up), 16% in the intervention group and 10% in the control group. Also, the outcome was a subjective measure and participants were not blinded to treatment allocation (reporter bias).
- b. Wide confidence interval with low number of participants for the optimal information size; also, crossing the null and the appreciable thresholds for benefit and harm.
- c. All studies have possible residual confounding. The non-randomized studies thus are downgraded two levels as in regular GRADE assessment. No further downgrading was considered necessary.
- d. All studies demonstrated a significant dose-response association at different levels of vitamin D supplementation on the risk of asthma or wheezing.
- e. No data on number of events on each arm; instead, all studies provide the adjusted odds ratios on the risk of asthma and its association with vitamin D intake. Baseline risk in control group was imputed from the rest of the studies, including the randomized controlled trial.

95%CI, 95% confidence interval

### References for table 3

- [1] 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:e66627.
- [2] 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. *J Dev Orig Health Dis*. 2015;6:308-16.
- [3] 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. *Am J Clin Nutr*. 2007;85:788-95.
- [4] Devereux G, Litonjua AA, Turner SW, Craig LC, McNeill G, Martindale S, et al. Maternal vitamin D intake during pregnancy and early childhood wheezing. *Am J Clin Nutr*. 2007;85:853-9.
- [5] 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. *Clin Exp Allergy*. 2009;39:875-82.
- [6] Maslova E, Hansen S, Jensen CB, Thorne-Lyman AL, Strom M, Olsen SF. Vitamin D intake in mid-pregnancy and child allergic disease - a prospective study in 44,825 Danish mother-child pairs. *BMC Pregnancy Childbirth*. 2013;13:199.
- [7] 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;35:1228-34.



Table 4

Table 4. Evidence profile (UPDATED July 2017). Randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Outcome: Asthma/recurrent wheezing.

N° of studies	Quality assessment							Effect		Quality	Importance	
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Vitamin D	Control	Relative (95% CI)			Absolute (95% CI)
<b>Asthma / recurrent wheezing – Randomized studies [1-3]</b>												
3	randomised trials	not serious <sup>a</sup>	not serious	not serious	serious <sup>b</sup>	none	162/805 (20.1%)	184/736 (25%)	RR 0.82 (0.68 to 0.99)	45 fewer per 1,000 (from 3 fewer to 80 fewer)	⊕⊕⊕○ MODERATE	CRITICAL
<b>Asthma / recurrent wheezing – Non-randomized studies [4-9]</b>												
6	non-randomized studies	not serious <sup>c</sup>	not serious	not serious	not serious	dose response gradient <sup>d</sup>	total 8,831 <sup>e</sup>	26,553 (risk 14%) <sup>e</sup>	OR 0.76 (0.69 to 0.84)	30 fewer per 1,000 (from 20 fewer to 39 fewer)	⊕⊕⊕○ MODERATE	CRITICAL

- a. In one study, the number of patients who were lost to follow-up were balanced (16% in the intervention group and 10% in the control group) and was not considered significant to change results.
- b. Confidence interval with low number of participants for the optimal information size; although not crossing the null and perhaps only the threshold for appreciable benefit.
- c. All studies have possible residual confounding. The non-randomized studies thus are downgraded two levels as in regular GRADE assessment. No further downgrading was considered necessary.
- d. All studies demonstrated a significant dose-response association at different levels of vitamin D supplementation on the risk of asthma or wheezing.
- e. No data on number of events on each arm; instead, all studies provide the adjusted odds ratios on the risk of asthma and its association with vitamin D intake. Baseline risk in control group was imputed from the rest of the studies, including the randomized controlled trial.

95%CI, 95% confidence interval

## References for table 4

- [1] 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:e66627.
- [2] Chawes BL, Bonnelykke K, Stokholm J, Vissing NH, Bjarnadottir E, Schoos AM, et al. Effect of Vitamin D3 Supplementation During Pregnancy on Risk of Persistent Wheeze in the Offspring: A Randomized Clinical Trial. *JAMA*. 2016;315:353-61.
- [3] 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;315:362-70.
- [4] 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. *J Dev Orig Health Dis*. 2015;6:308-16.
- [5] 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. *Am J Clin Nutr*. 2007;85:788-95.
- [6] Devereux G, Litonjua AA, Turner SW, Craig LC, McNeill G, Martindale S, et al. Maternal vitamin D intake during pregnancy and early childhood wheezing. *Am J Clin Nutr*. 2007;85:853-9.
- [7] 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. *Clin Exp Allergy*. 2009;39:875-82.

[8] Maslova E, Hansen S, Jensen CB, Thorne-Lyman AL, Strom M, Olsen SF. Vitamin D intake in mid-pregnancy and child allergic disease - a prospective study in 44,825 Danish mother-child pairs. *BMC Pregnancy Childbirth*. 2013;13:199.

[9] 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;35:1228-34.

Table 5

**Table 5. Example of Summary of Findings table with two bodies of evidence (RS and NRS) in two outcomes. Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks.**

Outcomes	N <sup>o</sup> of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects
				Risk with no probiotics
				Risk difference with probiotics
All cause neonatal mortality – RS [1]	5303 (17 RS)	⊕⊕⊕⊕ HIGH <sup>ab</sup>	<b>RR 0.70</b> (0.55 to 0.88)	68 per 1,000 <b>20 fewer per 1,000</b> (31 fewer to 8 fewer)
All cause neonatal mortality –NRS [2]	10768 (11 NRS)	⊕⊕○○ LOW <sup>c</sup>	<b>RR 0.72</b> (0.61 to 0.86)	66 per 1,000 <b>18 fewer per 1,000</b> (26 fewer to 9 fewer)
Sepsis – RS [1]	5338 (19 RS)	⊕⊕○○ LOW <sup>fg</sup>	<b>RR 0.92</b> (0.77 to 1.11)	162 per 1,000 <b>13 fewer per 1,000</b> (37 fewer to 18 more)
Sepsis – NRS [2]	6893 (7 NRS)	⊕⊕○○ LOW	<b>RR 0.86</b> (0.74 to 1.00)	185 per 1,000 <b>26 fewer per 1,000</b> (48 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; RR: Risk ratio; RS: Randomized studies; NRS: Non-randomized studies.

**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

**Explanations**

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process
- b. Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise
- c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.
- d. Includes both randomized and non-randomized studies
- e. The risk of bias in NRS stems from the residual confounding which is almost invariably impossible to discard.
- f. Statistical heterogeneity of 4.7% on the I square value.
- g. Confidence intervals still include the plausible harm and benefit thresholds

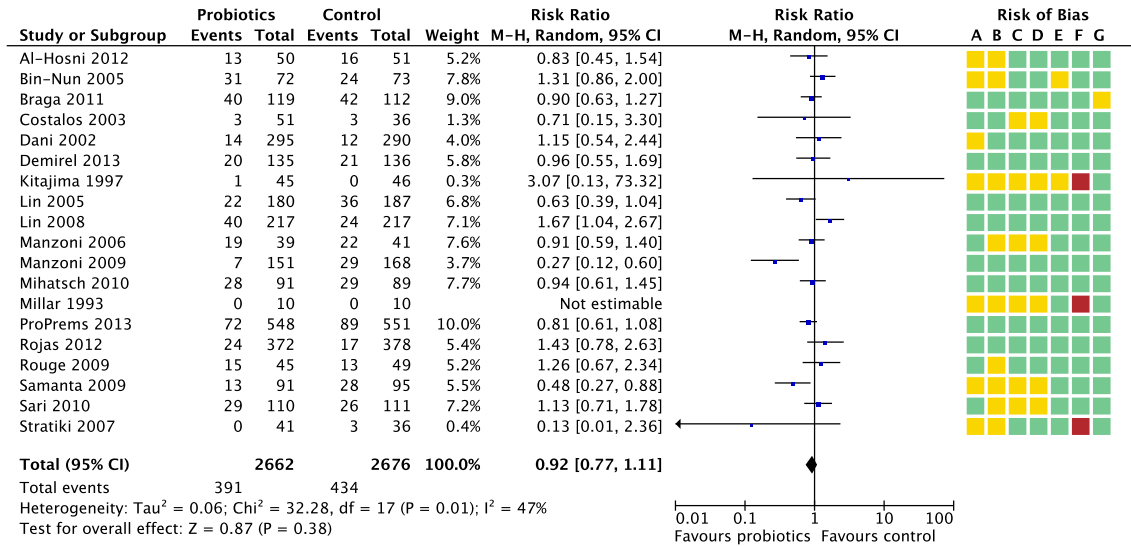
## References for table 5

- [1] AlFaleh K, Anabrees J. Probiotics for prevention of necrotizing enterocolitis in preterm infants. *Cochrane Database Syst Rev.* 2014;CD005496.
- [2] Olsen R, Greisen G, Schroder M, Brok J. Prophylactic Probiotics for Preterm Infants: A Systematic Review and Meta-Analysis of Observational Studies. *Neonatology.* 2016;109:105-12.

## Figures

Figure 1

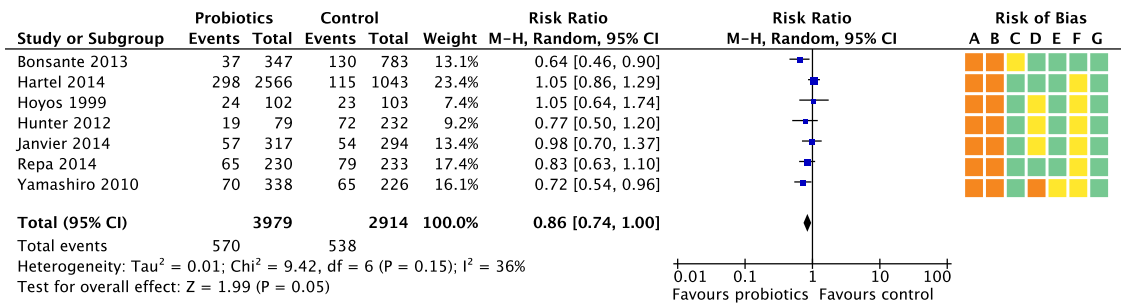
A) RANDOMIZED STUDIES



RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)

- 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

B) NON-RANDOMIZED STUDIES



RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)

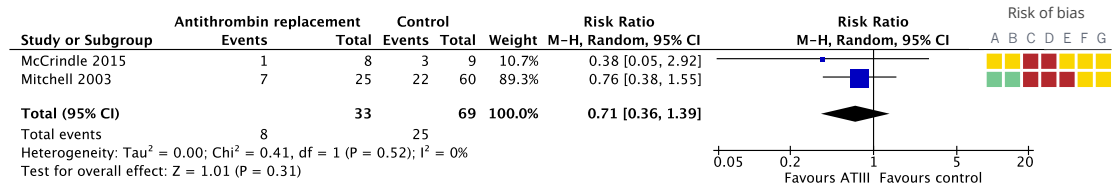
- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

Figure 1. Forest plots. Randomized and non-randomized studies. Probiotic supplementation in preterm infants in the neonatal intensive care unit. Outcome: culture proven sepsis. Colours indicate risk of bias (RoB) judgments. For randomized studies, low RoB=green; unclear RoB=yellow, and high RoB=red. For non-randomized studies, low RoB=green; moderate RoB=yellow; high RoB=orange; and critical RoB=red.

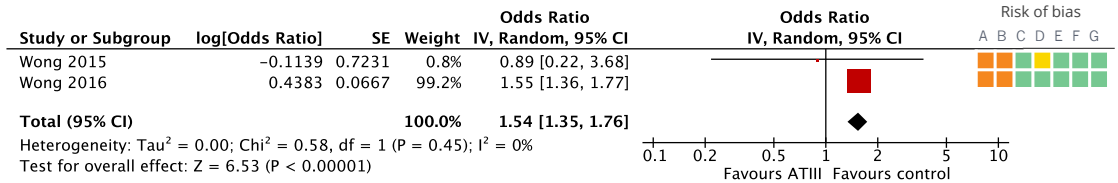


Figure 2

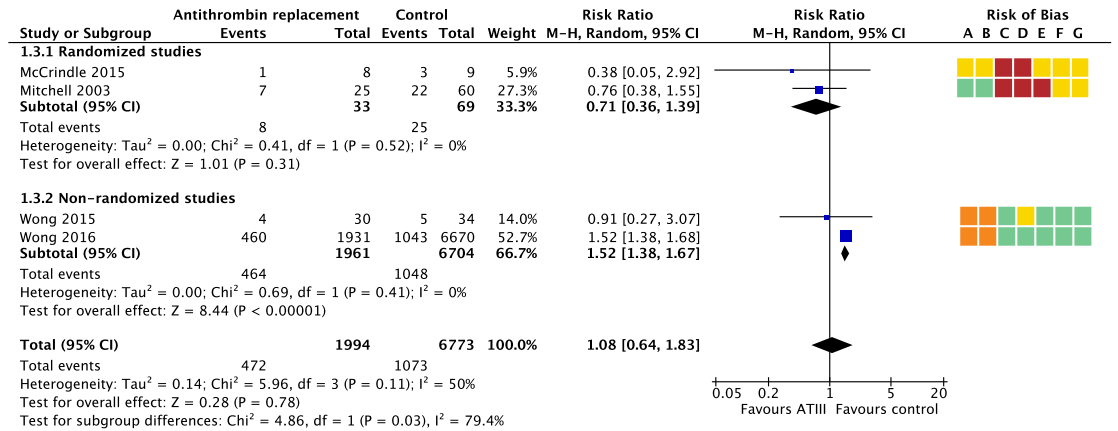
A) RANDOMIZED STUDIES



B) NON-RANDOMIZED STUDIES



C) RANDOMIZED AND NON-RANDOMIZED STUDIES



RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

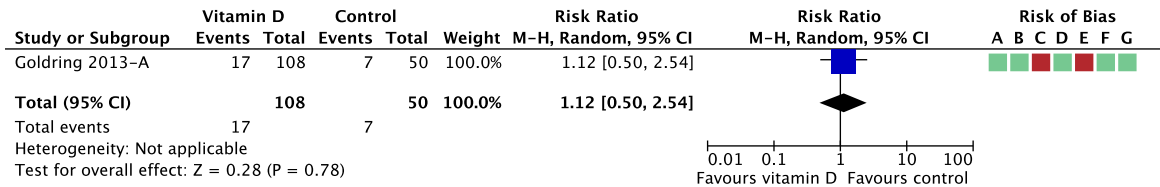
RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)

- 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

Figure 2. Forest plots of randomized and non-randomized studies. Antithrombin III replacement for the prevention of arterial or venous thrombosis in children undergoing extracorporeal membrane oxygenation therapy. (A) only RS, (B) only NRS, and (C) both RS and NRS by subgroups. Colours indicate risk of bias (RoB) judgments. For randomized studies, low RoB=green; unclear RoB=yellow, and high RoB=red. For non-randomized studies, low RoB=green; moderate RoB=yellow; high RoB=orange; and critical RoB=red.

Figure 3

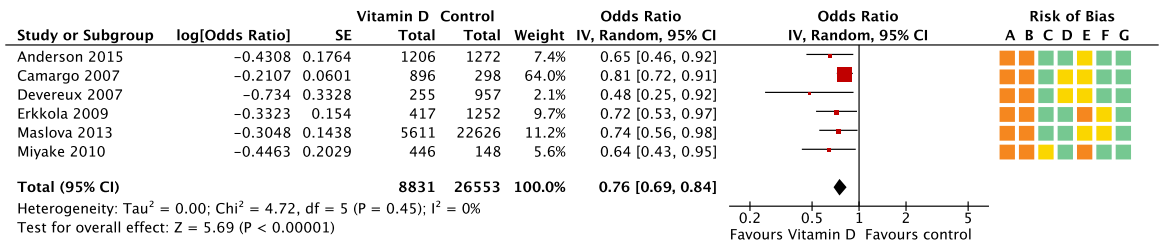
**A) RANDOMIZED STUDIES**



**RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)**

- 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

**B) NON-RANDOMIZED STUDIES**



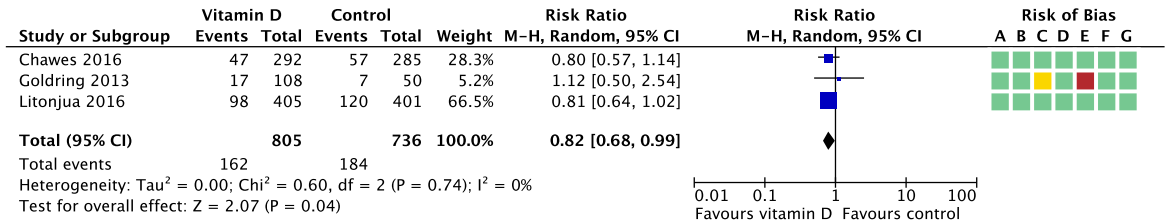
**RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)**

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

Figure 3. Forest plots, randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Colours indicate risk of bias (RoB) judgments. For randomized studies, low RoB=green; unclear RoB=yellow, and high RoB=red. For non-randomized studies, low RoB=green; moderate RoB=yellow; high RoB=orange; and critical RoB=red.

Figure 4

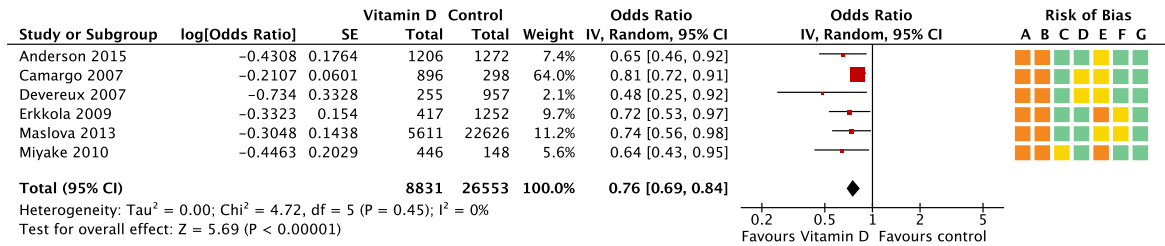
A) RANDOMIZED STUDIES



RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)

- 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

B) NON-RANDOMIZED STUDIES



RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

Figure 4. Forest plots updated –January 2017. Randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Colours indicate risk of bias (RoB) judgments. For randomized studies, low RoB=green; unclear RoB=yellow, and high RoB=red. For non-randomized studies, low RoB=green; moderate RoB=yellow; high RoB=orange; and critical RoB=red.

Figure 5

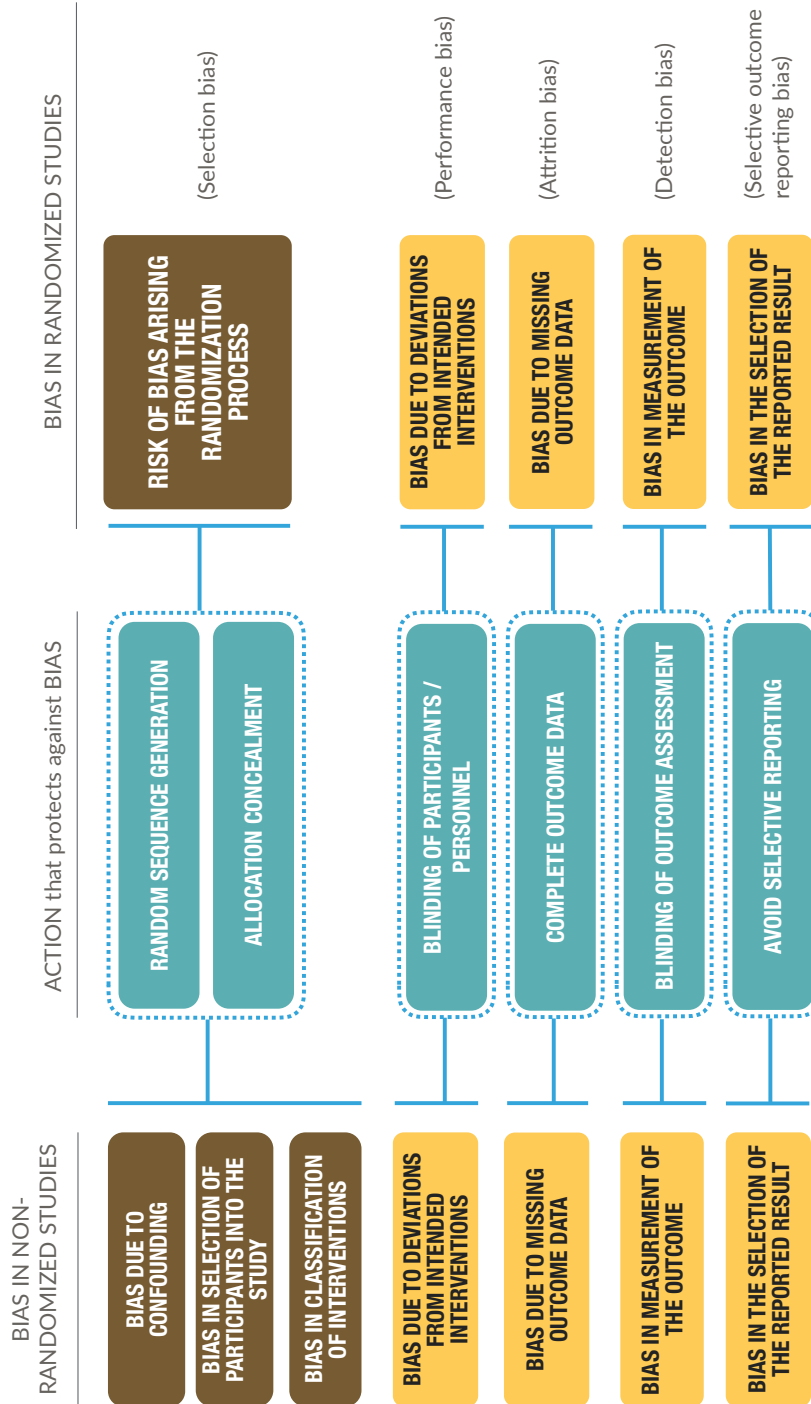


Figure 5. Types of bias met in non-randomized studies (left column) and randomized studies (right column) with the situations or actions performed in a randomized trial that protect against these biases on each type of study (center column). In parentheses are depicted other terms for biases.



Figure 6

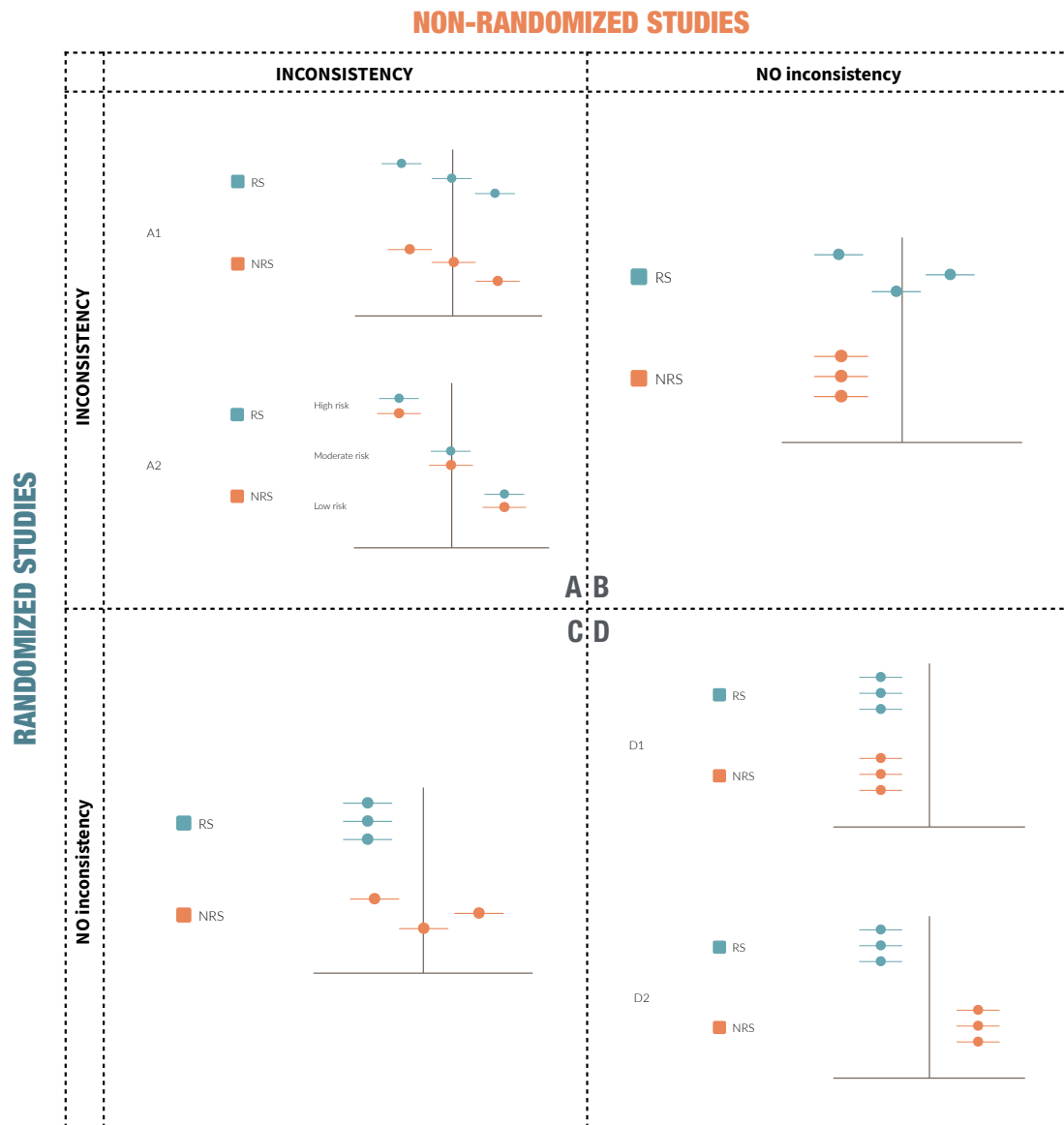


Figure 6. Inconsistency in randomized (blue colors) and non-randomized studies (red colors) distributed in four possible scenarios where A= a scenario where RS and NRS present both inconsistency; B= RS present inconsistency but NRS have similar results; C= RS present similar results among them, but NRS present inconsistency; D=both RS and NRS have no inconsistency. See text for discussion.

## Appendices

## **Supplementary material 1**

### **Example 1**

A clinical guideline for the prevention and treatment of necrotizing enterocolitis (NEC) in the neonatal intensive care unit is conducting a systematic review on the effects of prophylactic supplementation of probiotics to all premature infants on the outcomes NEC, overall mortality, and sepsis. The review team discusses whether to include NRS for any of these outcomes, realizing that there is an overall high certainty from RS for the outcomes NEC and mortality, so there is no need to look for NRS for these outcomes. However, for the outcome sepsis, the overall certainty of the evidence is deemed low mostly due to inconsistency and imprecision (figure 1a and table 1). The panel decides they would feel more comfortable by looking for NRS, especially when sepsis has been linked to the use of probiotics in very preterm babies in observational studies. [1] Authors find seven NRS with an overall low certainty of the evidence due to the inherent risk of bias, and no reasons to upgrade (table 1 and figure below).

### **Analysis**

For the outcome sepsis, guideline panelists can use NRS alone to generate a recommendation, most likely conditional in favor of supplementing probiotics given the precision provided; this is, probiotics would reduce to at best 48 fewer cases of sepsis and at worst zero (see table 1). Had authors looked only at the evidence from

RS, the recommendation could have been made either in favor, against, or even neutral, given the uncertainty of the evidence and the background information from case reports linking sepsis to probiotics. The use of NRS provides more certainty, and decision-makers can feel more comfortable with a recommendation in favor since probiotics is a low-cost intervention with better a balance of desirable vs undesirable effects. In this case, both bodies of evidence provide similar certainty, and authors could decide to present both or use the one that provides the highest confidence.

If the guideline panel and the review team decide to depict the two bodies of evidence in evidence profiles and summary of findings they have three options: (a) use two separate tables, one for each type of study; (b) display both RS and NRS separately in two rows and express that the recommendation was made based mainly on the effect from the NRS; or (d) use both designs, merging data in a single row in the EP, and even in a single pooled estimate (last row in red of table 1 and 2). This last option will require caution and methodological expertise to avoid misuse.

It is important to note several occurrences from this example. The effect estimates are similar in magnitude and direction, which give authors a sense of confidence that the estimated effect is close to the true estimate. When using ROBINS-I as the risk of bias tool for assessing NRS, the initial assessment of risk of bias should be rated as high (instead of low), and then downgrade accordingly for confounding and selection

of participants if authors consider that there is no reason against it. In table 2, we can see the risk of bias domain in NRS deemed very serious (highlighted in yellow) due to confounding and possible selection of participants; this rates down the certainty of the evidence of this outcome by two levels from high to low. More testing is needed in this area.

## REFERENCES

- [1] Dani C, Coviello CC, Corsini II, Arena F, Antonelli A, Rossolini GM. Lactobacillus Sepsis and Probiotic Therapy in Newborns: Two New Cases and Literature Review. *AJP Rep.* 2016;6:e25-9.

Table 1 of example 1

Table 1. Evidence profile. Randomized and non-randomized studies. Probiotic supplementation in preterm infants in the neonatal intensive care unit. Outcome: culture proven sepsis.

N° of studies	Study design	Quality assessment						N° of patients		Effect		Quality	Importance
		Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Probiotics	Control	Relative (95% CI)	Absolute (95% CI)			
<b>Sepsis – Randomized studies [1]</b>													
19	randomised trials	not serious	serious <sup>a</sup>	not serious	serious <sup>b</sup>	none	391/2662 (14.7%)	434/2676 (16.2%)	RR 0.92 (0.77 to 1.11)	13 fewer per 1,000 (from 18 more to 37 fewer)	⊕⊕○○ LOW	CRITICAL	
<b>Sepsis – Non-randomized studies [2]</b>													
7	observational studies	not serious <sup>c</sup>	not serious	not serious	not serious	none	570/3979 (14.3%)	538/2914 (18.5%)	RR 0.86 (0.74 to 1)	26 fewer per 1,000 (from 0 fewer to 48 fewer)	⊕⊕○○ LOW	CRITICAL	
<b>Sepsis – Randomized and non-randomized studies merged<sup>d</sup></b>													
26	Studies	not serious	not serious	not serious	not serious	none	9617/6641 (14.5%)	972/5590 (17.4%)	RR 0.90 (0.79 to 1.01)	17 fewer per 1,000 (from 2 more to 37 fewer)	⊕⊕⊕○ MODERATE	CRITICAL	

a. Statistical heterogeneity ( $I^2 = 47\%$ ) and visually, confidence intervals do not overlap in several studies (see forest plot). Possibly explained by different population risks and sample size.  
 b. Confidence intervals still include thresholds for an appreciable benefit or harm.  
 c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected. Nonetheless, residual confounding was not possible to rule out.  
 d. This merge is hypothetical and still subject to methodologists' preferences.

Table 2 of example 1.

Table 2. Evidence profile. Randomized and non-randomized studies with the use of ROBINS-I. Probiotic supplementation in preterm infants in the neonatal intensive care unit. Outcome: culture proven sepsis.

Nº of studies	Study design	Quality assessment						Nº of patients		Effect		Quality	Importance
		Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Probiotics	Control	Relative (95% CI)	Absolute (95% CI)			
<b>Sepsis – Randomized studies [1]</b>													
19	randomised trials	not serious	serious <sup>a</sup>	not serious	serious <sup>b</sup>	none	391/2662 (14.7%)	434/2676 (16.2%)	RR 0.92 (0.77 to 1.11)	13 fewer per 1,000 (from 18 more to 37 fewer)	⊕⊕○○ LOW	CRITICAL	
<b>Sepsis – Non-randomized studies [2]</b>													
7	observational studies	very serious <sup>c</sup>	not serious	not serious	not serious	none	570/3979 (14.3%)	538/2914 (18.5%)	RR 0.86 (0.74 to 1)	26 fewer per 1,000 (from 0 fewer to 48 fewer)	⊕⊕○○ LOW	CRITICAL	
<b>Sepsis – Randomized and non-randomized studies pooled<sup>d</sup></b>													
26	Studies	serious	not serious	not serious	not serious	none	961/6641 (14.5%)	972/5590 (17.4%)	RR 0.90 (0.79 to 1.01)	17 fewer per 1,000 (from 2 more to 37 fewer)	⊕⊕⊕○ MODERATE	CRITICAL	

a. Statistical heterogeneity ( $I^2 = 47%$ ) and visually, confidence intervals do not overlap in several studies (see forest plot). Possibly explained by different population risks and sample size.

b. Confidence intervals still include thresholds for an appreciable benefit or harm.

c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected. Nonetheless, residual confounding was not possible to rule out. In this case, ROBINS-I was used as the risk of bias tool.

d. This merge is hypothetical and subject to methodologists' preferences.

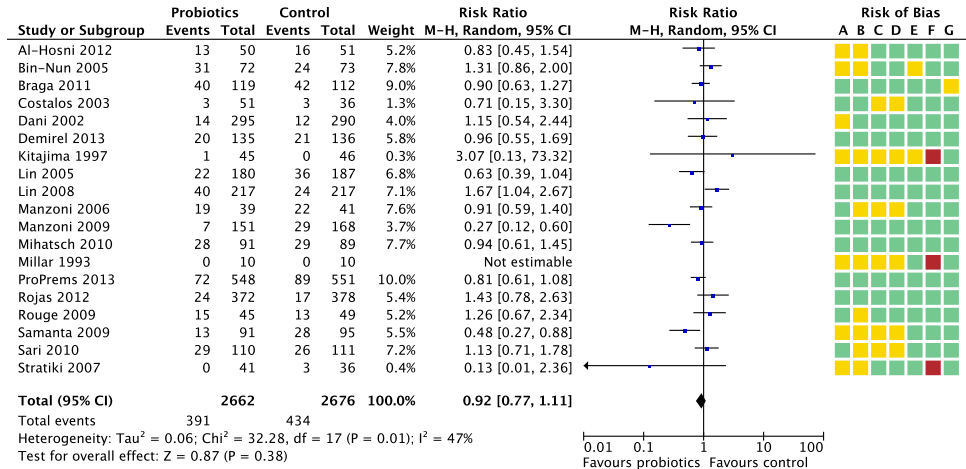
95%CI, 95% confidence interval



Figure 1 of example 1

Figure 1 of example 1. Forest plots. Randomized and non-randomized studies. Probiotic supplementation in preterm infants in the neonatal intensive care unit. Outcome: culture proven sepsis.

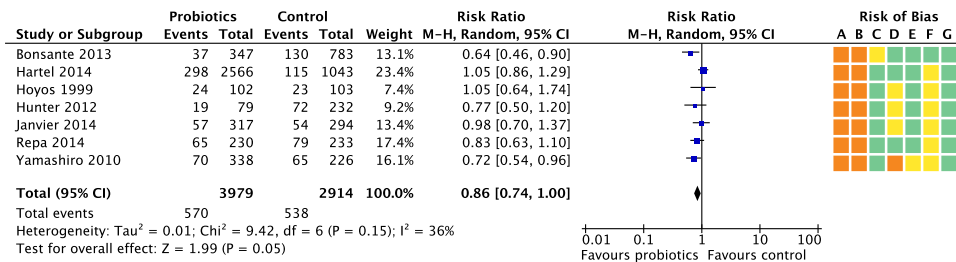
**A) RANDOMIZED STUDIES**



**RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)**

- 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

**B) NON-RANDOMIZED STUDIES**



**RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)**

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

## **Supplementary material 2.**

### **Example 2**

A panel of experts is conducting a clinical practice guideline on thromboprophylaxis in the Pediatric Intensive Care Unit. They are assessing the guideline question: “Should antithrombin III (AT-III) versus no AT-III be used in infants undergoing extra-corporeal membrane oxygenation (ECMO) for the prevention of thrombosis (arterial or venous)?” Their search strategy yields four randomized studies (RS) comparing AT-III to placebo, of which only two assess populations in the desired age group, i.e., children above one month of age (Table 1a). The review team decides to look for non-randomized studies (NRS), eventually finding eight studies, of which only two directly assess the population of interest. (Table 1b) By creating evidence profiles the team realizes that the certainty in the evidence from RS is very low due to risk of bias, imprecision, and indirectness, while the certainty from NRS is deemed low. (Table 2 and Figure)

### **Analysis**

In this case, searching for NRS was justified due to important indirectness from RS at the population and intervention level that together with risk of bias and imprecision give RS a certainty of very low, while that from NRS is deemed low certainty due to risk of bias (confounding). The authors’ objective is to create a clinical recommendation and its direction might be different if using RS or NRS. If

the research team utilizes evidence from RS alone (very low certainty) it could end with a conditional recommendation either against the intervention, in favor, or neutral. On the other hand, by using only

NRS, the team would certainly be more inclined for a recommendation against the intervention due to concerns over the increased number of thrombosis in the intervention group, yet with low certainty (Table 2 and figure). In this example, it is recommended to use only the evidence from NRS as it provides more certainty, also, indirectness has important influence on this decision.

**Table 1 of example 2.**

Table 1. Included and excluded individual studies for the research question, any outcome.

Study	Population	Outcomes	Notes
<b>A. Randomized studies [1-4]</b>			
Fulia 2003	Preterm infants ≤30 weeks of gestational age with more than 12 hours of postnatal age, and an ATIII activity ≤40%.	Mortality, bleeding, intra-ventricular hemorrhage	Excluded. Population age considered too indirect.
Schmidt 1998	Preterm infants (weight 750 - 1900 g); post-natal age 2 to 12 h; endotracheal intubation and mechanical ventilation for RDS, and indwelling arterial catheter.	Mortality, bleeding, intra-ventricular hemorrhage	Excluded. Population age considered too indirect.
McCrindle 2015	17 infants (8 in ATIII group, 9 in control) undergoing heart surgery and cardiopulmonary bypass.	Thrombosis, bleeding, infection	<b>Included.</b> Although indirect on how the intervention is administered.
Mitchell 2003	Children with acute leukemia and asparaginase treatment.	Thrombosis, bleeding	<b>Included,</b> although indirect population (children with leukemia)
<b>B. Non-randomized studies [5-12]</b>			
Hausmann 2006	Children (0.2 to 19.6 years of age) undergoing stem cell transplantation	Mortality, bleeding, thrombosis	Excluded due to indirectness of population.
Wong 2015	Children undergoing ECMO for respiratory failure	Mortality, bleeding, thrombosis	<b>Included</b>
Wong 2016	Children undergoing ECMO for respiratory failure	Mortality, bleeding, thrombosis	<b>Included</b>
Corder 2014	Neonates / infants with thrombosis and treated with enoxaparin (treatment, not prophylaxis)	Bleeding, thrombosis	Excluded. Population with thrombosis (i.e., not prophylaxis)
Niebler 2011	Pediatric patients on ECMO	Mortality, bleeding	Retrospective chart review of cases and controls. Study with zero thrombotic events. Excluded
Perry 2013	Neonates with congenital diaphragmatic hernia requiring ECMO	Bleeding	Excluded. Indirect population, and does not evaluate the outcome of VTE.
Petaja 1999	Neonates who underwent heart surgery	Mortality, thrombosis	Excluded due to indirectness of population (neonates undergoing heart surgery)
Stansfield 2016	Infants requiring ECMO.	Mortality, bleeding	Excluded due to indirectness (clots were measured in the ECMO circuit, and not in patients); also, zero events.

AT-III, antithrombin III; RDS, respiratory distress syndrome; ECMO, extra-corporeal membrane oxygenation; VTE, venous thromboembolism.

## References for example 2

- [1] Fulia F, Cordaro S, Meo P, Gitto P, Gitto E, Trimarchi G, et al. Can the administration of antithrombin III decrease the risk of cerebral hemorrhage in premature infants? *Biol Neonate*. 2003;83:1-5.
- [2] McCrindle BW, Manlhiot C, Holtby HM, Chan AK, Brandao LR, Rolland M, et al. Abstract 18061: Supplementation to Treat Antithrombin Deficiency Improves Sensitivity to Heparin, Anticoagulation and Decreased Thrombogenicity in Neonates and Infants Undergoing Cardiac Surgery With Cardiopulmonary Bypass. *Circulation*. 2015;132:A18061-A.
- [3] Mitchell L, Andrew M, Hanna K, Abshire T, Halton J, Wu J, et al. Trend to efficacy and safety using antithrombin concentrate in prevention of thrombosis in children receiving l-asparaginase for acute lymphoblastic leukemia. Results of the PAARKA study. *Thromb Haemost*. 2003;90:235-44.
- [4] Schmidt B, Gillie P, Mitchell L, Andrew M, Caco C, Roberts R. A placebo-controlled randomized trial of antithrombin therapy in neonatal respiratory distress syndrome. *Am J Respir Crit Care Med*. 1998;158:470-6.
- [5] Corder A, Held K, Oschman A. Retrospective evaluation of antithrombin III supplementation in neonates and infants receiving enoxaparin for treatment of thrombosis. *Pediatr Blood Cancer*. 2014;61:1063-7.
- [6] Haussmann U, Fischer J, Eber S, Scherer F, Seger R, Gungor T. Hepatic veno-occlusive disease in pediatric stem cell transplantation: impact of pre-emptive antithrombin III replacement and combined antithrombin III/defibrotide therapy. *Haematologica*. 2006;91:795-800.

- [7] Niebler RA, Christensen M, Berens R, Wellner H, Mikhailov T, Tweddell JS. Antithrombin replacement during extracorporeal membrane oxygenation. *Artif Organs*. 2011;35:1024-8.
- [8] Perry R, Stein J, Young G, Ramanathan R, Seri I, Klee L, et al. Antithrombin III administration in neonates with congenital diaphragmatic hernia during the first three days of extracorporeal membrane oxygenation. *J Pediatr Surg*. 2013;48:1837-42.
- [9] Petäjä J, Peltola K, Rautiainen P. Disappearance of symptomatic venous thrombosis after neonatal cardiac operations during antithrombin III substitution. *The Journal of Thoracic and Cardiovascular Surgery*. 1999;118:955-6.
- [10] Stansfield BK, Wise L, Ham PB, 3rd, Patel P, Parman M, Jin C, et al. Outcomes following routine antithrombin III replacement during neonatal extracorporeal membrane oxygenation. *J Pediatr Surg*. 2016;52:609-13.
- [11] Wong TE, Delaney M, Gernsheimer T, Matthews DC, Brogan TV, Mazor R, et al. Antithrombin concentrates use in children on extracorporeal membrane oxygenation: a retrospective cohort study. *Pediatr Crit Care Med*. 2015;16:264-9.
- [12] Wong TE, Nguyen T, Shah SS, Brogan TV, Witmer CM. Antithrombin Concentrate Use in Pediatric Extracorporeal Membrane Oxygenation: A Multicenter Cohort Study. *Pediatr Crit Care Med*. 2016;17:1170-8.

Table 2 of example 2

Table 2. Evidence profile. Randomized and non-randomized studies. Antithrombin III for the prevention of thrombosis in children undergoing ECMO. Outcome: Any Thrombosis.

Nº of studies	Study design	Quality assessment						Nº of patients		Effect		Quality	Importance
		Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Control	ATIII	Relative (95% CI)	Absolute (95% CI)			
<b>Thrombosis – Randomized Studies [1, 2]</b>													
2	randomized trials	serious <sup>a</sup>	not serious	serious <sup>b</sup>	serious <sup>c</sup>	none	8/33 (24.2%)	25/69 (36.2%)	RR 0.71 (0.36 to 1.39)	105 fewer per 1,000 (from 141 more to 232 fewer)	⊕○○○ VERY LOW	CRITICAL	
<b>Thrombosis – Non-randomized Studies [3, 4]</b>													
2	observational studies	not serious serious <sup>d</sup>	not serious	not serious	not serious	none	464/1961 (23.7%)	1048/6704 (15.6%)	OR 1.54 (1.35 to 1.76)	66 more per 1,000 (from 44 more to 90 more)	⊕⊕○○ LOW	CRITICAL	

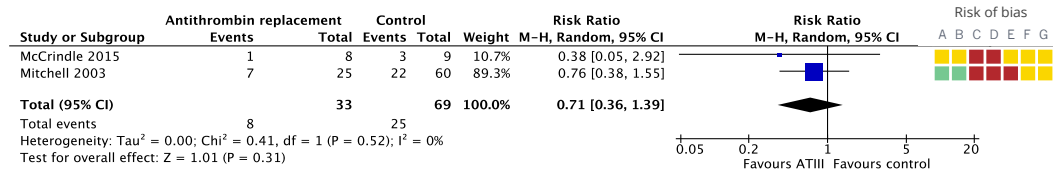
a. Both studies not blinded. There are concerns in one study about the random sequence generation and incomplete outcome data.  
 b. One study assesses children undergoing cardiac surgery while other includes children with leukemia and asparaginase treatment.  
 c. Wide confidence intervals that do not exclude a plausible harm or benefit. Also, very low number of events and participants.  
 d. Residual confounding was not possible to exclude even after authors perform a good assessment and adjustment of confounders (in this example ROBINS-i is not used).

95%CI, 95% confidence interval  
 ATIII: Antithrombin III

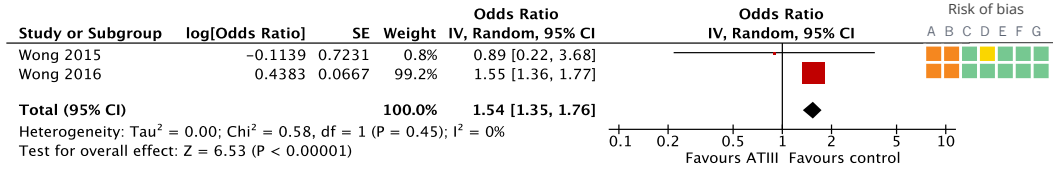
## Figure of example 2

Figure 1. Case 1 forest plots, randomized and non-randomized studies. Antithrombin III replacement for the prevention of arterial or venous thrombosis in children undergoing extracorporeal membrane oxygenation therapy. (A) only RS, (B) only NRS, and (C) both RS and NRS by subgroups.

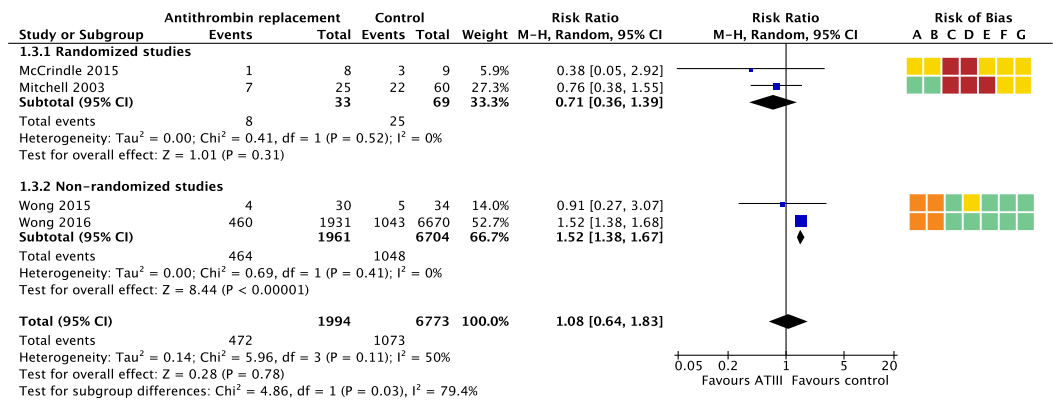
### A) RANDOMIZED STUDIES



### B) NON-RANDOMIZED STUDIES



### C) RANDOMIZED AND NON-RANDOMIZED STUDIES



#### RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

#### RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)

- 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



## **Supplementary material 3**

### **Example 3**

In January 2016, a Health Technology Assessment unit is working on the question: “Should vitamin D be supplemented to all pregnant women for the prevention of recurrent wheeze or asthma in their infants?”

Willing to include NRS, the review team evaluates the body of evidence from RS first. Only one published RS is found (figure 1) with an overall certainty in the evidence classified as very low due to serious risk of bias and very serious imprecision (table 1). The team decides to search NRS. They find and include six which are assessed and deemed low certainty of evidence due to risk of bias (confounding). However, in consultation with content experts, the authors judged that there is a dose-response effect (inverse relationship between the adjusted ORs and increased dosages or levels of vitamin D) which upgrade the Certainty by one level, from low to moderate.

### **Analysis**

In this case, the certainty in the evidence from NRS is higher than RS. The main difference between RS and NRS in the GRADE domains results from the very serious imprecision from RS and the risk of bias. It could be argued that the observed dose- response effect in NRS results from a dose dependent confounding, that is, the higher the vitamin D exposure the greater is the influence of residual confounding

on the outcome. If that were the case the raters should not upgrade for dose response effects and the overall Certainty from NRS would end up as low rather than moderate.

Authors should consider which option provide the least biased alternative that provides the highest confidence for decision-making. In this case, they could use only the evidence from NRS to provide a recommendation in favor of vitamin D given the moderate certainty in the evidence (and if we believe in the dose response effect) after comparing it with the body of evidence from RS (deemed very low certainty).

## **Epilogue**

A year later (January 2017) an updated search yields two more RS added to the meta- analysis (figure 2). Three studies now provide an effect estimate excluding the null (45 fewer cases of asthma/wheezing per 1,000 treated; 95% C.I. from 80 fewer to 3 fewer; see table 2) and a final certainty deemed as moderate (only downgraded one level due to imprecision). After adding this evidence, authors would be more comfortable with a conditional (or even strong) recommendation in favor of vitamin D.

Two important details to note: first, RS and NRS are now at the same certainty level, still with some (arguable) differences in the precision of the effect estimates and the lingering risk of bias (confounding) from NRS; second, the direction of both effect

estimates concur in favor of vitamin D. Authors could feel reassured in portraying both study designs in a single table (table 2) either in two rows, or even pooled into a single effect estimate, if feasible and sensitivity analyses are performed to assess the effects of the study designs after pooling. The only difference among the GRADE domains (besides the risk of bias) stems from the imprecision from RS, which can be averted with the incorporation of NRS.

Table 1 of example 3

Table 1. Evidence profile of randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Outcome: Asthma/recurrent wheezing.

Nº of studies	Quality assessment							Nº of patients		Effect		Quality	Importance
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Vitamin D	Control	Relative (95% CI)	Absolute (95% CI)			
<b>Asthma / recurrent wheezing - Randomized studies [1]</b>													
1	randomised trials	serious <sup>a</sup>	not serious	not serious	very serious <sup>b</sup>	none	17/108 (15.7%)	7/50 (14.0%)	RR 1.12 (0.50 to 2.54)	17 more per 1,000 (from 70 fewer to 216 more)	⊕○○○ VERY LOW	CRITICAL	
<b>Asthma / recurrent wheezing - Non-randomized studies [2-7]</b>													
6	non-randomized studies	not serious <sup>c</sup>	not serious	not serious	not serious	dose response gradient <sup>d</sup>	total 8,831 <sup>e</sup>	26,553 (risk 14%) <sup>e</sup>	OR 0.76 (0.69 to 0.84)	30 fewer per 1,000 (from 20 fewer to 39 fewer)	⊕⊕⊕○ MODERATE	CRITICAL	

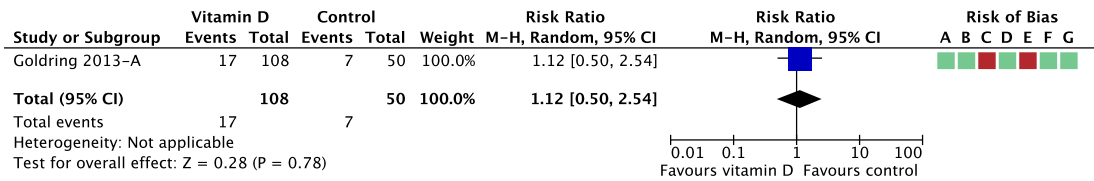
a. There were 22/180 participants who were not analyzed (lost to follow-up), 16% in the intervention group and 10% in the control group. Also, the outcome was a subjective measure and participants were not blinded to treatment allocation (reporter bias).  
 b. Wide confidence interval with low number of participants for the optimal information size; also, crossing the null and the appreciable thresholds for benefit and harm.  
 c. All studies have possible residual confounding. The non-randomized studies thus are downgraded two levels as in regular GRADE assessment. No further downgrading was considered necessary.  
 d. All studies demonstrated a significant dose-response association at different levels of vitamin D supplementation on the risk of asthma or wheezing.  
 e. No data on number of events on each arm; instead, all studies provide the adjusted odds ratios on the risk of asthma and its association with vitamin D intake. Baseline risk in control group was imputed from the rest of the studies, including the randomized controlled trial.

95%CI, 95% confidence interval

Figure 1 of example 3.

Figure 1. Forest plots, randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants.

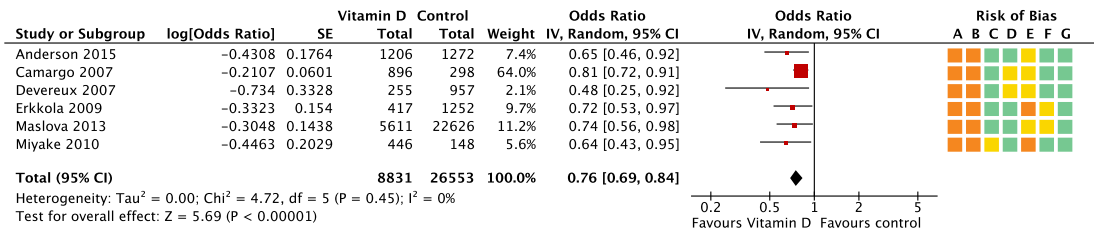
**A) RANDOMIZED STUDIES**



**RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)**

- 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

**B) NON-RANDOMIZED STUDIES**



**RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)**

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

Table 2 of example 3

Table 2. Evidence profile (UPDATED July 2017). Randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants. Outcome: Asthma/recurrent wheezing.

Nº of studies	Quality assessment							Nº of patients		Effect		Quality	Importance
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Vitamin D	Control	Relative (95% CI)	Absolute (95% CI)			
<b>Asthma / recurrent wheezing – Randomized studies [1-3]</b>													
3	randomised trials	not serious <sup>a</sup>	not serious	not serious	serious <sup>b</sup>	none	162/805 (20.1%)	184/736 (25%)	<b>RR 0.82</b> (0.68 to 0.99)	<b>45 fewer per 1,000</b> (from 3 fewer to 80 fewer)	⊕⊕⊕○ MODERATE	CRITICAL	
<b>Asthma / recurrent wheezing – Non-randomized studies [4-9]</b>													
6	non-randomized studies	not serious <sup>c</sup>	not serious	not serious	not serious	dose response gradient <sup>d</sup>	total 8,831 <sup>e</sup>	26,553 (risk 14%) <sup>e</sup>	<b>OR 0.76</b> (0.69 to 0.84)	<b>30 fewer per 1,000</b> (from 20 fewer to 39 fewer)	⊕⊕⊕○ MODERATE	CRITICAL	

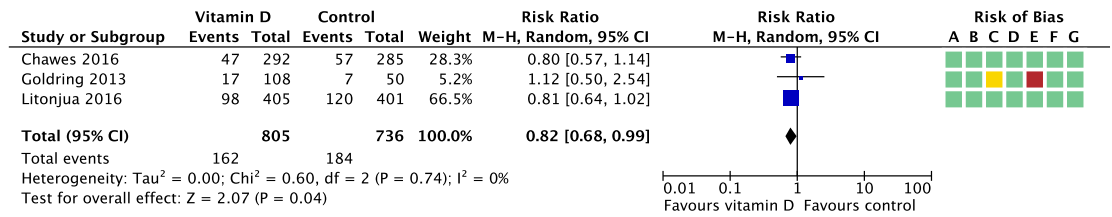
- a. In one study, the number of patients who were lost to follow-up were balanced (16% in the intervention group and 10% in the control group) and was not considered significant to change results.
- b. Confidence interval with low number of participants for the optimal information size; although not crossing the null and perhaps only the threshold for appreciable benefit.
- c. All studies have possible residual confounding. The non-randomized studies thus are downgraded two levels as in regular GRADE assessment. No further downgrading was considered necessary.
- d. All studies demonstrated a significant dose-response association at different levels of vitamin D supplementation on the risk of asthma or wheezing.
- e. No data on number of events on each arm; instead, all studies provide the adjusted odds ratios on the risk of asthma and its association with vitamin D intake. Baseline risk in control group was imputed from the rest of the studies, including the randomized controlled trial.

95%CI, 95% confidence interval

Figure 2 of example 3

Figure 2. Forest plots updated –January 2017. Randomized and non-randomized studies. Vitamin D supplementation in pregnant women for the prevention of asthma/wheezing in their infants.

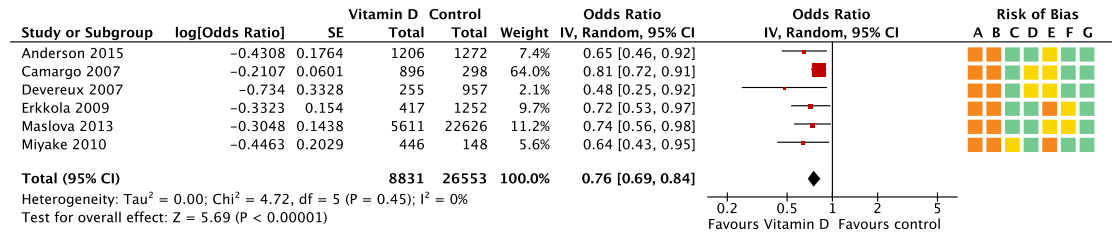
**A) RANDOMIZED STUDIES**



**RISK OF BIAS IN RANDOMIZED STUDIES (Cochrane tool)**

- 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

**B) NON-RANDOMIZED STUDIES**



**RISK OF BIAS IN NON-RANDOMIZED STUDIES (ROBINS-I)**

- A. Bias due to confounding
- B. Bias in selection of participants
- C. Bias in classification of interventions
- D. Bias due to departures from intended interventions
- E. Bias due to missing data
- F. Bias in measurement of outcomes
- G. Bias in selection of the reported result

# CHAPTER 4. GRADE GUIDANCE: THE ROLE OF RANDOMIZED AND NON- RANDOMIZED STUDIES IN KNOWLEDGE SYNTHESES OF HEALTH INTERVENTIONS

## **AUTHORS**

Carlos A. Cuello<sup>a</sup>, Rebecca L. Morgan<sup>a</sup>, Jan Brozek<sup>a</sup>, Nancy Santesso<sup>a</sup>, Jos Verbeek<sup>b</sup>, Kris Thayer<sup>c</sup>, Mohammed T. Ansari,<sup>d</sup> Gordon Guyatt<sup>a</sup>, Holger J. Schünemann<sup>a</sup>

## **AUTHORS AFFILIATIONS**

- a. Department of Health Research Methods, Evidence, and Impact, McMaster University. Hamilton Ontario, Canada.
- b. Cochrane Work Review Group. Finnish Institute of Occupational Health, Helsinki, Finland
- c. Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment. Environmental Protection Agency. USA
- d. School of Epidemiology and Public Health. University of Ottawa, Ottawa, Ontario. Canada.

## **CORRESPONDING AUTHOR**

Holger J. Schünemann, M.D., Ph.D.

Chair, Department of Health Research Methods, Evidence, and Impact.  
McMaster University. Health Sciences Centre Room 2C16.

1280 Main Street West. Hamilton, Ontario Canada. L8N 4K1



**Article history:**

- Survey Slides and case studies presented at GRADE meetings in Washington, D.C., U.S. (2016), Seoul, South Korea (2016), Rome, Italy (2017), and Hamilton, Canada (2017).
- Article first draft on 31-May-2017.
- Reviewed by Holger Schunemann, Gordon Guyatt, and Jan Brozek on July-2017.

Word count: 2,678

## Abstract

This is the 20th in the ongoing series of articles describing the GRADE approach to systematic reviews, guidelines, and health technology assessment.

Systematic review authors, guideline developers, and other knowledge synthesers' practitioners use randomized studies (RS) and non-randomized studies (NRS) as sources of evidence for questions about health interventions. Well conducted RS represent the most reliable individual source of evidence for estimating relative effects, primarily because of protection against confounding. NRS, however, can provide valuable information as complementary, sequential, or replacement evidence for RS.

This article provides guidance on how to integrate NRS with RS in a body of evidence for questions about health interventions, focusing on the implications of using one or both type of studies on the overall certainty of the evidence, and on the decision to include them in health recommendations. This guidance provides a framework to help authors, guideline panelists, and methodologists conducting knowledge syntheses using GRADE. The final sections of this article deal with requirements for further work.

**Keywords:** GRADE, quality of evidence, certainty of the evidence, risk of bias, non-randomized studies, ROBINS.

## Background

Randomized studies (RS) provide the best source of evidence for estimating effects<sup>5</sup> on outcomes in knowledge syntheses and health guidelines. Non-randomized studies (NRS) of representative populations provide the best evidence with respect to prognosis and baseline risk,<sup>1</sup> and are useful in many situations as replacement, sequential, or complementary evidence for using with a body of evidence of RS.<sup>2</sup> They are, however, limited by potential confounding and other biases.

Authors of knowledge syntheses evaluating a health question of an intervention require the most complete and least biased studies to present estimate of effects with the highest certainty, and guideline developers will need these syntheses to generate trustworthy recommendations. This is why most experts consider incorporation of NRS with RS in systematic reviews about interventions desirable.<sup>3</sup> In this article, although we will at times mention that NRS are ideal for assessing baseline risk, our focus is primarily on the use of NRS to generate relative estimates of effect of interventions.

If authors identify and decide to include both types of studies for their PICO question, they could face several challenges, specifically: how will their conclusions

---

<sup>5</sup> For the remainder of the discussion, we will use the term “estimates of relative effects”. The reader can assume we are referring to estimates of relative effect of interventions on binary outcomes or absolute effect from studies using continuous variables.

be affected if differences in the direction and magnitude of effects between study designs are found? what would be the effect of the differences in the individual GRADE domains, including risk of bias? and how should authors present results in evidence profiles and summary of findings tables? For this guidance, we will consider knowledge synthesis as any systematic review, rapid review, health technology assessment, or any other attempt to summarize all pertinent studies on a specific question.<sup>4</sup> This guidance will look from both the perspective of a knowledge synthesis author and the clinical guideline developer who aims at generating a recommendation.

This guidance is based on previous works, meetings, webinars, and workshops with members of Cochrane, the Guidelines International Network (GIN), and GRADE, with feedback and refinement from the GRADE project group on NRS and other GRADE members.<sup>2, 3, 5</sup> The Cochrane Methods Innovation grant, the National Toxicology Program in the U.S., and the McMaster GRADE centre have provided support for this project.

The first section of this guidance will consider reasons for integrating<sup>6</sup> NRS at the early stages of formulating a research question for a knowledge synthesis. The second section deals with the possible scenarios encountered when evaluating a body of

---

<sup>6</sup> \* For this discussion, when we use the term “integration” it will refer to any form of using RS and NRS together, either in the same synthesis, in the same summary of findings (same table but separated in rows), or in the same analysis (pooled into a single estimate).

evidence with RS and NRS. The third section explains how to portray RS and NRS in GRADE summary of findings tables and evidence profiles and the implications on the GRADE domains for certainty of evidence. Finally, we will discuss future areas of research and next steps, including the use of new tools for assessing the risk of bias of NRS.

## **1. How to consider inclusion of non-randomized studies in knowledge syntheses**

### **1.1. The importance and role of a protocol and search strategy**

Authors of knowledge syntheses must decide and declare from the outset (i.e., in the protocol of a systematic review) any pre-specified criteria about the type of study (NRS or RS) for which they will search and under which circumstance these articles should be evaluated or included (figure 1). This offers transparency and increases confidence in the results.<sup>6, 7</sup> It is important for authors to detail in the protocol stage their PICO question (patient, intervention, comparisons, and outcomes) and describe how a study that answers this question would be conducted by randomized controlled experimentation, regardless of the feasibility to do it.<sup>8</sup>

Authors may have reasons to search and include NRS irrespective of the availability of RS that yield high certainty evidence for primary intervention effect. The most

common would be looking for evidence regarding baseline risk, and including outcomes for which RS evidence would be sparse or unavailable (e.g. rare adverse outcomes). Another reason may be serious indirectness in the RS evidence. Systematic review authors who decide to include NRS should search for both types of evidence, with a filter that differentiates the two. Current reference managers, search strategies, and filters make this objective achievable.<sup>9,10</sup>

## **1.2. When to include NRS (eligibility criteria)**

The remainder of this discussion focuses on situations in which authors have concluded that they might plausibly find NRS that complement or replace RS with respect to estimates of relative effect (i.e. experts know of NRS that yield at least low quality evidence). Having completed the search, authors should first do a complete assessment of the RS, including the GRADE assessment of certainty of evidence (see figure 1). If the GRADE assessment reveals high certainty of evidence, further evaluation of NRS to complement estimates of relative effect are not necessary, except in extraordinary circumstances when authors are aware of NRS that are likely to yield moderate certainty evidence, this also applies when RS assessments show moderate certainty. Authors must consider this issue for all patient-important outcomes; RS may provide high certainty for benefit but not for harm outcomes, particularly when these outcomes are rare. When authors face evidence from RS deemed moderate, low, or very low certainty, the evaluation of the eligible NRS has

the potential to be helpful. Eligibility criteria for the NRS should be restricted to studies that will plausibly yield evidence of equal certainty to the RS – for instance, unbiased NRS with adequate sample size that undertook adjustment for key prognostic variables and achieved satisfactory follow-up.

## **2. Optimal use of randomized and non-randomized studies**

### **2.1. Possible scenarios when dealing with two bodies of evidence**

Figure 2 presents the possible combinations of results that may emerge from certainty ratings of the RS and NRS bodies of evidence. Although 16 possible combinations are theoretically possible, a number are extremely implausible (i.e., that both NRS and RS are high certainty – cell A), highly unlikely (NRS with high certainty – the first column, cells E, I, and M), or straightforward (if RCTs have high certainty – the first row, cells A to D – under most circumstances one shouldn't bother assessing the NRS).

### **2.2. Using non-randomized rather than randomized studies**

When NRS provide higher certainty than RS one needn't present the RS results (green cells from figure 2). Take for instance the case of the use of vitamin D in

pregnant women for the prevention of asthma or wheezing in infants, where RS provide very low certainty evidence (due to very serious imprecision and risk of bias) while NRS provide moderate certainty evidence (due to very serious risk of bias that eventually is upgraded because of a dose-response effect); in this case NRS can be utilized alone without considering RS.<sup>11</sup> However, one should always consider exceptions where RS could still be included, especially when they provide valid information or improvements in GRADE domains.

Consider, for example, one recent systematic review<sup>12</sup> comparing the failure rate of antibiotics versus appendectomy in children with uncomplicated appendicitis (table 1), which includes a body of evidence of RS deemed low certainty (only one RS included with very serious imprecision), while the body of evidence from NRS is considered moderate certainty (downgraded two levels for serious risk of bias, then upgraded one level for strong association). In this case (corresponding to cell J from figure 2) some experts using this review for a recommendation in a health guideline might judge that the reason to downgrade the RS (very serious imprecision without concerns in the other GRADE domains) makes the integration of RS and NRS possible, either in a single summary of findings table (keeping separated in two rows) or merged into a single pooled estimate. Other experts might judge that the body of evidence of NRS is sufficient to make a recommendation in favor of surgery and will discard the RS.



### **2.3. Using either or both types of studies**

Previous GRADE guidance suggested, when certainty of evidence was the same in RS and NRS (orange cells from figure 2), to present both bodies of evidence in separate rows in evidence profiles.<sup>13</sup> Here we provide new guidance for presenting and integrating both types of studies and associated considerations.

When bodies of evidence from RS and NRS provide the same level of certainty, the authors of a systematic review or a health guideline can use both RS and NRS in a single summary of findings, separated or pooled (see 3.2 below). However, on occasions they must decide which of the two bodies of evidence leaves them with higher certainty, always considering each GRADE domain affected and the implications on the final recommendation.

Consider, for instance, a guideline panel assessing a question about routine use of probiotics in preterm infants in the neonatal intensive care unit with the intention of preventing necrotizing enterocolitis (table 2). When evaluating the outcome of culture proven sepsis, the authors rate down the body of evidence of RS for imprecision and inconsistency (giving a low certainty) while NRS are also deemed low certainty due to lack of randomization and consequential confounding, but no other concerns. In this case, both bodies of evidence are at the same level of certainty (low) and in most situations, it would be preferable to present both RS and NRS in summary of findings so decision-makers can reach a recommendation by viewing

both; however, in this example, the panel could feel that the most trustworthy information comes from NRS given the precision provided –excluding the null and the appreciable threshold of harm– and the notion that NRS are preferable for evaluating harm outcomes.

### **3. Presenting in GRADE tables**

If authors, by following guidance from figure 1 and considering the points described above (2.1 to 2.3), decide to include RS and NRS for estimating relative effects, they have several options to portray both bodies of evidence in evidence profiles and summary of findings tables (figure 3 and appendices 1, 2, and 3).

#### **3.1. Alternative presentations**

Consistent with prior GRADE guidance,<sup>13</sup> and preferred by experts<sup>3</sup> especially when confronting RS and NRS with similar certainty, presenting the findings from the two bodies of evidence in adjacent rows of the summary of findings represents the preferred approach. Presenting the two bodies of evidence in separate tables represents a reasonable alternative (figure 3).

#### **3.2. Combining both type of study designs**

Considerations motivating this guidance have included differences in risk of bias and the possibility of large NRS dominating small RS. If, however, authors have rated

down the RS for risk of bias, and the NRS residual confounding and selection bias within the ROBINS-I framework leave NRS at a similar risk of bias as RS then one can contemplate pooling all the studies in the same analysis<sup>8</sup>. This would require not only similar overall risk of bias, but also similar results. We anticipate such situations will be unusual. Nonetheless, we present an example on how a summary of findings table like this would appear (appendix 3), and we require additional examples to provide more detailed guidance on these specific situations.

## **4. The role of ROBINS-I**

Until now, we have assessed the integration of both bodies of evidence in GRADE irrespective of which risk of bias assessment tools had been used. GRADE does not suggest using one risk of bias tool over another. However, the use of ROBINS-I in GRADE assessments may facilitate comparison of evidence between RS and NRS because they are placed on a common metric for the assessment of risk of bias.<sup>5</sup> As detailed in section 3.2, integrating both study designs will require considering the methodological similarities between RS and NRS. ROBINS-I suggests that a low risk of bias NRS is equivalent to a well conducted RS answering the same PICO question. In such cases, the main difference between RS and NRS results from the randomization process, which protects essentially from the first three biases depicted in figure 4. If the assumption that NRS have none or minimal concerns regarding

confounding and selection bias holds (e.g. in well conducted interrupted time series), there should be no concerns when these NRS are integrated with RS, especially if other GRADE domains are similar and/or improved. Again, we have not yet identified compelling examples.

## 5. Summary and next steps

The use of GRADE can guide authors of knowledge syntheses in considering RS and NRS to inform health questions (see table 3). In some situations, authors will decide not to search for NRS to address issues of relative effect, e.g. when the intervention is a well-known treatment and they anticipate identifying large well-conducted randomized trials evaluating its efficacy. Under such circumstances, searching, screening, analyzing, and presenting evidence from NRS unnecessarily adds substantial work. – though it may still be desirable to search of other sorts of NRS that address issues of baseline risk. However, health questions exist and require answers regardless of the current underlying evidence and, thus, authors will have to look for the highest quality evidence available, which when high certainty RS are not available may be NRS to further complement the body of evidence from RS (e.g. when indirect or imprecise) or replace RS if the overall certainty of evidence is higher from NRS than RS. We provide guidance in this article to accomplish this. If authors have initially decided to search for NRS, but then rejected using that entire

body of evidence, they may consider reporting the NRS studies in their documentation of “excluded studies”.<sup>14</sup>

## **5.1. Unresolved issues and next steps**

We have based this guidance on performing scoping reviews, surveying experts to obtain their advice, preferences, rationale and through feedback and refining using qualitative methods during meetings and online discussions with the GRADE NRS project group specifically and the GRADE Working group more broadly. Further research is needed to test the main premise that using a strategy of looking for NRS when reviewers anticipate that RS will yield very low, low, or on rare occasions moderate certainty of evidence, versus the current strategy of looking for only for RS from the outset, will result in important gains in evidence summaries to support decision making. Such research might address the distribution of GRADE certainty of evidence levels in systematic reviews that includes RS and NRS, or which GRADE domains prove to have serious limitations in NRS and RS when reviewer authors consider both bodies of evidence.

As ROBINS-I will be utilized more frequently as the main risk of bias tool in Cochrane and non-Cochrane reviews, we will need to explore and test new GRADE metrics for downgrading and upgrading NRS.

## 5.2. Summary points

- The GRADE approach supports authors in deciding whether to look and integrate NRS with RS in knowledge syntheses about health interventions.
- We suggest not searching for NRS in a knowledge synthesis if authors anticipated that will identify RS that prove to have high certainty evidence. When authors anticipate very low, low or moderate certainty evidence from RS, they should consider also searching for relevant NRS.
- Bodies of evidence from NRS will generally be classified as high certainty only when authors can identify reasons for rating up (typically very large effects and dose response relationships). Without such rating up, high certainty is theoretically possible yet very unlikely to occur.
- NRS that are higher certainty than RS can be used as a single body of evidence, in the same manner as RS that are higher certainty than NRS can be used as a single body of evidence.
- If authors decide to include both bodies of evidence, the preferred strategy is to present both bodies of evidence in adjacent rows of an evidence profile or SoF table, though other options are available.
- We suggest that authors are cautious in pooling RS and NRS, transparently report their reasons for doing so and seek expert methodological advice.

The use of ROBINS-I requires the same caution in this context but may help with the choices for doing so.

--

## **Acknowledgments and conflict of interest**

CCG, NS, RM, JV and HJS have received funding from the Methods Innovation Fund from Cochrane for the development of this guidance. There is no direct financial conflict of interest.

HJS has no direct financial conflict of interest.

Part of this work has been presented in scientific conferences and at GRADE working group meetings and Cochrane symposia.

## References

1. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ*. 2015;350:h870.
2. 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. *Res Synth Methods*. 2013;4(1):49-62.
3. Cuello-Garcia C, Morgan R, Santesso N, Thayer K, Verbeek JH, Brozek J, et al. A scoping review and survey provides the rationale, perceptions, and preferences for the integration of randomized and non-randomized studies in evidence syntheses (in publication). 2017.
4. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*. 2016;353:i2016.
5. Schunemann H, Akl EA, Morgan R, Cuello-Garcia C. GRADE Guidelines 19. How new tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence [in publication]. 2017.
6. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;349:g7647.



7. Editors PM. Best practice in systematic reviews: the importance of protocols and registration. *PLoS Med.* 2011;8(2):e1001009.
8. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *BMJ.* 2016(355):i4919.
9. McKibbin KA, Wilczynski NL, Haynes RB, Hedges T. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. *Health Info Libr J.* 2009;26(3):187-202.
10. Furlan AD, Irvin E, Bombardier C. Limited search strategies were effective in finding relevant nonrandomized studies. *J Clin Epidemiol.* 2006;59(12):1303-1311.
11. Cuello-Garcia C, Morgan R, Santesso N, Thayer K, Verbeek JH, Brozek J, et al. Strategies to optimize use of randomized and non-randomized studies in evidence syntheses using GRADE (in publication). 2017.
12. Huang L, Yin Y, Yang L, Wang C, Li Y, Zhou Z. Comparison of Antibiotic Therapy and Appendectomy for Acute Uncomplicated Appendicitis in Children: A Meta-analysis. *JAMA Pediatr.* 2017;171(5):426-434.
13. 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):158-172.

14. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from <http://www.handbook.cochrane.org/>.

## Tables

Table 1

**Table 1. Evidence profile. Antibiotics compared to appendectomy for children with uncomplicated appendicitis (Huang 2017).**

№ of studies	Quality assessment					№ of patients		Effect		Quality	Importance	
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	antibiotics	appendectomy	Relative (95% CI)			Absolute (95% CI)
Failure rate – RS												
1	randomised trials	not serious	not serious	not serious	very serious <sup>a</sup>	none	2/24 (8.3%)	0/26 (0.0%)	RR 5.40 (0.27 to 108.00)	0 fewer per 1,000 (from 0 fewer to 0 fewer)	⊕⊕⊕⊕ LOW	CRITICAL
								0.5%		22 more per 1,000 (from 4 fewer to 535 more)		
Failure rate – NRS												
4	observational studies	not serious <sup>b</sup>	not serious	not serious	not serious	strong association <sup>c</sup>	14/144 (9.7%)	1/210 (0.5%)	RR 9.43 (2.52 to 35.30)	40 more per 1,000 (from 7 more to 163 more)	⊕⊕⊕⊕ MODERATE	CRITICAL

CI: Confidence interval; RR: Risk ratio

**Explanations**

- a. Only one randomized controlled trial with wide confidence intervals including the thresholds for appreciable benefit and harms.
- b. NRS classified as very serious risk of bias due to lack of randomization (confounding and selection of participants) if ROBINS-I is used.
- c. A relative risk of 9.4 was considered a large increase in the risk of failure rates, including the lower C.I. of 2.52.

Table 2.

№ of studies	Quality assessment							Effect		Quality	Importance	
	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Probiotics	no probiotics	Relative (95% CI)			Absolute (95% CI)
All cause neonatal mortality – Randomized studies												
17	randomised trials	not serious serious <sup>a</sup>	not serious	not serious	not serious <sup>b</sup>	none	118/2635 (4.5%)	181/2668 (6.8%)	RR 0.70 (0.61 to 0.88)	20 fewer per 1,000 (from 8 fewer to 31 fewer)	⊕⊕⊕ HIGH	CRITICAL
All cause neonatal mortality – Non-randomized studies												
11	non-randomized studies	not serious serious <sup>c</sup>	not serious	not serious	not serious	none	358/5126 (7.0%)	372/5642 (6.6%)	RR 0.72 (0.61 to 0.86)	18 fewer per 1,000 (from 9 fewer to 26 fewer)	⊕⊕○○ LOW	CRITICAL
Severe necrotizing enterocolitis (stage II-III) – Randomized studies												
20	randomised trials	not serious	not serious	not serious	not serious	none	68/2761 (2.5%)	159/2768 (5.7%)	RR 0.47 (0.35 to 0.63)	30 fewer per 1,000 (from 21 fewer to 37 fewer)	⊕⊕⊕⊕ HIGH	CRITICAL
Severe necrotizing enterocolitis (stage II-III) – Non-randomized studies												
12	non-randomized studies	not serious	not serious	not serious	not serious	none	169/5144 (3.3%)	325/5656 (5.7%)	RR 0.55 (0.39 to 0.78)	26 fewer per 1,000 (from 13 fewer to 35 fewer)	⊕⊕○○ LOW	CRITICAL
Sepsis – Randomized studies												
19	randomised trials	not serious	serious <sup>d</sup>	not serious	serious <sup>e</sup>	none	391/2662 (14.7%)	434/2676 (16.2%)	RR 0.92 (0.77 to 1.11)	13 fewer per 1,000 (from 18 more to 37 fewer)	⊕⊕○○ LOW	CRITICAL
Sepsis – Non-randomized studies												
7	non-randomized studies	not serious	not serious	not serious	not serious	none	570/3979 (14.3%)	538/2914 (18.5%)	RR 0.86 (0.74 to 1.00)	26 fewer per 1,000 (from 0 fewer to 48 fewer)	⊕⊕○○ LOW	CRITICAL

CI: Confidence interval; RR: Risk ratio

**Explanations**

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process
- b. Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise
- c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.
- d. Statistical heterogeneity of 47% on the I square value.
- e. Confidence intervals still include the plausible harm and benefit thresholds

**Table 3**

GRADE judgements and implications for integration of randomized and non-randomized studies.

<b>GRADE DOMAIN</b>	<b>JUDGEMENT BETWEEN RS AND NRS</b>	<b>IMPLICATIONS FOR INTEGRATION*</b>
<b>Risk of bias</b>	Risk of bias is similar between RS and NRS	<p>Using GRADE, NRS are rated down by two levels due to absence of randomization; if other biases (e.g., missing data, unblinded outcome assessment, etc.) are deemed unlikely, no further downgrading is undertaken. If RS present no concerns about risk of bias they remain at high certainty.</p> <p>In the context of GRADE, confounding bias should prompt rating NRS as serious risk of bias on ROBINS-I (that is rating down by two levels). With GRADE this leads to a rating of ‘very serious’ unless authors have strong justification to not consider risk of bias due to confounding (e.g., in a study with strong interrupted time series design); this is rare because bias due to confounding and other bias such as due to selection of participants are rarely eradicated, even with good adjustment techniques. With ROBINS-I we consider RS and NRS at the same metric, using an “ideal” randomized trial as benchmark, and on occasions, they can have the same risk of bias; for example, if RS are deemed ‘very serious’ due to poor description of the randomization process, and NRS are without concerns of bias other than the confounding and end up as ‘very serious’ too. A similar risk of bias between RS and NRS will make integration into a single pooled estimate more feasible and appropriate (considering other GRADE domains).</p>
	Risk of bias is different between RS and NRS	<p>If RS have less concern of bias than NRS, there will be compelling reasons to use RS only; however, exceptions can occur when other GRADE domains are considered (e.g., indirectness from RS vs direct evidence from NRS) and NRS can still provide complementary evidence.</p> <p>If NRS have less concern of risk of bias than RS, either because RS have very serious risk of bias, or because NRS have good reasons to not rate down (i.e., no reasons to suspect residual confounding), then NRS can be used alone or even pooled with RS if it is considered sensible (e.g., no important differences on other GRADE domains, similar direction/magnitude of effects, etc.)</p>
<b>Inconsistency</b>	Similar concerns of inconsistency	<p>If both RS and NRS have no inconsistency, the integration into a single summary of findings separated in two rows or into a single pooled estimate may be appropriate, although it may be restricted to judgments that influence ratings on other GRADE domains, in particular, indirectness and imprecision.</p> <p>If both RS and NRS have concerns of inconsistency, any form of integration will be less appropriate.</p>

<b>GRADE DOMAIN</b>	<b>JUDGEMENT BETWEEN RS AND NRS</b>	<b>IMPLICATIONS FOR INTEGRATION*</b>
	Different concerns of inconsistency	If the body of evidence from RS and NRS indicates inconsistent results that cannot be explained other than by risk of bias considerations, then RS and NRS should be considered separately. If one body of evidence is clearly leaving us with higher certainty of evidence our answer to a health question will rely on that body of evidence.
<b>Indirectness</b>	Similar concerns of indirectness	If authors have concerns of indirectness from both RS and NRS, they will have to rely on the body of evidence with highest certainty by assessing other GRADE domains.
	Different concerns of indirectness	Direct evidence from NRS can provide equivalent or potentially higher certainty compared to indirect evidence from RS. In such cases, using both or only NRS may be appropriate. If using both, however, the option to integrate into a single pooled estimate will be less appropriate.
<b>Imprecision</b>	Similar concerns of imprecision	If imprecision is the only affected GRADE domain in a body of evidence of RS or NRS, their integration may be feasible and appropriate. If both RS and NRS have imprecise results, the integration will depend mostly on other GRADE domains and in the overall certainty.
	Different concerns of imprecision	Precise results in one body of evidence can complement imprecise results in another and may influence our decision to use one over another. However, even in they differ, it is still feasible to integrate both types of studies if sensible by considering other GRADE domains affected and the overall certainty.
<b>Publication bias</b>	Similar or different concerns of publication bias	Both RS and NRS are prone to this type of bias. Publication bias has less weight on the decision to integrate RS and NRS in any form, and authors should base their choice for integration based on the overall certainty of evidence for each outcome.
<b>Large effects</b>	Only applicable to NRS	A large effect (strong association) can increase the certainty in the body of evidence of NRS. On occasions this will make it more likely to integrate with RS or using only NRS.
<b>Dose-response</b>	Only applicable to NRS	Dose response can increase the certainty in the body of evidence of NRS and the appropriateness of integration with RS, or even the consideration for using only NRS over RS.
<b>Opposing residual confounding</b>	Only applicable to NRS	If opposing plausible residual confounding is suspected, authors can rate up one level the certainty in NRS and apply other GRADE criteria to evaluate the appropriateness of integration with RS. This domain is optionally included on each ROBINS-I item, as an add-on for signaling questions; therefore, authors may evaluate opposing residual confounding in the risk of bias GRADE domain, and not as a stand-alone domain; more testing and empirical observations are needed.

## Figures



Figure 1.

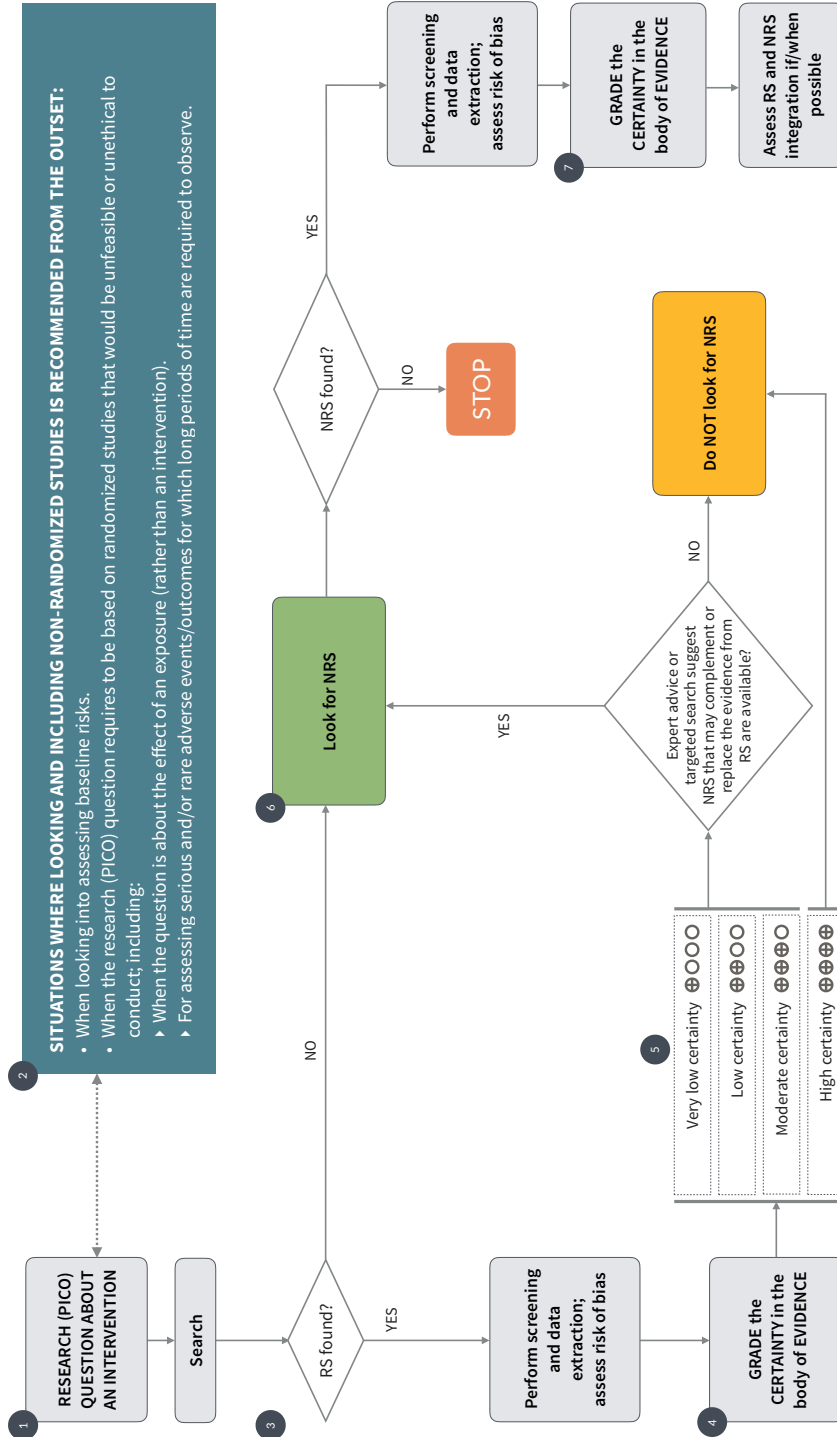


Figure 1. Flowchart regarding the process of conducting a systematic review about a health intervention considering the role of randomized and non-randomized studies. Points 1 and 2 can be addressed from the protocol stage; the rest are in relation to the conduction of the systematic review.

- a. In the protocol stage for the systematic review, state your PICO eligibility criteria, which can apply to either RS or NRS. Consider if you want to search for NRS to inform baseline risk (which will involve separate eligibility criteria).
- b. If you are looking for NRS to inform baseline risks, or if expert advice suggests that any of the points in this box apply, look for NRS from the outset, regardless of your intention to look for RS.
- c. If you have decided to possibly include NRS, a comprehensive search should produce a database of references that include both RS and NRS and filters to differentiate if/when needed. We suggest screening titles and abstracts for RS first and, if found, proceed to full-text screening, data extraction, and assessment of risk of bias. If you find no RS, we suggest seeking NRS unless you have a clear reason for not doing so, which should be declared in the systematic review document.
- d. Assess the GRADE certainty in the evidence of RS by outcome.

- e. With high certainty evidence from RS, there is no reason to search for NRS (unless authors are seeking NRS to inform baseline risk). With very low and low certainty of evidence, authors should evaluate the available NRS for eligibility if experts suggest that informative NRS are available. With moderate certainty of evidence from RS, it is unlikely that authors will find NRS with similar or higher certainty of evidence (i.e., NRS classified as moderate or high), and evaluation of NRS for relative effects should proceed only if there is knowledge that very exceptional NRS are available.
- f. Consider the eligibility criteria of NRS, which should yield evidence of similar certainty to the RS – for example, NRS with adequate sample size that undertook adjustment for key prognostic variables and achieved satisfactory follow-up.
- g. Once NRS are considered applicable for the research question and data extraction is completed, authors should GRADE the certainty of the body of evidence from NRS.

Figure 2

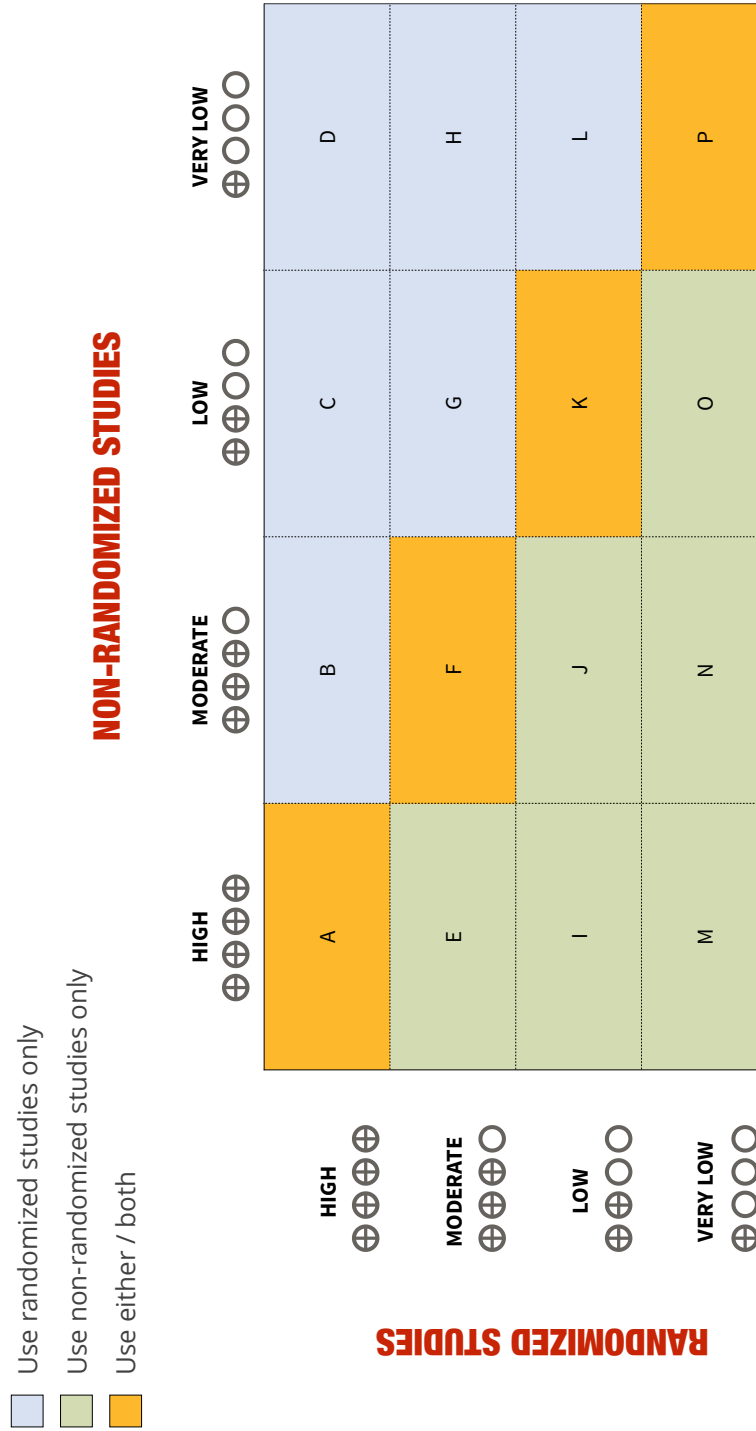


Figure 2. Sixteen possible scenarios to encounter when evaluating bodies of evidence of RS and NRS. See text for full description.

Figure 3

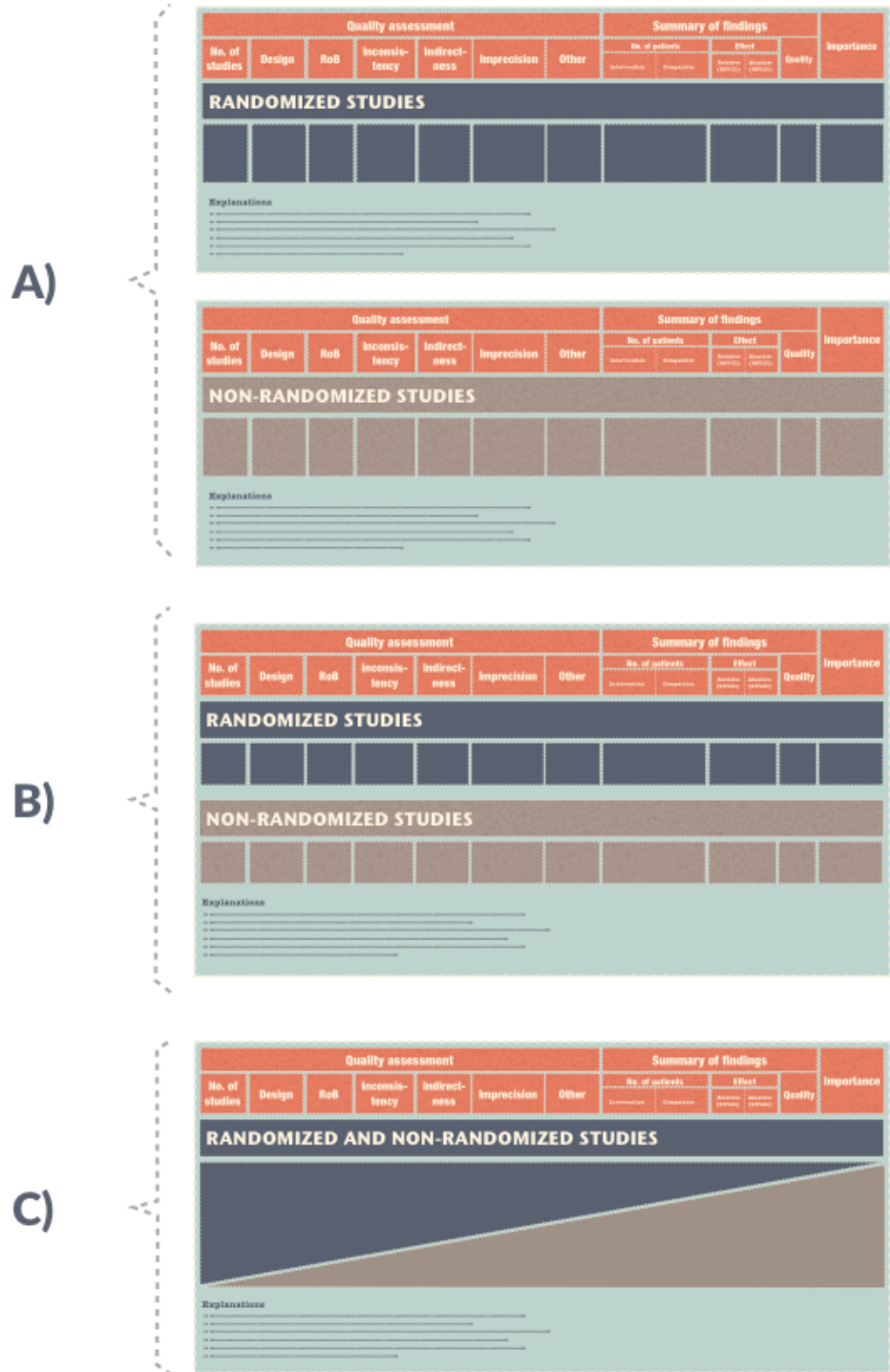


Figure 3. Three possible presentations of both RS and NRS in GRADE evidence profiles. For an example with summary of findings tables see appendices 1, 2, and 3. See also text for full description.

Figure 4

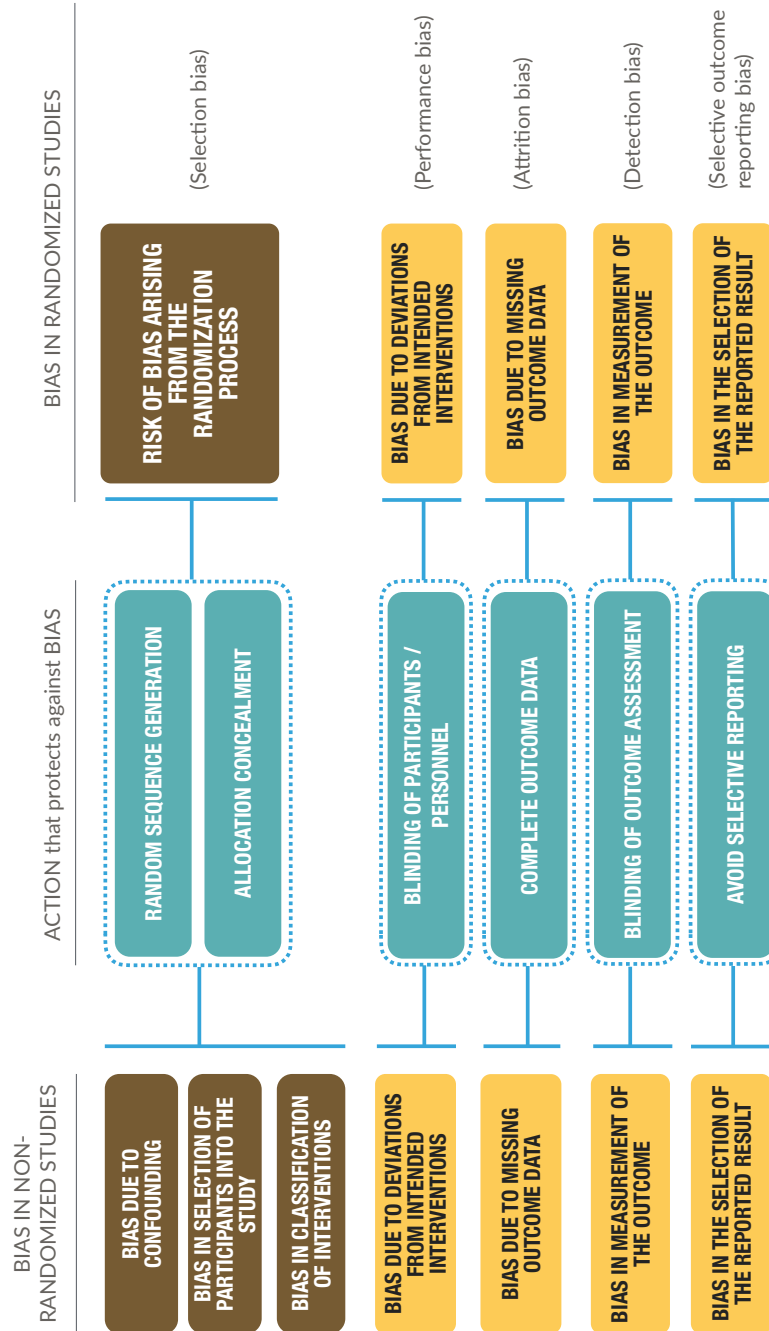




Figure 4. Types of bias met in non-randomized studies (left column) and randomized studies (right column) with the situations or actions performed in a randomized trial that protect against these biases on each type of study (center column). In parentheses are depicted other nomenclatures for biases.

## Appendices

## Supplementary material 1

### Appendix 1a. Example of presentation in summary of findings with RS and NRS still with distinction among the two.

SUMMARY OF FINDINGS TABLE. Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks					
Outcomes	N <sup>o</sup> of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no probiotics	Risk difference with Probiotics
All cause neonatal mortality – Randomized studies (mortality)	5303 (17 RCTs)	⊕⊕⊕⊕ HIGH <sup>a,b</sup>	RR 0.70 (0.55 to 0.88)	68 per 1,000	20 fewer per 1,000 (31 fewer to 8 fewer)
All cause neonatal mortality – Non-randomized studies	10768 (11 non-randomized studies)	⊕⊕○○ LOW <sup>c</sup>	RR 0.72 (0.61 to 0.86)	66 per 1,000	18 fewer per 1,000 (26 fewer to 9 fewer)
Severe necrotizing enterocolitis (stage II-III) – Randomized studies	5529 (20 RCTs)	⊕⊕⊕⊕ HIGH	RR 0.47 (0.35 to 0.63)	57 per 1,000	30 fewer per 1,000 (37 fewer to 21 fewer)
Severe necrotizing enterocolitis (stage II-III) – Non-randomized studies	10800 (12 non-randomized studies)	⊕⊕○○ LOW	RR 0.55 (0.39 to 0.78)	57 per 1,000	26 fewer per 1,000 (35 fewer to 13 fewer)
Sepsis – Randomized studies	5338 (19 RCTs)	⊕⊕○○ LOW <sup>d,e</sup>	RR 0.92 (0.77 to 1.11)	162 per 1,000	13 fewer per 1,000 (37 fewer to 18 more)
Sepsis – Non-randomized studies	6893 (7 non-randomized studies)	⊕⊕○○ LOW	RR 0.86 (0.74 to 1.00)	185 per 1,000	26 fewer per 1,000 (48 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; RR: Risk ratio

#### Explanations

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process  
b. Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise  
c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.  
d. Statistical heterogeneity of 47% on the I square value.  
e. Confidence intervals still include the plausible harm and benefit thresholds

## Appendix 1b. Example of presentation in evidence profile with RS and NRS still with distinction among the two.

**Table 2. Evidence profile.** Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks

№ of studies	Study design	Quality assessment							№ of patients		Effect		Quality	Importance
		Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Probiotics	no probiotics	Relative (95% CI)	Absolute (95% CI)				
All cause neonatal mortality – Randomized studies														
17	randomised trials	not serious <sup>a</sup>	not serious	not serious <sup>b</sup>	not serious	none	118/2635 (4.5%)	181/2668 (6.8%)	RR 0.70 (0.55 to 0.88)	20 fewer per 1,000 (from 8 fewer to 31 fewer)	⊕⊕⊕⊕ HIGH	CRITICAL		
All cause neonatal mortality – Non-randomized studies														
11	non-randomized studies	not serious <sup>c</sup>	not serious	not serious	not serious	none	358/5126 (7.0%)	372/5642 (6.6%)	RR 0.72 (0.61 to 0.86)	18 fewer per 1,000 (from 9 fewer to 26 fewer)	⊕⊕○○ LOW	CRITICAL		
Severe necrotizing enterocolitis (stage II-III) – Randomized studies														
20	randomised trials	not serious	not serious	not serious	not serious	none	68/2761 (2.5%)	159/2768 (5.7%)	RR 0.47 (0.35 to 0.63)	30 fewer per 1,000 (from 21 fewer to 37 fewer)	⊕⊕⊕⊕ HIGH	CRITICAL		
Severe necrotizing enterocolitis (stage II-III) – Non-randomized studies														
12	non-randomized studies	not serious	not serious	not serious	not serious	none	169/5144 (3.3%)	325/5656 (5.7%)	RR 0.55 (0.39 to 0.78)	26 fewer per 1,000 (from 13 fewer to 35 fewer)	⊕⊕○○ LOW	CRITICAL		
Sepsis – Randomized studies														
19	randomised trials	not serious	serious <sup>d</sup>	not serious	serious <sup>e</sup>	none	391/2662 (14.7%)	434/2676 (16.2%)	RR 0.92 (0.77 to 1.11)	13 fewer per 1,000 (from 18 more to 37 fewer)	⊕○○○ LOW	CRITICAL		
Sepsis – Non-randomized studies														
7	non-randomized studies	not serious	not serious	not serious	not serious	none	570/3979 (14.3%)	538/2914 (18.5%)	RR 0.86 (0.74 to 1.00)	26 fewer per 1,000 (from 0 fewer to 48 fewer)	⊕○○○ LOW	CRITICAL		

CI: Confidence interval; RR: Risk ratio

### Explanations

- 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process
- Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise
- All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.
- Statistical heterogeneity of 47% on the I square value.
- Confidence intervals still include the plausible harm and benefit thresholds

## Supplementary material 2

### Appendix 2. Portrayal of randomized and non-randomized studies separated in two different GRADE summary of findings tables.

**SUMMARY OF FINDINGS TABLE 1. Randomized studies. Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks**

Outcomes	Nº of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no probiotics	Risk difference with Probiotics
All cause neonatal mortality – Randomized studies (mortality)	5303 (17 RCTs)	⊕⊕⊕⊕ HIGH <sup>a,b</sup>	RR 0.70 (0.55 to 0.88)	68 per 1,000	20 fewer per 1,000 (31 fewer to 8 fewer)
Severe necrotizing enterocolitis (stage II-III) – Randomized studies	5529 (20 RCTs)	⊕⊕⊕⊕ HIGH	RR 0.47 (0.35 to 0.63)	57 per 1,000	30 fewer per 1,000 (37 fewer to 21 fewer)
Sepsis – Randomized studies	5338 (19 RCTs)	⊕⊕○○ LOW <sup>c,d</sup>	RR 0.92 (0.77 to 1.11)	162 per 1,000	13 fewer per 1,000 (37 fewer to 18 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

#### Explanations

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process  
 b. Considering a reduction of at least of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise  
 c. Statistical heterogeneity of 47% on the I square value.  
 d. Confidence intervals still include the plausible harm and benefit thresholds

**SUMMARY OF FINDINGS TABLE 2. Non-randomized studies. Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks**

Outcomes	Nº of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no probiotics	Risk difference with Probiotics
All cause neonatal mortality – Non-randomized studies	10,768 (11 non-randomized studies)	⊕⊕○○ LOW <sup>a</sup>	RR 0.72 (0.61 to 0.86)	66 per 1,000	18 fewer per 1,000 (26 fewer to 9 fewer)
Severe necrotizing enterocolitis (stage II-III) – Non-randomized studies	10,800 (12 non-randomized studies)	⊕⊕○○ LOW	RR 0.55 (0.39 to 0.78)	57 per 1,000	26 fewer per 1,000 (35 fewer to 13 fewer)
Sepsis – Non-randomized studies	6,893 (7 non-randomized studies)	⊕⊕○○ LOW	RR 0.86 (0.74 to 1.00)	185 per 1,000	26 fewer per 1,000 (48 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; RR: Risk ratio

#### Explanations

- a. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.

## Supplementary material 3

### Appendix 3. Portrayal of randomized and non-randomized studies separated in a single summary of findings table with a single pooled estimate for both.

**SUMMARY OF FINDINGS TABLE. Probiotics compared to no probiotics for premature newborns less than 1500 grams and/or less than 34 weeks**

Outcomes	N <sup>a</sup> of participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects	
				Risk with no probiotics	Risk difference with Probiotics
All cause neonatal mortality – RS+NRS	13,442 (28 Studies) <sup>d</sup>	⊕⊕⊕⊕ HIGH <sup>ab,cd</sup>	<b>RR 0.72</b> (0.63 to 0.81)	86 per 1,000	<b>24 fewer per 1,000</b> (32 fewer to 16 fewer)
Severe necrotizing enterocolitis (stage II-III) – RS+NRS	16,329 (32 Studies) <sup>d</sup>	⊕⊕⊕⊕ HIGH	<b>RR 0.52</b> (0.42 to 0.65)	57 per 1,000	<b>28 fewer per 1,000</b> (33 fewer to 20 fewer)
Sepsis – RS+NRS	12,231 (26 Studies)	⊕⊕⊕○ MODERATE <sup>e</sup>	<b>RR 0.90</b> (0.79 to 1.01)	174 per 1,000	<b>17 fewer per 1,000</b> (37 fewer to 2 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; **RS:** Randomized studies; **NRS:** Non-randomized studies.

#### Explanations

- a. 5 studies with unclear (no adequate description) of the random sequence generation and seven with no adequate description of the allocation concealment process
- b. Considering a reduction of at best of 8 fewer deaths (per 1000 treated children) as clinically important or not this might be considered imprecise
- c. All studies were retrospective cohorts with historical controls with one arm where all patients received probiotics routinely while the historic control did not. Most studies used adequate methods to adjust for baseline confounding when suspected, except for two studies that were classified as serious risk of bias (ROBINS-I) for not using adequate strategies to adjust baseline confounding domains and variables. Residual confounding was considered unlikely in nine of the studies.
- d. Includes both randomized and non-randomized studies
- e. Statistical heterogeneity of 47% on the I square value.

## CHAPTER 5. CONCLUSIONS

---

## Summary of findings

This work presents three main pieces of research and analyses. Through these, the main findings can be summarized as follows:

- a. New methods in the field of observational studies have emerged generating new opportunities to use NRS with RS in knowledge syntheses of health interventions.
- b. Experts in knowledge syntheses are willing to use NRS with RS when facing a research question about a health care intervention, either in a single synthesis, in a single summary of findings, or in a single pooled estimate. In fact, many experts already do integrate these two types of studies, although, on most occasions, with rather different methods and without specific guidance.
- c. To evaluate the appropriateness of their integration, it is fundamental to consider the certainty of the evidence (per outcome) of both RS and NRS, and not just the risk of bias.
- d. If integration of both type of studies in a single summary of findings is deemed appropriate, most experts prefer to draw a distinction between RS and NRS (i.e., separated in two rows in GRADE summary of findings tables). However,



in special circumstances pooling both designs may be appropriate; for example, if both have same direction of effects, and direct evidence from NRS do not suffer from additional risk of bias and RS are deemed not to have higher certainty. More testing and more examples are needed, as well as an appropriate statistical assessment (see future research direction and needs below).

- e. We created guidance for authors of knowledge syntheses who wish to use RS and NRS in knowledge syntheses. In this, we discuss how GRADE can help assessing the appropriateness of integrating both bodies of evidence in different ways. Also, we provide insights about the ROBINS-I tool for assessing the risk of bias in NRS, highlighting the opportunities that this novel instrument represents due to its assessment of both types of evidence in the same absolute scale of risk of bias.

## **Implications for researchers, guideline developers, clinicians, patients, and policy-makers**

The implications for patients, researchers, guideline panelists, practitioners, and policy-makers, arise from the opportunity created to increase the certainty of the evidence. Systematic reviewers will benefit from the framework presented in chapter

4, that assess the pertinence and role of NRS in knowledge syntheses. The framework starts at the protocol stage of a systematic review, evaluating the appropriateness of incorporating NRS with RS.

With current concerns about over-diagnosis, misguided treatments, and research waste,<sup>1, 2</sup> following guidance that helps incorporating RS and NRS in knowledge syntheses can increase comprehensiveness and completeness on the topic or research question, helping researchers widen their field of studies to be included and perhaps to reach earlier to the point of “no further research is needed” earlier.

Currently, many guideline developers include NRS, mostly because they provide vital information such as baseline risks, adverse events, or rare outcomes. The application of the information and advice presented here, will help guideline panelists increase their evidence base, gain certainty, and reach an adequate recommendation. In consequence, clinicians will have a more sensible and complete synthesis to use and share with their patients to help them understand their treatment options the consequences of their decisions.

The implications for policy-makers stem from the fact that health policies are increasingly being based on comparative (or relative) effectiveness of interventions to inform decisions, and NRS are a key study design used for comparative effectiveness research (CER) because they are conducted in what is considered ‘real

world' settings.<sup>3</sup> Incorporating both RS and NRS in knowledge syntheses will increase the armamentarium for policy-makers to make informed health-policy decisions.

## **Strengths and challenges of this work**

This work is the first to review and analyze the preferences and practices of experts on how to integrate RS and NRS in knowledge syntheses. It is also the first work to explore the integration of RS and NRS by considering their similarities and dissimilarities within the GRADE domains and how these differences will affect the appropriateness of integrating RS and NRS. Its strength relies on the transparency and structure that the GRADE approach provides.

Previous guidance<sup>4-6</sup> related to the integration of both bodies of evidence has approached the issue by focusing on the risk of bias, study design, and the possibility to find RS related to the PICO question. The approach here presented evaluates also the differences in the certainty in the evidence between RS and NRS, acknowledging that RS and NRS' main differences might rely on domains other than risk of bias.

This work also evaluates and provides guidance for the conduction of an adequate systematic review that considers incorporating RS and NRS. This includes a framework (as described in chapter 4) suggesting a search strategy (with support from

an information specialist) to tackle the issue of increased workload when NRS are included in a review. This could be viewed as guidance to improve efficiency in the conduct of a knowledge synthesis.

There are also several challenges to address. We understand that evaluating the appropriateness and integrating NRS will require effort and will be time consuming. However, we believe that the guidance for making choices when to search for NRS and the comprehensiveness and the increased certainty are worth the effort.

There can be errors or misapplications from users who do not follow guidance (presented in chapter 4) properly. For example, by using GRADE and ROBINS-I, users can erroneously rate the certainty of the evidence from NRS higher than really needed. We acknowledge the criticism from authors that by using GRADE, no observational study will ever obtain a ‘high’ (or even ‘moderate’) certainty rating. GRADE highlights that one may still rate up certainty for large effects, a dose-response gradient, or if all plausible biases will strengthen rather than undermine inferences from study results.<sup>7</sup>

One common criticism in the field of research syntheses, and biomedical research itself, is the poor quality of individual studies that feed any synthesis of RS and NRS. This important issue is not unique to our topic of integrating evidence, but to all knowledge syntheses, and it must be emphasized. There is also the concern that

researchers, after evaluating both RS and NRS, could decide to include one body of evidence over another based on the results that would be more suitable for their interests. Selecting outcomes and results also goes beyond the integration of RS and NRS. Transparency is key to avoid selecting evidence inappropriately and a priori documented decisions criteria are a way to protect against this.

## **Further research directions**

The first documented clinical trial was conducted 270 years ago, the first randomized controlled trial 70 years ago, and methods for systematic reviews in health sciences have been around for more than four decades. Yet, the appropriate methods to incorporate RS and NRS in knowledge syntheses are still in their initial stages. Several areas of opportunity remain worth exploring:

There have been methodological overviews assessing how systematic reviews incorporate NRS.<sup>5, 8</sup> It would be worth adding to this knowledge an analysis of the differences on each GRADE domain between bodies of evidence of RS and NRS; this could help elucidate why these differences occur beyond randomization, where are the main problems, what do these differences imply when NRS and RS present with differences in the magnitude and direction of effect estimates, and possibly provide more insights for appropriate integration.

Further testing of our suggested approaches to integrate RS and NRS is necessary, starting with an assessment of different search strategies that could provide the most sensible and specific strategy for looking for both RS and NRS, and some testing on the management of the search results, including timing and effort during the screening and data extraction processes.

More testing on the use of the ROBINS-I tool will be necessary in the context of integrating NRS with RS using GRADE. For example, more detailed guidance will be needed for the GRADE domain plausible residual confounding as to when one would not rate down for confounding bias by two levels when using ROBINS-I.

Examples of bodies of evidence from NRS that would not be rated down for confounding bias when using ROBINS-I, such as strong interrupted time series or other non-randomized designs are lacking. Furthermore, when it comes to research questions for which RS and NRS evidence is available, there is currently no real-life example where both (or at least the NRS) are classified as ‘high’ certainty. The reason for this could be that such a situation would reach unethical grounds; for instance, if a body of evidence from RS is considered ‘high’ certainty, there would be no reason to conduct a NRS for the same outcomes and *vice versa*.

Another question that remains, is whether similar RS and NRS be pooled in a single meta-analysis? On theoretical and practical grounds, there are no barriers to

proceed if the certainty for the two bodies of evidence is judged similarly (e.g. both end up as low certainty) but more examples are required.

## **Final remarks**

NRS can provide valuable information for knowledge syntheses of interventions when being used as sequential, supplemental, or replacement of RS.<sup>9</sup> Clinicians, researchers, and policy-makers can benefit from the guidance provided here to integrate RS and NRS in any type of knowledge synthesis about an intervention.

Better health outcomes can be achieved with the help of better information obtained through high quality research syntheses. This dissertation represents an effort for increasing the quality of research syntheses of interventions by attaining comprehensiveness of the evidence about interventions so all stake-holders can reach decisions with higher confidence.

## References

1. Altman DG. The scandal of poor medical research. *BMJ*. 1994;308(6924):283-284.
2. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276.
3. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report. *Value Health*. 2012;15(2):217-230.
4. Reeves BC, Higgins JP, Ramsay C, Shea B, Tugwell P, Wells GA. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Res Synth Methods*. 2013;4(1):1-11.
5. Peinemann F, Kleijnen J. Development of an algorithm to provide awareness in choosing study designs for inclusion in systematic reviews of healthcare interventions: a method study. *BMJ Open*. 2015;5(8):e007540.
6. Norris S, Atkins D, Bruening W, Fox S, Johnson E, Kane R, et al. Selecting Observational Studies for Comparing Medical Interventions. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville (MD)2008.
7. Schunemann H, Akl EA, Morgan R, Cuello-Garcia C. GRADE Guidelines 19. How new tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence [in publication]. 2017.



8. Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. *J Clin Epidemiol.* 2014;67(6):645-653.
  
9. 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. *Res Synth Methods.* 2013;4(1):49-62.