THE ROLE OF PRAGMATISM IN EXPLAINING HETEROGENEITY IN META-ANALYSES

THE ROLE OF PRAGMATISM IN EXPLAINING HETEROGENEITY IN META-ANALYSES OF RANDOMIZED TRIALS: A METHODOLOGICAL REVIEW

By THERESA AVES, BSC (HONS)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for a Master of Science Degree

McMaster University MASTER OF SCIENCE (2017) Hamilton, Ontario (Health Research Methodology)

TITLE: The Role of Pragmatism in Explaining Heterogeneity in Meta-Analyses of Randomized Trials: A Methodological Review

AUTHOR: Theresa Aves, BSc (Western University)

SUPERVISOR: Dr. Lawrence Mbuagbaw

NUMBER OF PAGES: x, 90

Lay Abstract:

Systematic reviews and meta-analyses of randomized controlled trials (RCTs) are an important scientific activity that can lead to changes in health care. However, there is concern whether it is appropriate to meta-analyze data from RCTs that are performed under more controlled conditions (explanatory RCTs) and RCTs that are performed under more real world conditions (pragmatic RCTs) since there may be variability between them. The purpose of this research was to explore how much these trial types affect variability, otherwise known as heterogeneity, in systematic reviews. We applied a scoring tool called the Pragmatic-Explanatory Continuum Indicator Summary-2 (PRECIS-2) to RCTs within 10 systematic reviews with at least moderate heterogeneity and performed statistical modelling to determine how much heterogeneity could be explained by a trial being more or less pragmatic. Results showed that trial type did not explain heterogeneity therefore it is probably reasonable to meta-analyze data from pragmatic and explanatory RCTs.

Abstract:

Introduction
There has been increasing interest in evidence from pragmatic trials as healthcare providers and decision makers must determine if available evidence can be translated and used in real world practice. As a result, a number of tools have been developed to help researchers design and appraise randomized controlled trials (RCTs) within the pragmatic-explanatory continuum. It is unclear what role pragmatism plays in heterogeneity and if pragmatic and explanatory trials should be pooled in meta-analyses of systematic reviews.

Objectives
Our primary objective was to explore the role of pragmatism (based on the Pragmatic-Explanatory Continuum Indicator Summary-2 [PRECIS-2] score) as a source of heterogeneity in Cochrane systematic reviews with at least substantial heterogeneity ($I^2 \geq 50\%$). Our secondary objective was to compare and contrast the application of the established PRECIS-2 tool to the newly developed Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES) tool.

Methods
We conducted a cross-sectional methodological review on systematic reviews of RCTs published in the Cochrane Library from January 1, 2014 to January 1, 2017. Included systematic reviews had a minimum of 10 RCTs in the meta-analysis of the primary outcome and at least moderate heterogeneity ($I^2 \geq 50\%$). Of the eligible systematic reviews, a random selection of 10 were included for quantitative evaluation. In each systematic review, RCTs were scored using the PRECIS-2 and RITES tools, in duplicate, to determine the amount of pragmatism. Meta-regression modelling was performed to evaluate how much variability in heterogeneity (quantified by I^2) was due to pragmatism. Inter-rater reliability of both PRECIS-2 and RITES was measured using the intraclass correlation coefficient and Spearman's correlation coefficient was used to determine the strength of the relation between PRECIS-2 and RITES.

Results
Ten systematic reviews from nine Cochrane Review Groups were included in the quantitative analysis. The reviews included an average of 13 RCTs (standard deviation=2.6) for a total of 132 RCTs of which 128 could be obtained. When the PRECIS-2 summary score was entered as a covariate in random effects meta-regression models for each systematic review, there were minimal changes in heterogeneity. The changes in I^2 ranged from 0.2% to 13.3%.

Conclusion
Based on these findings it appears pragmatism as measured by PRECIS-2 does not explain heterogeneity in systematic reviews, therefore pooling of pragmatic and explanatory RCTs is unlikely to be detrimental to meta-analyses.

Acknowledgements:

Many thanks to my supportive, encouraging, and patient supervisor- Dr. Lawrence Mbuagbaw. You have been a clear source of guidance in my academic endeavours and have provided me with many opportunities and lessons in which I will always be grateful. I have learned so much from your mentorship and am so honoured to be your inaugural graduate student.

Thank you to my committee members, Dr. Robby Nieuwlaat, Dr. Joseph Beyene and Dr. Elizabeth Alvarez. Your input and feedback have been valuable contributions to this research and I am so appreciative of your guidance and expertise.

My gratitude to Katherine Allan and Daeria Lawson. You both were an integral part of this research and I would not have been able to complete this project without you. You were there for me during the most critical moments, unwavering in your support and ability to score and extract data.

Special thanks to my wonderful family, particularly my parents who have always been steadfast in supporting my goals both academic and otherwise. You have always told me that I could do anything I put my mind to. Thank you for your constant encouragement starting from when I was young and continuing on today. Thank you to my siblings, for being my biggest cheerleaders, continually showing me support, and lending me an ear when the journey was sometimes challenging.

Thank you to the faculty and students of the HRM program. You have provided me with a constant reminder of how much I enjoy learning and scientific research. You have challenged me to think outside the box, to ask questions, and to be a better researcher.

Most importantly, a big thank you to my partner. Your support has led me to where I am today. I consider myself very lucky to have gone through this journey with you by my side. Thank you for always being there for me and for always believing in me.

Table of Contents:

List of Figures and Tables:

List of Figures:

List of Tables:

List of Abbreviations:

Abbreviation	Definition
ASPECT-R	A Study Pragmatic-Explanatory Characterization Tool-Rating
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
GEECT	Grading of Efficacy-Effectiveness in Clinical Trials
HR	Hazard Ratio
ICC	Intraclass Correlation Coefficient
MD	Mean Difference
OR	Odds Ratio
PRECIS	Pragmatic-Explanatory Continuum Indicator Summary
PRECIS-2	Pragmatic-Explanatory Continuum Indicator Summary-2
PR-Tool	Pragmatic-Explanatory Continuum Indicator Summary Review Tool
RCT	Randomized Controlled Trial
RR	Risk Ratio
ROB	Risk of Bias
RITES	Rating of Included Trials on the Efficacy-Effectiveness Spectrum
SMD	Standardized Mean Difference
T1, T2, T3	Tertile 1, Tertile 2, Tertile 3

Declaration of Academic Achievement:

The original idea for this research was conceptualized by Dr. Lawrence Mbuagbaw who along with myself, designed and developed the project. Scoring of PRECIS-2, RITES, and data abstraction was carried out by myself, Katherine Allan and Daeria Lawson. I performed statistical analyses based on guidance from Dr. Mbuagbaw and Behnam Sadeghirad. I wrote the first draft of this thesis and received subsequent feedback and suggestions from Dr. Mbuagbaw and my committee members (Dr. Robby Nieuwlaat, Dr. Joseph Beyene, and Dr. Elizabeth Alvarez). All individuals as mentioned above will be included as co-authors on the published manuscript(s) resulting from this work.

Chapter 1: Background and Objectives

1.1 Pragmatic and Explanatory Trials

In clinical research, randomized trials are often categorized as either pragmatic or explanatory.[1] In broad terms, pragmatic trials are designed to determine the effects of an intervention under the usual or real world conditions in which it will be applied whereas explanatory trials are designed to determine the effects of an intervention under ideal or controlled circumstances.[2] The distinction between pragmatic and explanatory trials was first introduced by Schwartz and Lellouch nearly a half century ago.[3] In their seminal article, they described differing approaches to pragmatic and explanatory trials with the former aimed at clinical *decision* making and the latter aimed at u*nderstanding* treatment effects.[3] The authors suggested that most trials adopt an explanatory approach without question even though a pragmatic approach may be more justifiable,[3] resulting in trial methodology that may not be reflective of its intended purpose.

Interest in the design of pragmatic trials has become increasingly widespread in the scientific community as health care providers and decision makers seek to maximize the applicability of clinical trial results to usual practice.[4] In 2008, an extension to the Consolidated Standards of Reporting Trials (CONSORT) statement was published aimed at improving the reporting of pragmatic trials.[5] The extension provides specific guidance for 8 of the 22 standard CONSORT checklist items including those pertaining to scientific background and rationale, eligibility criteria, details of the intervention(s), outcome measurement, sample size, blinding, flow of study participants through the trial, and generalizability of the findings.[5] With steady growth in the use of pragmatic trial designs[6], the CONSORT extension is a valuable tool to promote transparency and consistency in communication of pragmatic trial research findings.

There are key design features where pragmatic and explanatory trials differentiate. A highly pragmatic approach would be one that considers the setting, participants, and intervention with the same flexibility as would occur in usual care.[5] The trial outcome(s) would be directly relevant to patients, health care providers, decision makers, and clinical practice.[5] Conversely, a highly explanatory approach would be one that highly selects the setting and participants, with an intervention that is closely monitored.[5] The trial outcome(s) may be short-term surrogate measures with indirect relevance to clinical practice.[5] However, the distinction between an pragmatic and explanatory trial in real life is not so simple nor clear cut.[6] To remedy this, several tools have been developed to aid researchers in characterizing and designing pragmatic trials.

1.2 Tools for Characterizing and Designing Pragmatic Trials

Over the past 10 years, a number of tools have been developed to assess the degree of pragmatism in randomized controlled trials (RCTs; Table 1). In 2006, Gartlehner et al. published a tool to distinguish pragmatic from explanatory trials in an effort to provide authors of systematic reviews a means to quantify generalizability of included studies.[7] Thorpe et al. followed this with the Pragmatic-Explanatory Continuum Indicator Summary (PRECIS) tool which was developed to inform study design rather than a

means of classifying trials within systematic reviews.[1] The PRECIS tool has 10 domains which include key trial design considerations such as participant eligibility, interventions and expertise, follow-up and outcomes, compliance/adherence and analysis.[1] The authors discussed the use of a pragmatic-explanatory continuum rather than a dichotomy as Gartlehner et al. had proposed[7] and as such, a formal scoring system was not developed.[1]

Shortly thereafter, Tosh et al. utilized the PRECIS framework to develop the Pragmascope tool, which was designed assess the applicability of RCT results, according to what was planned at the protocol stage.[8] Unlike PRECIS, the Pragmascope had a formal scoring system where each of the 10 PRECIS domains were rated from 1=most explanatory to 5=most pragmatic.[8] Similar to the Pragmascope, El Dib et al. used the PRECIS domains to develop the Grading of Efficacy-Effectiveness in Clinical Trials (GEECT) tool where they scored RCTs rather than protocols on a scale of 0 to 10 for each PRECIS domain where 0=high efficacy and 10=high effectiveness.[9] Recently, Bossie et al. developed an adapted version of the PRECIS tool called A Study Pragmatic-Explanatory Characterization Tool-Rating (ASPECT-R) and reduced the number of domains from 10 to 6, due to perceived redundancy.[10] The domains included participant eligibility, flexibility, setting, follow-up intensity, primary outcomes, and participant compliance.[10] Like the Pragmascope and GEECT tools, the ASPECT-R tool provided a distinct scoring system where each domain received a score from 0=extremely explanatory to 6=extremely pragmatic.[10]

In addition to application retrospectively at the protocol and main publication stages, PRECIS has been applied a number of times in the systematic review setting in an effort to quantify how pragmatic primary RCTs and systematic reviews are.[11 12] This quantification may provide additional guidance for healthcare providers and decision makers regarding the applicability of the RCTs and systematic reviews in routine practice.[11] In cases where PRECIS was applied to systematic reviews, a scoring system was utilized which ranged from either 0 to 4, or 1 to 5 with the lowest number representing a more explanatory RCT or review and the highest number representing a more pragmatic RCT or review.[11 12]

Koppenaal et al. applied a modified version of PRECIS which they called the PRECIS Review tool (PR-tool) to two systematic reviews of primary care interventions.[11] The authors discussed noteworthy observations such as the assumption of equal weighting across the 10 domains and that PRECIS-2 cannot always provide an assessment of pragmatism that is applicable to multiple settings such as different countries or types of healthcare services.[11] Yoong et al. applied an adapted version of PRECIS to a systematic review of interventions for preventing obesity in children.[12] Independent raters gave scores of 0 to 4 for each PRECIS domain within a primary RCT.[12] The authors developed cut offs to classify primary RCTs as predominantly explanatory (0 to 1.7), combined explanatory/pragmatic (>1.7 to ≤2.2) and mostly pragmatic (>2.2 to 4).[12] They explored the impact of study classification on intervention effect sizes by age group (0-5 years, 6-12 years and 13-18 years), and found that pragmatic trials had the smallest effect sizes compared to explanatory trials.[12] However, the authors stopped

short of exploring the effect of pragmatism on heterogeneity (I^2) which was substantial among each age group and overall (I^2=79%).[13] Yoong et al. suggested reporting the results of PRECIS with other subgroup analyses in systematic reviews and discussed the need to further explore the impact of pragmatism across a broad range of systematic review topics and large number of trials.[12]

While Koppenaal et al. and Yoong et al. applied modified versions of PRECIS to previously conducted systematic reviews, Witt et al. conducted a systematic analysis in trials of acupuncture for lower back pain with the intention of applying the PRECIS tool.[14] The authors used a similar scoring system as Koppenaal et al. which was performed independently by five raters followed by consensus discussions to resolve disagreements. The authors discussed missing information as a limitation of applying PRECIS which appeared as such in both Koppenaal et al. and Yoong et al.[11 12 14] Nonetheless, each research group acknowledged that the modification of PRECIS was useful and may provide important insight regarding the quantification of pragmatism at both the RCT and systematic review level.[11 12 14] Among each of the tools, inter-rater reliability was diverse across domains and summary scores from poor to almost perfect. Agreement may have been affected by missing data, differential clinical expertise, and/or difficulty applying definitions of the domains.

In 2015, a revised version of the PRECIS tool was published by Loudon et al. called PRECIS-2 which addressed the weaknesses of the original tool such as unclear inter-rater reliability, lack of a scoring system and redundancy in some PRECIS domains.[5] Currently, there are 9 domains in the PRECIS-2 tool including eligibility, recruitment, setting, organization, flexibility, follow-up, primary outcome, and primary analysis.[15] Each domain is scored using a 5 point Likert scale where 1=a very explanatory trial and 5=a very pragmatic trial.[15] Scores from each domain may be graphically displayed using the PRECIS-2 wheel where points closer to the center of the wheel depict a more explanatory trial and points at the outer area of the wheel depict a more pragmatic trial.[15] Since studies are rarely entirely pragmatic or explanatory, one domain may be more or less pragmatic than another (Figure 1).[1] While the tool is intended to be used at the design stage of a trial, the authors believe PRECIS-2 may have a role in critical appraisal and systematic reviews.[15]

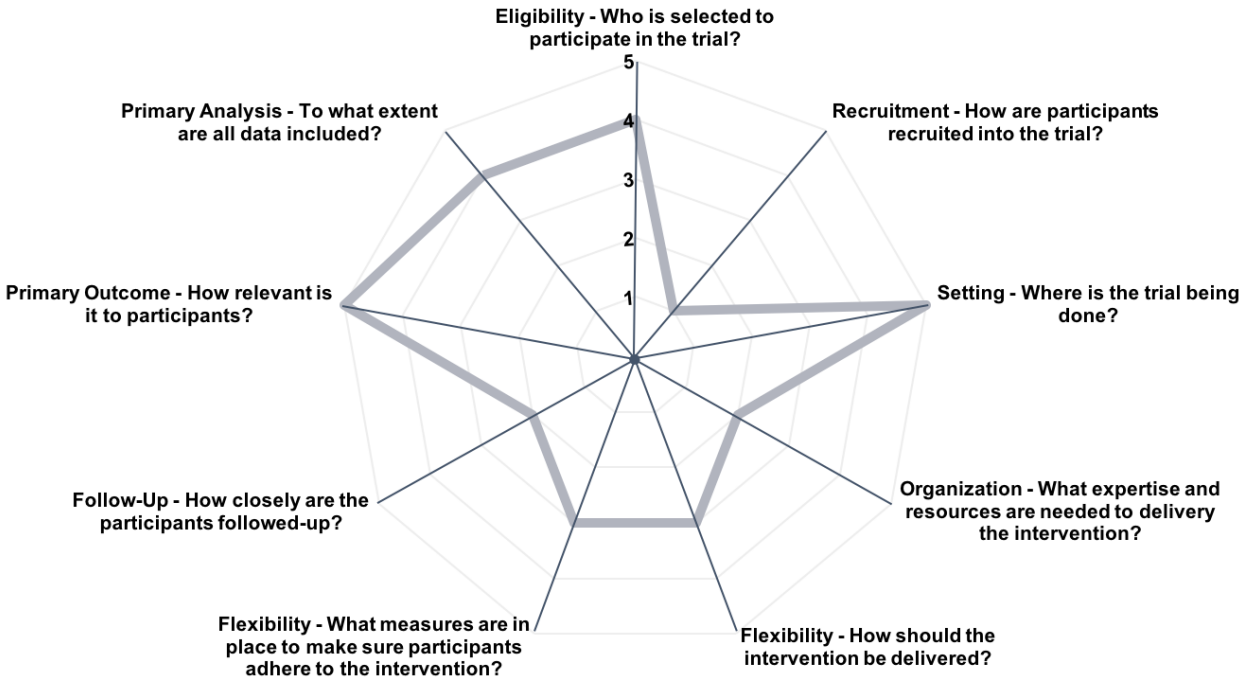


Figure 1. An example PRECIS-2 wheel showing domains that are more and less pragmatic.

More recently, Loudon et al. undertook an in depth assessment of inter-rater reliability and discriminant validity of the PRECIS-2 tool.[16] Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) which ranged from 0.24 to 0.94 indicating diverse agreement. [16] Further assessment of inter-rater reliability may be beneficial, particularly with the use of main trial publications. Additional assessment would provide inter-rater reliability information for a systematic review setting and complementary information for how PRECIS-2 could be applied when RCT protocols are not available.

Following the development and publication of this research protocol, we discovered a new tool entitled Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES). The objective of RITES is to retrospectively characterize primary RCTs in systematic reviews on the efficacy-effectiveness spectrum.[17] With RITES, the authors use the terms efficacy and effectiveness to address RCT evidence instead of the terms pragmatic and explanatory which are used to address trials and their design.[17] The RITES tool has 4 domains and focuses on essential elements of the efficacy-effectiveness spectrum that are likely to be reported in the RCTs of systematic reviews such as participants characteristics, trial setting, flexibility of intervention(s), and clinical relevance of experimental and comparison intervention(s).[17] Domains are scored using a 5 point Likert scale where 1=strong emphasis on efficacy and 5=strong emphasis on effectiveness.[17] The tool was pilot tested among 12 researchers using three small Cochrane systematic reviews to assess feasibility and inter-rater reliability.[17] Results of the pilot testing showed a wide variation in inter-rater reliability with ICCs ranging from 0.23 to 0.45.[17] The developers are yet to provide guidance for how to incorporate RITES into the conduct and reporting of systematic reviews, and further work is

necessary to determine how RITES may be of practical use to users of systematic reviews.[17] Although only preliminary work has been conducted using RITES, we have included it in this review to ensure a robust assessment of pragmatism across a subset of systematic reviews and to compare the newly developed RITES tool with the current PRECIS-2 tool.

Table 1. Summary of tools available for the design and characterization of pragmatic trials

Primary Author, Year	Name of Tool	Purpose of Use	Domains (Number)	Scoring System	Inter-Rater Reliability
Gartlehner G, 2006	Simple tool to distinguish efficacy from effectiveness studies	Characterization of RCTs	Population, Eligibility, Outcomes, Study Duration, Adverse Events, Sample Size, Analysis (7)	Dichotomous (yes or no)	Overall 0.42 (Kappa)
Thorpe K, 2009	Pragmatic-Explanatory Continuum Indicator Summary (PRECIS)	RCT design	Eligibility, Interventions and Expertise, Follow-Up and Outcomes, Compliance/Adherence, Analysis (10)	None	Not done
Koppenaal T, 2011	PRECIS Review Tool (PR-tool)	Characterization of systematic reviews	Same as PRECIS (10)	1 to 5 where 1=extreme explanatory study; 5=extreme pragmatic study	Not done
Tosh G, 2011	Pragmascope	Characterization of RCT protocols	Same as PRECIS (10)	1 to 5 where 1=most explanatory; 5=most pragmatic	Overall 0.72 (Weighted Kappa)
Witt C, 2012	Adapted PRECIS	Systematic analysis of RCTs	Same as PRECIS (10)	1 to 5 where 1=maximal efficacy; 5=maximal effectiveness	0.02 to 0.60 across domains (ICC)
Yoong S, 2014	Adapted PRECIS	Characterization of systematic reviews	Same as PRECIS (10)	0 to 4 where 0=completely explanatory; 4=completely pragmatic	0.23 to 0.75 across domains (Weighted Kappa)
El Dib R, 2015	Grading of Efficacy-Effectiveness in Clinical Trials (GEECT)	Characterization of RCTs	Same as PRECIS (10)	0 to 10 where 0=high efficacy; 10=high effectiveness	Overall 0.11 (Kappa)
Loudon K, 2015	PRECIS-2	RCT design	Eligibility, Recruitment, Setting, Organization, Flexibility, Follow-Up, Primary Outcome, Primary Analysis (9)	1 to 5 1=very explanatory; 5=very pragmatic	0.24 to 0.94 across domains (ICC)
Bossie C, 2016	A Study Pragmatic-Explanatory Characterization Tool-Rating (ASPECT-R)	Characterization of RCTs	Eligibility, Flexibility, Setting, Follow-Up, Primary Outcomes, Compliance (6)	0 to 6 where 0=extremely explanatory; 6=extremely pragmatic	Overall 0.87 (ICC)
Wieland S, 2017	Rating of Included Trials on the Efficacy-	Characterization of Systematic Reviews	Participants Characteristics, Trial Setting, Flexibility of	1 to 5 where 1=strong emphasis on	0.23 to 0.45 across domains (ICC)

	Effectiveness Spectrum (RITES)		Intervention(s), Clinical Relevance (4)	efficacy; 5=strong emphasis on effectiveness	

RCT: Randomized controlled trial; ICC: Intraclass correlation coefficient

1.3 Pragmatic Trials as a Source of Heterogeneity in Systematic Reviews

Although the evaluation of pragmatic and explanatory primary trials in systematic reviews is an emerging topic, researchers have focused mainly on how pragmatism impacts treatment effect sizes in addition to the application and reliability of the available tools. While important developments have been made related to the design and characterization of pragmatic trials, there remains a lack of information regarding how pragmatism may contribute as a source of heterogeneity among studies utilizing similar or the same interventions.

Heterogeneity can be considered as any kind of variability between studies in a systematic review due to variability in participants, interventions, and outcomes (clinical heterogeneity), variability in study design and risk of bias (ROB; methodological heterogeneity), and variability in intervention effects being assessed in each trial (statistical heterogeneity) resulting from clinical or methodological heterogeneity, or both.[18] Heterogeneity is often measured by I^2, a statistic for quantifying inconsistency which describes the percentage of variability in effect estimates due to heterogeneity.[18] The Cochrane Handbook provides rough thresholds for interpreting I^2 where 0% to 40% might not be important, 30% to 60% may represent moderate heterogeneity, 50% to 90% may represent substantial heterogeneity, and 75% to 100% may represent considerable heterogeneity.[18] Literature has suggested that intervention effects decrease as trials become more pragmatic in design due to greater diversity in both the participants and the delivery of such trials.[12 19] Since systematic reviews of RCTs typically include studies with similar or the same intervention(s) and outcome(s), trials that are more pragmatic and more explanatory may be pooled in a meta-analysis consequently affecting heterogeneity.

Subgroup analyses and meta-regression are ways to explore heterogeneity and gain insight into why results from outcomes may be inconsistent between studies.[18] If heterogeneity is substantial, due to the degree of pragmatism, it might not be appropriate to pool data from pragmatic and explanatory trials. The use of the PRECIS-2 tool, and secondarily, the RITES tool could provide important information for authors of systematic reviews with regards to pooling data from primary RCTs based on the degree of pragmatism.

1.4 Objectives

The primary objective of this research is to investigate the role of pragmatism as a source of heterogeneity in systematic reviews through the use of PRECIS-2. This will be accomplished by:

1. Identifying systematic reviews with meta-analyses of RCTs with moderate to considerable heterogeneity ($I^2 \geq 50\%$)
2. Applying the PRECIS-2 scoring system to RCTs of 10 randomly selected systematic reviews to assess the contribution of pragmatism

3. Determining how much of this heterogeneity may be explained by pragmatism as assessed by PRECIS-2

The secondary objective is to compare and contrast the application of PRECIS-2 with RITES. This will be accomplished by:

4. Applying the RITES scoring system to RCTs in a subset of 5 randomly selected systematic reviews to assess the contribution of pragmatism
5. Evaluating the correlation between PRECIS-2 and RITES

Chapter 2: General Methods

2.1 Study Design

This study was designed as a cross-sectional methodological review. A literature search using the Cochrane Library was conducted for published reviews of RCTs from January 1, 2014 to January 1, 2017. The Cochrane database was selected based on the consistency of methodology and the quality of the systematic reviews.[20] We limited the search to the Cochrane Reviews Database and included the key terms *randomize* and *RCT** in titles, abstracts and keywords with word variations in an effort to capture all systematic reviews of RCTs published during the selected timeframe. Inclusion criteria were systematic reviews of RCTs from any Cochrane Review Group with at least 10 studies considered in one pooled effect relating to the primary outcome and at least moderate heterogeneity ($I^2 \geq 50\%$).[18] Exclusion criteria were systematic reviews of non-randomized, quasi-randomized or crossover trials.

2.2 Screening

Two reviewers (TA, KA) independently screened systematic review titles and abstracts retrieved by the search. Screening took place over the course of two stages, the first being a calibration stage and the second being an independent screening stage with the remaining citations. Following screening, we identified full texts of potentially eligible systematic reviews for data abstraction. We resolved disagreements about review inclusion by consensus and expert advice (LM) was available if a consensus could not be reached. Of the eligible systematic reviews, 10 were selected at random to keep the data manageable. We performed random selection using a random numbers generator in Statistical Package for Social Sciences (SPSS) v.23 (IBM Corp, Armonk, NY, USA).

2.3 Data Abstraction

Three reviewers (TA, KA, DL) used standardized data abstraction forms to independently extract data from each systematic review and its included trials, in duplicate. We extracted data at the systematic review level including information pertaining to the number of primary RCTs, types of participants, intervention and comparator, primary outcome, type of effect measure, number of participants, point estimate, 95% confidence intervals, heterogeneity (I^2) and if heterogeneity was explored. We extracted data at the primary study level including information pertaining to ROB, year of publication, number of sites and number of participants randomized to the intervention and comparator groups. We extracted ROB per the judgements of the review authors however if an overall ROB assessment was not provided by the authors,

one was assigned. The core ROB domains included random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, and selective reporting. We assessed an overall low ROB if 4 or more domains were low ROB and none high ROB; unclear if 3 or more domains were unclear ROB and none high ROB; and high if any of the 6 domains were high ROB. We resolved disagreements regarding data abstraction and assessment of overall ROB by consensus. Expert advice (LM) and additional data extraction and/or ROB assessment from the uninvolved reviewer were available if a consensus could not be reached. When there was missing or unclear information at the systematic review level, we contacted the authors of the review for clarification. When there was missing or unclear information at the primary study level, we contacted the authors of the primary study. We performed title and abstract screening, full text screening and data abstraction at the systematic review level in Distiller SR (Evidence Partners, Ottawa, Canada). We performed primary study data abstraction in Excel (Microsoft, Redmond, USA).

2.4 Application of PRECIS-2 and RITES Tools

Two teams of two reviewers (KA, TA; DL, TA) applied PRECIS-2 to all primary studies within their respective systematic reviews. Studies were scored across each of the 9 PRECIS-2 domains and a summary score was provided for each study ranging from 9 (very explanatory) to 45 (very pragmatic). A calibration phase with all reviewers took place using a minimum of 10 primary RCTs to ensure consistency in scoring across each PRECIS-2 domain. Following calibration, PRECIS-2 domains for the remainder of the included primary RCTs were scored independently, in teams of two reviewers. Scores of 3 were given where there was missing information, a method akin to that of Loudon et al. in their evaluation of inter-rater reliability.[16]

We applied the RITES tool to primary studies in a subset of 5 included systematic reviews. We scored studies across each of the 4 domains and a summary score was provided for each study ranging from 4 (strong emphasis on efficacy) to 20 (strong emphasis on effectiveness). Similar to scoring PRECIS-2, a calibration phase with all reviewers took place using a minimum of 10 primary RCTs to ensure consistency in scoring. Rating took place independently, in teams of two reviewers with at least a week time period between scoring PRECIS-2 and RITES, to ensure reviewers did not recall and replicate their responses. Scores of 3 were given where there was missing information. Although this method of missing data imputation was not outlined in the article by Wieland et al. it was used to maintain consistency in how missing data were dealt with between the two tools.

The ICC was used to measure inter-rater reliability between independent reviewers on PRECIS-2 and RITES domains and their summary scores. We considered an ICC of 0.21 to 0.40 as fair agreement, 0.41 to 0.60 as moderate agreement, 0.61 to 0.80 as substantial agreement, and 0.81 to 1.0 as almost perfect agreement.[21] Scoring disagreements were resolved by consensus and additional scoring from the uninvolved reviewer was an option, if needed. Inter-rater reliability for PRECIS-2 and RITES is

described by the ICC and 95% confidence intervals for all domains and summary score for each team of reviewers and by systematic review.

For each rating tool, a guidance document with descriptions of the domains and examples was provided to aid reviewers in appropriately selecting a score. Following independent rating, consensus meetings were held and a single score for each domain of PRECIS-2 and RITES was determined. The results thus included independent PRECIS-2 and RITES scores, and consensus PRECIS-2 and RITES scores as agreed upon by two reviewers.

Chapter 3. Statistical Methods

3.1 Description of Systematic Reviews and Primary Studies

We describe general characteristics of included systematic reviews by author information, number of primary RCTs, types of participants, intervention and comparator, primary outcome, type of effect measure, number of participants, point estimate with 95% confidence intervals, heterogeneity (I^2) and methods used to explore heterogeneity. We describe general characteristics of primary studies by author information, year of publication, number of sites, number of participants randomized to the intervention and comparator groups, and overall ROB assessments. Data are reported as total counts (percentages) or text, whichever is most appropriate to use for the characteristic.

3.2 Description of PRECIS-2 and RITES

PRECIS-2 scores achieved by consensus are described by systematic review for each domain and summary score in mean and standard deviation (SD), and range using minimum to maximum scores. We use the PRECIS-2 'wheel' to visually depict how explanatory or pragmatic a primary RCT is based on scores from each of the 9 domains. We describe RITES scores as achieved by consensus in the same manner as with PRECIS-2. However, instead of a 'wheel' to depict the degree of pragmatism, a spectrum is visually presented for each domain where primary studies are plotted to show differences between the trials.

3.3 Approaches to Dealing with Heterogeneity

Several statistical approaches were undertaken to explore pragmatism as a potential source of heterogeneity. As a primary analysis, we built linear random effects meta-regression models for each systematic review. The RCT was considered the unit of analysis and the dependent variable for each study was the mean difference or standardized mean difference, log odds or log hazard ratio depending on the nature of the outcome, accompanied by the standard error. First, meta-regression models were built in the absence of any covariates primarily to generate estimates of residual I^2. Second, PRECIS-2 as a continuous variable (9 to 45) was included in the model to determine its effect on heterogeneity. We describe meta-regression results by effect size with 95% confidence intervals in the absence of any covariates, and I^2 with and without PRECIS-2. These analyses were repeated for each systematic review and

across systematic reviews, by pooling similar effect measures (odds ratio, hazard ratio or standardized mean difference) using the systematic review as a grouping variable.

As there are no specific cut-off values for what is considered a pragmatic or explanatory trial, we classified RCTs in three categories, similar to Yoong et al.[12] Tertiles were based on the range of PRECIS-2 summary scores within this review (15 to 36) rather than the full PRECIS-2 range (9 to 45) since primary studies were not found to be at either extreme end. We considered an RCT in tertile 1 (T1; 15 to 21) as having more explanatory tendencies, tertile 2 (T2; 22 to 27) as having equally explanatory and pragmatic tendencies, and tertile 3 (T3; 28 to 36) as having more pragmatic tendencies. As a secondary analysis, these classifications were used as a categorical covariate in random effects meta-regression models. An additional exploratory secondary analysis included using ROB and individual PRECIS-2 domains as independent covariates for each systematic review. We included ROB as a covariate since ROB may be higher in pragmatic RCTs than explanatory ones due to higher risk of selection bias, challenges with blinding, and possibly more loss to follow-up. We considered PRECIS-2 and ROB for entry to subsequent multivariate models if $p<0.1$.[22]

We converted odds, risk or hazard ratios to their logarithmic estimates, and 95% confidence intervals to standard errors in Review Manager (RevMan) 5.3 (The Nordic Cochrane Centre, Copenhagen, DK). Systematic reviews using standardized mean differences were reported with 95% confidence intervals, however there lacks a conversion option to standard error in RevMan. We consulted the Cochrane Handbook[18] and an Excel calculator provided by the University of Portsmouth for The Cochrane Collaboration was used to make the conversions (University of Portsmouth, Portsmouth, UK). When conducting meta-regression across systematic reviews, risk ratios were converted to odds ratios and further converted to log odds ratios for the purposes of pooling. Similarly, we converted mean differences to standardized mean differences for pooling. When primary studies were unattainable, we repeated a random effects meta-analysis in RevMan to obtain a new pooled effect estimate. All other statistical tests including inter-rater reliability, correlation, and meta-regression were performed in Stata/IC 15.0 (Statacorp, Texas, USA).

3.4 Correlation

Spearman's correlation coefficient was used to quantify the correlation between PRECIS-2 and RITES. A Spearman's rank correlation coefficient of 0 to 0.3 was considered negligible correlation, 0.3 to 0.5 was considered low correlation, 0.5 to 0.7 was considered moderate correlation, 0.7 to 0.9 was considered high correlation, and 0.9 to 1.0 was considered very high correlation (either positive or negative correlation).[23] The results are described by Spearman's rho with 95% confidence intervals for the correlation across systematic reviews.

Meta-regression, correlation, and inter-rater reliability were performed in Stata/IC 15.0 (Statacorp, Texas, USA). Figure 2 outlines the statistical plan with descriptions of the primary and secondary analyses.

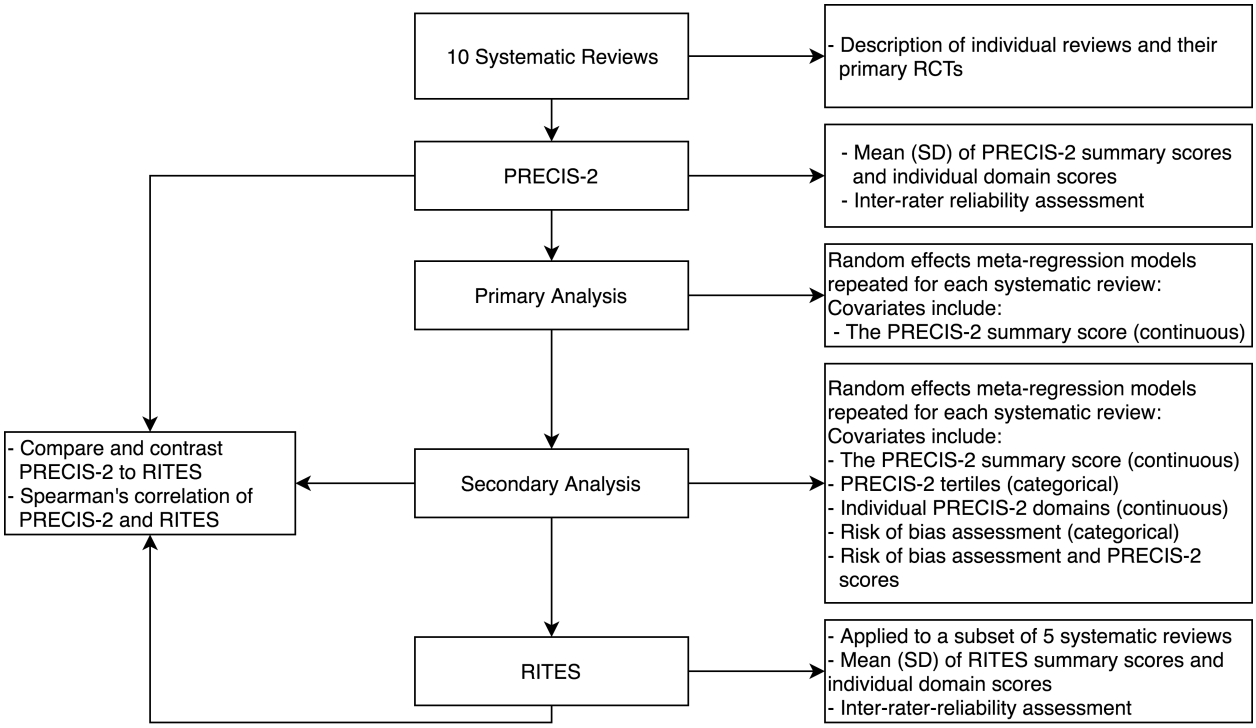


Figure 2. Flow diagram of the analysis plan.

Chapter 4: Results

4.1 Results of the Search

Figure 3 outlines the results of the systematic review search and selection process. The Cochrane Library search strategy was conducted on February 12, 2017 which resulted in 2617 citations of systematic reviews of RCTs from January 1, 2014 to January 1, 2017. Of the 2617 citations, 256 were retained for full text screening. Full text screening identified 52 systematic reviews which were considered for random selection. Of the 52 systematic reviews, 10 were randomly selected for inclusion in this methodological review.

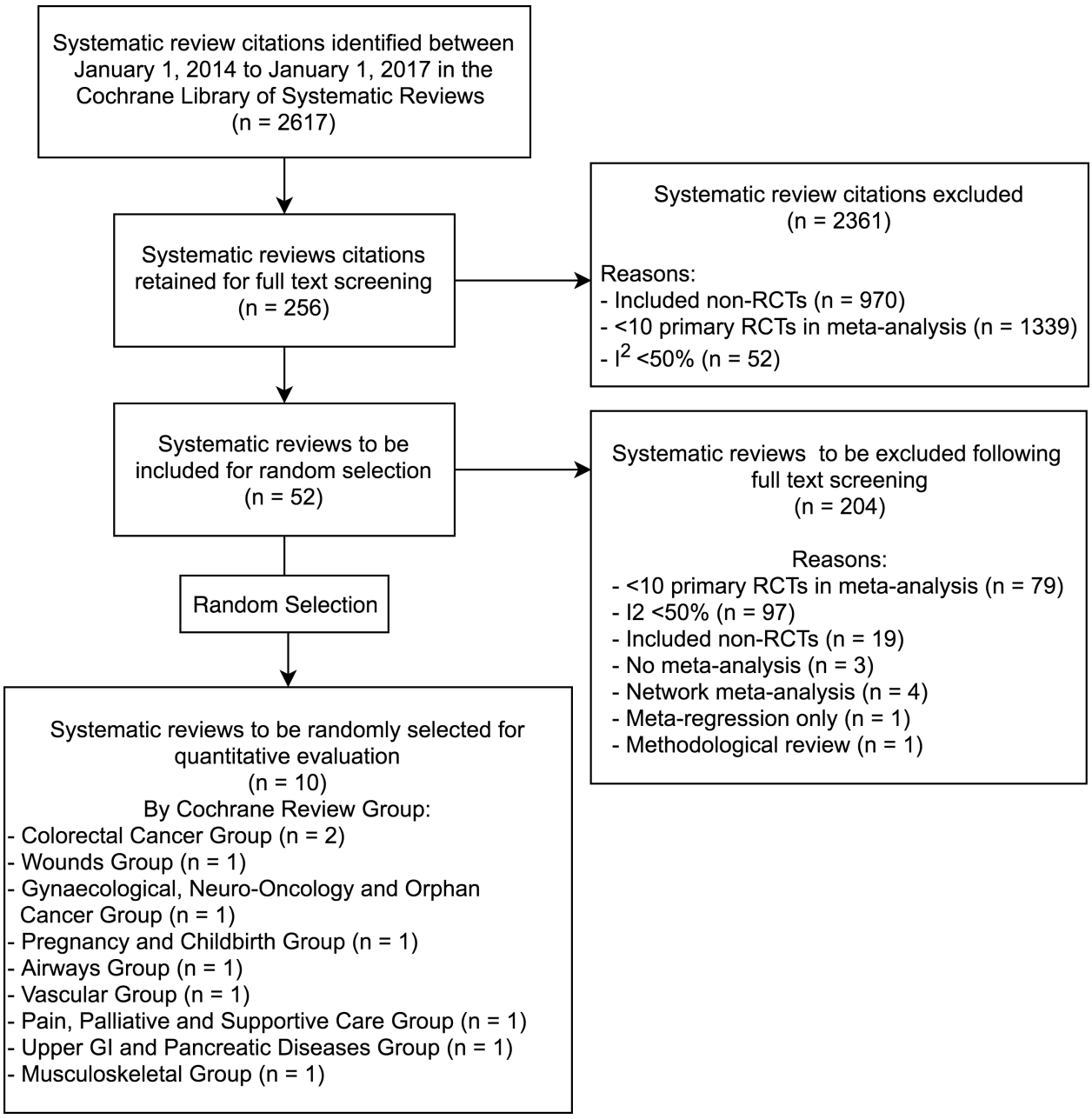


Figure 3. Flow diagram of the study selection procedure.

4.2 Characteristics of Systematic Reviews and Primary Studies

Characteristics of the 10 randomly selected systematic reviews studies can be found in Table 2 and characteristics of each primary study can be found in Appendix 1. In summary, across 10 systematic reviews, there were 132 primary RCTs with 134 comparisons of which 128 articles with 130 comparisons were obtained and are included in the analysis. Authors of the unattainable primary studies and their accompanying systematic reviews were contacted. All primary studies were randomized at the individual level. Three systematic reviews compared interventions head to head,[24-26] five compared interventions to either a placebo or usual care,[27-31] one

compared an intervention to either usual care, control or an alternative intervention[32], and one compared an intervention to usual care or no treatment.[32 33] Three of the systematic reviews involved surgical interventions[24-26] while the remaining seven systematic reviews involved interventions administered in an outpatient setting.[27-33] Of the systematic reviews with interventions in an outpatient setting, two involved behavioural interventions,[31 32] and five involved drug interventions as topical,[27] oral[29 30], injectable,[28] or combination delivery[33] treatments.

Table 2. Characteristics of included systematic reviews (N=10)

Primary Author, Year (Number of Articles)	Types of Participants	Intervention	Comparator	Primary Outcomes	Effect Estimate	Number of Participants	I^2 (%)	Heterogeneity Explored (Yes or No)
Hafner S, 2015 (16)	Adult male and female participants undergoing colonoscopy	Colonoscopy with water infusion in lieu of air (water exchange or water immersion) during insertion	Standard colonoscopy with air insufflation	Cecal intubation rate	RR: 1.0 (0.97, 1.03)	2933	72	Yes
Martí-Carvajal AJ, 2015 (12)	Adults (>18 years of age) with a diabetic foot ulcer of any aetiology	Any growth factor	Standard care (i.e. antibiotic therapy, debridement, wound dressings) alone or plus placebo	Complete wound healing	RR: 1.51 (1.31, 1.73)	1139	51	Yes
Akl E, 2014 (11)	Patients with cancer with no standard indication for prophylactic anticoagulation or for therapeutic anticoagulation	Parenteral anticoagulants such as unfractionated heparin, low molecular weight heparin and fondaparinux	Placebo or no intervention	All-cause mortality over the duration of the trial	HR: 0.84 (0.74, 0.96)	5254	58	Yes
Buppasiri P, 2015 (13)	Pregnant women who received any calcium supplementation	Calcium supplementation during pregnancy	Placebo or no treatment	Preterm birth less than 37 weeks' gestation	RR: 0.86 (0.70, 1.05)	16 139	57	Yes
Hnin K, 2015 (13)	Adult and paediatric participants diagnosed with bronchiectasis who reported daily sputum expectoration for at least three months	Any dose of prolonged antibiotic therapy of four or more weeks	Placebo or as required treatment	Exacerbations	OR: 0.31 (0.19, 0.52)	884	51	Yes
Birch DW, 2016 (19)	Adults and children undergoing laparoscopic abdominal surgery	Heated (with or without humidification) gas insufflation	Cold gas insufflation	Change in intra-operative core temperature	MD: 0.21 (0.06, 0.36)	1100	86	Yes
Lane R, 2014 (12)	Participants with symptomatic intermittent claudication due to	Any exercise programme used in the treatment of intermittent	Usual care or placebo	Maximal walking time in minutes	MD: 4.51 (3.11, 5.92)	577	82	Yes

	atherosclerotic disease	claudication was included, such as walking, skipping and running						
Bennett S, 2016 (12)	Adults aged 18 years and older were who could have been receiving curative or palliative treatment or long-term follow-up, or could have had no evidence of active disease	Educational interventions designed specifically to manage cancer-related fatigue, or educational interventions targeting a constellation of physical symptoms or quality of life where fatigue was the primary focus	Usual care or wait list controls, attention controls, or an alternative intervention for cancer-related fatigue	General fatigue assessed by validated fatigue scales or by any method of self-evaluation	SMD: -0.27 (-0.51, -0.04)	1680	80	Yes
Song H, 2016 (10)	Adults with histologically-confirmed adenocarcinoma of the stomach or of the gastro-esophageal junction with locally advanced un-resectable or metastatic disease and esophageal adenocarcinoma	Molecular-targeted agents (i.e. anti-EGFR agents, VEGF-targeting agents) plus conventional chemotherapy	Conventional chemotherapy alone or no treatment	Overall survival (survival until death from all causes)	HR: 0.92 (0.80, 1.05)	3843	61	Yes
Hofstede SN, 2015 (14)	Patients undergoing total knee arthroplasty for osteoarthritis or rheumatoid arthritis	Mobile bearing (meniscal or rotational) implant	Fixed bearing polyethylene implant	Knee Society clinical score	MD: -1.06 (-2.87, 0.75)	1845	77	No

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference

4.3 PRECIS-2 Results

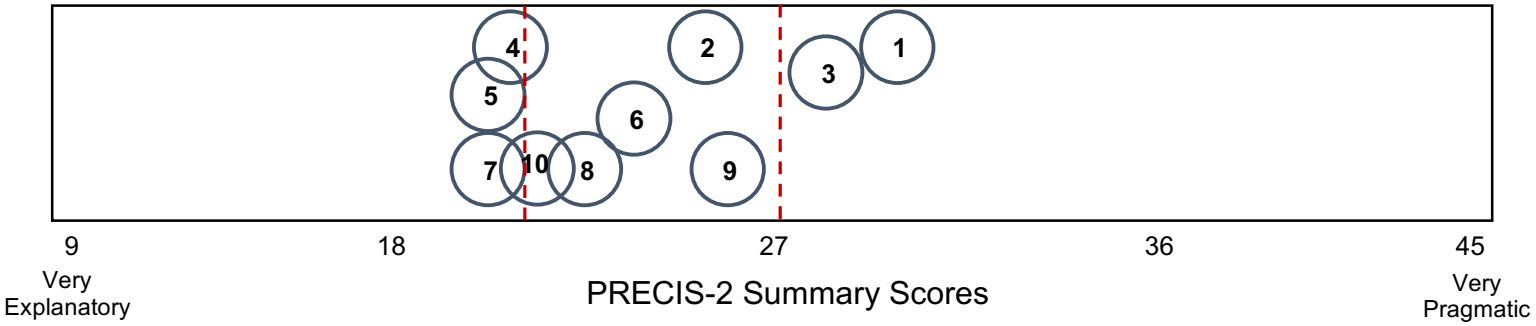
The PRECIS-2 tool was used to score a total of 128 primary studies across the 10 included systematic reviews.

4.3.1 Missing Data

There were missing data among primary RCTs of systematic reviews for several of the PRECIS-2 domains. Scores of 3 were given when there was not enough information to adequately assess a domain. The eligibility, organization, flexibility (delivery), and primary analysis domains had full data across systematic reviews. The domain that consistently had missing data was recruitment, affecting 100% of systematic reviews with missing data between 13% to 100% within a review. Recruitment techniques were rarely discussed as the emphasis on study enrollment was usually placed on eligibility criteria. Missing data for setting was minimal, only two RCTs did not provide information about the setting of the trial. Missing data for flexibility (adherence) was minimal as well, with only four primary RCTs lacking information about adherence measures; three of the four primary studies were from a single systematic review. Follow-up measurements and intensity of visits were not specified in one RCT. Lastly, the primary outcome was not explicitly reported in five RCTs, two of which were from one review. There were no distinct patterns of missing information, primary RCTs did not have more than two domains with missing information. Mean scores and frequency of missing data by domain for each systematic review are described in Appendix 2.

4.3.2 Mean PRECIS-2 Summary Scores

Across systematic reviews, PRECIS-2 scores of primary RCTs ranged from a minimum summary score of 15 to a maximum summary score of 36, moderately encompassing the full pragmatic-explanatory continuum. Mean PRECIS-2 summary scores according to systematic review had a much smaller range of between 20 to 30, denoting average scores that were all equally pragmatic-explanatory with no reviews having mean scores that were predominantly explanatory or predominately pragmatic however, within the context of this research, reviews by Hafner et al., and Akl et al. had pragmatic tendencies based on our division of tertiles (mean summary scores ≥28). Furthermore, the reviews by Hafner et al., and Birch et al. were the only two with pragmatic primary studies as measured by PRECIS-2. The remainder of the reviews included primary studies that were scored as either explanatory or equally pragmatic-explanatory (Figure 4).



Systematic Review Primary Author, Year	Number	Systematic Review Primary Author, Year	Number
Hafner S, 2015	1	Birch DW, 2016	6
Martí-Carvajal AJ, 2015	2	Lane R, 2014	7
Akl EA, 2014	3	Bennett S, 2016	8
Buppasiri P, 2015	4	Song H, 2016	9
Hnin K, 2015	5	Hofstede SN, 2015	10

Figure 4. Mean PRECIS-2 scores as represented by numbered circles according to corresponding systematic review. Dotted lines represent division of tertiles where T1=15 to 21; T2=22 to 27; T3=28 to 36. (N=10).

4.3.3 Mean PRECIS-2 Scores by Domain

Mean PRECIS-2 scores and range of summary scores according to systematic review can be found in Table 3. Briefly, mean scores across the majority of domains were equally pragmatic-explanatory. Exceptions were the primary outcome domain which had mean scores that were mostly pragmatic suggesting the majority of the reviews included primary RCTs with patient important outcomes, and the organization domain which had mean scores that were mostly explanatory suggesting the majority of reviews included trials that required additional expertise and/or resources in order to provide the intervention(s).

The systematic review by Hafner et al. included primary trials with a mostly pragmatic approach to recruitment while the remaining reviews were mainly equally pragmatic-explanatory in scoring suggesting a mix of approaches were used to recruit participants across primary studies such as consecutive enrollment and invitation. However, it is important to note that there was a large amount of missing data for this domain. Reassuringly, when we removed primary studies with imputed scores of 3 from the analysis, the means and SDs for the recruitment domain remained virtually unchanged. The exception was the systematic review by Song et al., which had 100% missing data for the recruitment domain.

Two systematic reviews by Akl et al. and Song et al. included studies that were conducted in mostly pragmatic settings such as health care centers aligned with usual care for the study population. The review by Hafner et al. included studies that generally had flexibility in the delivery of the intervention leaving administration of the intervention(s) up to the healthcare providers to decide. Two systematic reviews by Hafner et al. and Birch et al. had primary trials where flexibility of adherence was primarily pragmatic with little, if any, measures in place to ensure participant adherence.

Mean scores for follow-up were pragmatic in reviews by Hafner et al. and Martí-Carvajal et al. where follow-up measures and visit intensity were mostly similar or the same to that of usual care. Three systematic reviews by Hafner et al., Akl et al., and Song et al. included studies that mainly followed a pragmatic approach to primary analysis through the intention to treat principle. Mean scores for eligibility were either explanatory or equally explanatory-pragmatic suggesting the primary studies, in general, included eligibility criteria such that only a select sample of the population who would potentially receive and benefit from the treatment if it were part of usual care, were considered for participation.

Individual PRECIS-2 domain scores for primary RCTs can be found in Appendix 3. PRECIS-2 wheels visually describing the maximum pragmatic and explanatory trials according to systematic review can be found in Appendix 4.

Table 3. PRECIS-2 scores for systematic reviews (N=10)

Primary Author, Year (number of articles)	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	Mean PRECIS-2 Score	(Min, Max Score)
Hafner S, 2015 (16*)	2.8 (0.9)	3.6 (0.8)	2.6 (1.4)	2.3 (0.8)	3.6 (0.5)	4.6 (1.0)	3.9 (0.6)	2.9 (1.6)	3.8 (1.1)	29.9 (3.9)	24, 36
Martí-Carvajal AJ, 2015 (11*)	1.6 (0.7)	3.0 (0.4)	3.5 (1.7)	2.5 (1.0)	2.2 (0.8)	2.8 (1.2)	3.5 (1.0)	4.1 (1.4)	3.1 (1.4)	26.2 (4.5)	17, 32
Akl EA, 2014 (11)	2.9 (0.9)	3.0 (0.4)	4.2 (1.6)	1.9 (0.9)	2.1 (0.8)	3.1 (0.7)	3.2 (1.2)	4.4 (0.9)	4.1 (0.9)	28.8 (3.1)	25, 33
Buppasiri P, 2015 (12*)	2.7 (1.2)	2.6 (1.2)	2.2 (1.6)	1.1 (0.3)	2.1 (0.8)	1.8 (1.0)	3.3 (1.0)	3.2 (0.7)	2.8 (1.3)	21.6 (4.7)	15, 30
Hnin K, 2015 (13)	2.1 (1.0)	3.4 (0.9)	3.1 (1.8)	1.2 (0.4)	1.5 (0.5)	2.1 (0.8)	1.9 (1.2)	2.2 (1.3)	3.0 (1.6)	20.4 (3.9)	15, 28
Birch DW, 2016 (18*)	2.9 (1.2)	3.4 (0.5)	1.5 (1.1)	2.3 (0.9)	2.8 (0.8)	3.6 (1.1)	3.3 (0.8)	2.6 (1.3)	2.9 (0.8)	25.4 (4.6)	19, 36
Lane R, 2014 (12)	2.3 (1.3)	2.8 (0.8)	1.6 (1.2)	1.8 (0.9)	1.3 (0.6)	2.4 (0.5)	1.5 (0.7)	3.8 (1.5)	2.8 (0.8)	20.2 (3.1)	17, 26
Bennett S, 2016 (12)	2.5 (0.8)	2.5 (0.9)	3.0 (1.8)	1.3 (0.5)	1.7 (1.0)	1.9 (1.0)	2.8 (0.9)	4.6 (0.8)	2.9 (0.9)	23.3 (4.0)	19, 31
Song H, 2016 (10)	1.7 (0.5)	3.0 (0)	4.8 (0.6)	1.9 (0.9)	1.5 (0.8)	3.3 (1.1)	1.5 (0.5)	4.6 (1.3)	4.3 (0.8)	26.6 (2.0)	23, 30
Hofstede SN, 2015 (13*)	2.9 (1.1)	3.0 (0.9)	1.9 (1.3)	1.5 (0.7)	2.9 (0.9)	2.8 (0.8)	2.7 (1.3)	2.5 (1.1)	2.3 (1.1)	22.5 (3.8)	18, 30
Overall	2.5 (1.1)	3.1 (0.8)	2.7 (1.7)	1.8 (0.9)	2.2 (1.0)	2.9 (1.3)	2.8 (1.2)	3.4 (1.5)	3.2 (1.2)	24.6 (5.0)	15, 36

All data are described as mean (standard deviation) unless otherwise indicated, where there was missing data a score of 3 was given; *1 primary RCT missing

4.3.4 Inter-Rater Reliability of PRECIS-2

Inter-rater reliability was measured as a quality indicator of PRECIS-2 scoring. Two teams of reviewers (KA, TA; DL, TA) scored five systematic reviews each. For both teams, agreement was substantial (ICC: 0.64; 95% CI: 0.40, 0.78; ICC: 0.73; 95% CI: 0.55, 0.83). Agreement varied by PRECIS-2 domain and by systematic review, with an ICC as low as -0.01 (poor agreement) for the domains of recruitment, organization and flexibility of delivery (Table 4). Across systematic reviews, agreement ranged from fair (ICC: 0.21) to almost perfect (ICC: 0.92). The ICCs by individual PRECIS-2 domain for each systematic review can be found in Appendix 5.

Table 4. Inter-rater reliability of PRECIS-2 domains and summary score (N=128 articles)

Domain	ICC Team 1†	95% CI Team 1†	ICC Team 2‡	95% CI Team 2‡
Eligibility	0.35	-0.03, 0.60	0.60	0.35, 0.76
Recruitment	0.17*	-0.21, 0.46	0.82	0.71, 0.89
Setting	0.90	0.82, 0.04	0.92	0.86, 0.95
Organization	-0.01*	-0.41, 0.31	0.50	0.19, 0.69
Flexibility: Delivery	0.19*	-0.26, 0.49	0.29	-0.10, 0.55
Flexibility: Adherence	0.68	0.41, 0.81	0.64	0.42, 0.78
Follow-Up	0.24*	-0.27, 0.54	0.55	0.26, 0.73
Primary Outcome	0.74	0.57, 0.84	0.87	0.79, 0.92
Primary Analysis	0.69	0.48, 0.81	0.87	0.79, 0.92
Overall Score	0.64	0.40, 0.78	0.73	0.55, 0.83

CI: confidence interval; ICC: intraclass correlation coefficient; †Reviewers KA, TA, 63 articles; ‡Reviewers TA, DL, 65 articles; *p>0.05 not statistically significant

4.4 Primary Analysis

4.4.1 Random Effects Meta-Regression with PRECIS-2

Pragmatism as measured by PRECIS-2 did not explain heterogeneity in any of the 10 systematic reviews. Small reductions in heterogeneity occurred when including the PRECIS-2 summary score as a covariate in reviews by Lane et al, Hafner et al, Birch et al, and Hofstede et al. however these reductions were 13% or less resulting in heterogeneity of the primary outcome that was still substantial or considerable (Table 5). R^2 was zero across all meta-regression models of systematic reviews with the exception of Birch et al. and Lane et al. suggested that for the most part, PRECIS-2 explained less heterogeneity than would be expected by chance, resulting in poor model fit. Graphs of random effects meta-regression with PRECIS-2 as a covariate can be found in Appendix 6.

Table 5: Exploring heterogeneity through random effects meta-regression methods using PRECIS-2 as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Pooled Effect Size without PRECIS-2 (95% CI)	I^2 without PRECIS-2 (%)	I^2 with PRECIS-2 (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	67.5	-0.8
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	58.8	3.8
Akl EA, 2014 (11)	HR: 0.84 (0.71, 1.00)	58.7	62.3	3.6
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	58.1	4.0
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	54.6	3.6
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	78.7	-5.3
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	68.9	-13.3
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	80.8	0.7
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	61.6	1.6
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	78.2	-0.2

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing, †Statistically significant, p=0.042

4.4.2 Random Effects Meta-Regression by Effect Measure with PRECIS-2

The linear effects meta-regression using standardized mean difference as the outcome measure included four systematic reviews with a total of 55 primary studies.[25 26 31 32] Three of the reviews reported their results using mean difference however these were converted to standardized mean difference for the purpose of pooling. The linear effects meta-regression using log odds ratio as the outcome measure included four systematic reviews with a total of 47 primary studies.[24 27 29 30] Three of the reviews reported their results using risk ratio however these were converted log odds ratios for pooling.[24 27 29] Four primary studies had risk ratio of 1.0 therefore were excluded from the analysis. Another primary study had a risk ratio that was not estimable and was also excluded from the analysis. The linear effects meta-regression using log hazard ratio as the outcome measure included two systematic reviews with a total of 21 primary studies.[28 33] Both reviews reported their results in log hazard ratios thus no conversions were required. Among all random effects meta-regression models, heterogeneity was either substantial or considerable prior to including PRECIS-2. When the PRECIS-2 summary score was entered as a covariate, I^2 remained virtually unchanged in each model indicating pragmatism as measured by PRECIS-2 did not reduce heterogeneity. R^2 was either zero or remained unchanged with PRECIS-2, suggesting the summary score explained less heterogeneity than would be anticipated by chance, or did not improve goodness of fit.

Table 6: Exploring heterogeneity through random effects meta-regression methods using PRECIS-2 as a covariate by type of effect measure

Effect Measure (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with PRECIS-2 (%)	I^2 Difference (%)
Standardized Mean Difference (55)	0.35 (0.01, 0.69)	83.9	84.2	0.3
Odds Ratio (47)	1.03 (0.56, 1.90)	62.7	63.3	0.6
Hazard Ratio (21)	0.84 (0.72, 0.98)	59.3	60.1	0.8

4.5 Secondary Analysis:

4.5.1 Random Effects Meta-Regression with PRECIS-2 Tertiles

Pragmatism as measured by PRECIS-2 tertiles did not explain heterogeneity in any of the 10 systematic reviews. There were incremental reductions in I^2 however the decreases were less than 5% resulting in heterogeneity that was still substantial or considerable (Appendix 7).

4.5.2 Random Effects Meta-Regression with Individual PRECIS-2 Domains

When including individual PRECIS-2 domains as independent covariates, I^2 reduced from 54% to 23% with the flexibility (adherence) domain the review by Buppasiri et al. I^2 reduced from 55% to 37% with the setting domain in the review by Martí-Carvajal et al., and I^2 decreased from 51% to 38% with the organization domain in the review by Hnin et al. In each case, I^2 was 40% or less, the threshold for which heterogeneity might not be important.[18] For the most part though, when including individual PRECIS-2 domains as covariates, heterogeneity was largely unaffected (Appendix 8).

4.5.3 Random Effects Meta-Regression with Risk of Bias

Similar to PRECIS-2 tertiles, ROB did not explain heterogeneity in any of the 10 systematic reviews. There were small reductions in I^2, less than 15%, resulting in heterogeneity that was still substantial or considerable. ROB as a categorical covariate did not result in $p<0.1$ in any meta-regression models among systematic reviews thus multivariate modelling with ROB and PRECIS-2 was not pursued as an exploratory analysis.

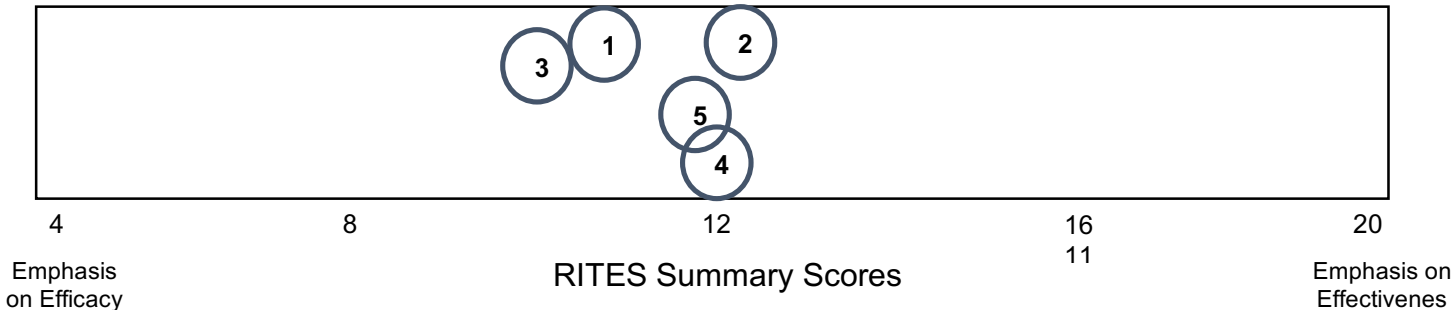
4.6 RITES Scores

A secondary objective of this review was to compare and contrast the application of RITES with PRECIS-2. The RITES tool was applied to 59 primary RCTs in a subset of 5 systematic reviews. [26-28 31 32]

4.6.1 Mean RITES Summary Scores

Across the reviews, RITES scores of primary studies ranged from a minimum summary score of 7 to a maximum summary score of 16, with no scores at either extreme end of the efficacy-effectiveness spectrum. Mean RITES summary scores according to systematic review had a very narrow range from 11 to 13, resulting in average scores that had an equal emphasis on efficacy-effectiveness. The majority of the systematic

review included primary studies that had an emphasis on effectiveness as assessed by RITES. There was only a single review with primary studies that only had emphasis on efficacy or equal emphasis on efficacy-effectiveness (Figure 4).



Primary Author, Year	Number
Martí-Carvajal AJ, 2015	1
Akl E, 2014	2
Lane R, 2014	3
Bennett S, 2016	4
Hofstede SN, 2015	5

Figure 5. Mean RITES scores as represented by numbered circles according to corresponding systematic review (N=5)

4.6.2 Mean RITES Scores by Domain

Mean RITES scores and range of summary scores according to systematic review can be found in Table 7. The participant characteristics domain included mean scores that had an emphasis on efficacy and an equal emphasis on efficacy-effectiveness (mean score and SD range: 1.5 [0.7] to 3.4 [1.0]) suggesting participants involved in the primary trials were, in general, more homogenous than those treated in usual care. The trial setting domain included mean scores that varied along the efficacy-effectiveness spectrum. Lane et al. and Hofstede et al. had more primary studies that were efficacy oriented (mean scores and SDs: 1.3 [0.9] and 1.9 [1.3], respectively) which took place at single academic or highly specialized centers whereas the reviews by Martí-Carvajal et al. and Akl et al. were effectiveness oriented (mean scores and SDs: 3.7 [1.6] and 4.3 [1.3], respectively) and mostly took place at multiple centers that were similar to usual care for the intervention.

Across systematic reviews, the flexibility of intervention domain mainly had an equal emphasis on efficacy-effectiveness with mean scores and SDs ranging from 2.1 (1.0) to 3.3 (1.0). Since the flexibility domain included multiple components such as monitoring, adherence, and co-interventions, the tendency was to score it as equal efficacy-effectiveness particularly if there was flexibility in one or two of the components but no flexibility in another. Notably, the clinical relevance domain had emphasis on effectiveness across systematic reviews (mean score and SD range: 3.5 [0.9] to 4.5 [0.5]) where both the intervention and comparator have the potential to be the gold standard treatment for the participant population. Individual RITES domain scores for each primary RCT can be found in Appendix 9. A visual description of primary trials on

the efficacy-effectiveness spectrum according to systematic review can be found in Appendix 10.

Table 7: RITES scores by systematic review (N=5)

Primary Author, Year (Number of Articles)	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	Mean RITES Score	Range of Total (Min, Max)
Martí-Carvajal AJ, 2015 (11*)	1.8 (1.2)	3.7 (1.6)	2.1 (1.0)	3.7 (0.9)	11.4 (1.8)	9, 16
Akl E, 2014 (11)	1.5 (0.7)	4.3 (1.3)	2.2 (0.9)	4.5 (0.5)	12.5 (1.6)	10, 16
Lane R, 2014 (12)	2.6 (1.2)	1.3 (0.9)	3.3 (1.0)	3.5 (0.9)	10.8 (2.6)	7, 15
Bennett S, 2016 (12)	2.5 (0.9)	2.9 (1.8)	2.8 (0.8)	3.9 (0.7)	12.1 (2.3)	9, 16
Hofstede SN, 2015 (13*)	3.4 (1.0)	1.9 (1.3)	3.1 (0.6)	3.5 (0.9)	11.9 (1.9)	9, 16
Overall	2.4 (1.2)	2.8 (1.7)	2.7 (1.0)	3.8 (0.8)	2.9 (1.3)	7, 16

All data are described as mean (standard deviation) unless otherwise indicated; *1 primary RCT missing

4.6.3 Inter-Rater Reliability of RITES

Two teams of reviewers (KA, TA; DL, TA) scored five systematic reviews in total. For both teams, agreement was almost perfect (ICC: 0.94; 95% CI: 0.85, 0.97; ICC: 0.88; 95% CI:0.77, 0.94). Agreement was almost perfect (>0.80) in the majority of domains with the exception of flexibility of intervention(s) which was substantial (ICC: 0.66) and clinical relevance which was moderate (ICC: 0.54; Table 8). Across systematic reviews, agreement ranged from substantial (ICC: 0.67) to almost perfect (ICC: 0.94). The ICCs by individual RITES domain and systematic review can be found in Appendix 11.

Table 8. Inter-rater reliability of RITES domains and summary score (N=22 articles)

Domain	ICC Team 1†	95% CI Team 1†	ICC Team 2‡	95% CI Team 2‡
Participant Characteristics	0.80	0.76, 0.96	0.84	0.70, 0.92
Trial Setting	1.0	-	0.98	0.96, 0.99
Flexibility of Intervention(s)	0.88	0.56, 0.96	0.66	0.32, 0.82
Clinical Relevance of Intervention(s)	0.85	0.59, 0.94	0.54	0.10, 0.76
Overall Score	0.94	0.85, 0.97	0.88	0.77, 0.94

CI: confidence interval; †Rater 1: KA, Rater 2: TA

4.7 Correlation of PRECIS-2 and RITES

Although PRECIS-2 and RITES evaluate similar concepts, the distinction is in the intended stage at which the tools may provide the most benefit. PRECIS-2 was developed with the intention for it to be applied at the design phase, however the authors and other researchers believe it may have a place in research synthesis.[11 12 34] RITES was recently developed with the intention for it to be applied at the systematic

review stage, however there is limited information to date on how it may be of use to users of systematic reviews.[17] In an effort to quantify the relationship between PRECIS-2 and RITES, a Spearman's rank correlation was undertaken for the 59 articles that scored PRECIS-2 and RITES, across four systematic reviews. Results showed a moderately positive relationship (r_s: 0.55; 95% CI: 0.34, 0.70; Figure 5).

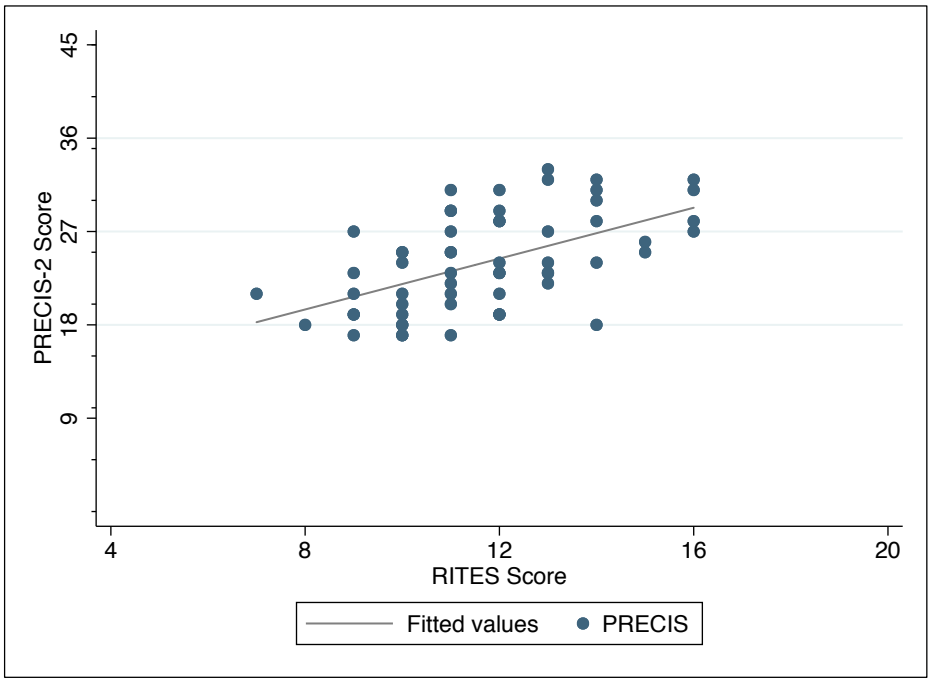


Figure 6. Spearman's rank correlation for PRECIS-2 and RITES summary scores (N=59).

Chapter 5: Discussion

5.1 Summary

Although the concept of pragmatism was first described in 1967, the design and conduct of pragmatic trials has recently gained momentum as health care providers and decision makers seek to determine whether available evidence may be translated and used in real world practice.[6] Thus, the evaluation of pragmatism in primary RCTs of systematic reviews is a novel and relevant topic. Over the past decade, several tools have been developed in an effort to quantify pragmatism; the majority of these deriving from the PRECIS tool published in 2009.[8-12 14] When PRECIS-2 was published in 2015, the authors addressed many of the criticisms the original tool faced and provided a revised tool intended for researchers to align their RCT design to a context in which they believe the intervention would be useful and RCT results applicable. [35] However, decision makers are the ones who will evaluate the RCT and make decisions regarding the implementation of the tested intervention.[35] Systematic reviews of RCTs are an essential scientific activity and the evidence upon which clinical and health system decisions are made.[36] With this in mind, it is important to consider the degree of pragmatism as a source of heterogeneity in systematic reviews as unexplained

heterogeneity can lead to downgrading the body of evidence which in turn could affect whether or not the tested intervention is implemented in a health care system.[37]

5.2 Meta-Regression with PRECIS-2

Overall, pragmatism as measured by PRECIS-2 did not explain heterogeneity. I^2 decreased by no more than 15% when PRECIS-2 was entered as a covariate in random effects meta-regression models for each of the 10 systematic reviews. Adjusted R^2 percentages were indicative of poor model fit. These results were the same for random effects meta-regression models built across systematic reviews using similar effect measures and again when ROB was included as a covariate.

When individual PRECIS-2 domains were included as independent covariates, there was no single one that explained heterogeneity in the systematic reviews. In fact, most reviews had different PRECIS-2 domains that explained heterogeneity though the reductions in I^2 were small. Certain domains explaining more or less heterogeneity among systematic reviews, may help to highlight main methodological differences between primary RCTs on the pragmatic-explanatory continuum. However, only three systematic reviews (Marti-Carvajal et al., Buppasiri et al., and Hnin et al.) resulted in reduced $I^2 \leq 40\%$ when PRECIS-2 domains were included as covariates. Based on these findings, pragmatism appears complex to model and its role in heterogeneity may not be important. Our current understanding is that pragmatism plays a minor part in explaining heterogeneity, if any, and when it does, it is likely to be influenced by other variables that we do not know of. As such, there may be limited or no benefit of using the PRECIS-2 summary score, tertiles or individual domains as covariates in random effects meta-regression models as a method of exploring heterogeneity in systematic reviews.

5.3 Application and Reliability of PRECIS-2

PRECIS-2 was applied to 128 RCTs across 10 systematic reviews resulting in a moderate diversity of trials along the pragmatic-explanatory continuum. Agreement was assessed among two teams where ICCs of PRECIS-2 domains ranged from -0.01 to 0.90 for the first team (KA, TA) and from 0.29 to 0.92 for the second team (TA, DL). These results are comparable to previous findings from inter-rater reliability assessments of PRECIS, including those from Yoong et al. and Bossie et al. who both reported wide variability in agreement between individual PRECIS domains.[10 12] Additionally, Loudon et al. reported similar agreement results with ICCs ranging from 0.24 to 0.94 across PRECIS-2 domains.[16] Despite the variability in agreement between the domains, agreement of the summary score was substantial for both teams, with an ICC of 0.64 for the first team and 0.73 for the second team. Bossie et al. noted the same pattern of results with variability among domain scores yet substantial agreement of the summary score.[10] The authors discussed the domain scores were psychometrically less stable than the summary score and that a larger sample of raters and wider collection of publications would provide greater confidence for the assessment of agreement of individual domains.[10] This review applied PRECIS-2 to primary studies in systematic reviews among 9 different Cochrane Review Groups however, was limited to 3 reviewers. A larger number of review teams would have

provided a more robust assessment of agreement both between domains and in the summary score.

Reasons for lower agreement may be attributed to several factors. First, a lack of clinical expertise could have resulted in differential scoring between reviewers. In circumstances where there was limited knowledge surrounding the health condition and intervention of interest, clinical guidelines were accessed and summarized to provide context. Unfortunately, guidelines do not provide context for each PRECIS-2 domain therefore in some cases, a certain amount of subjectivity was exercised which may have impacted agreement. Second, missing information may have affected how a domain was scored. Although scores of 3 were given when a domain could not be reasonably assessed, every effort was made to provide a score rather than indicate it as missing. In some situations, assumptions may have been made about one or more domains, leading to discrepancies in agreement. Third, there was difficulty in deciding what constituted a one point difference in score among reviewers. For some domains, such as primary analysis, there were clear criteria for what was considered a very pragmatic or a very explanatory study (intention to treat principle=very pragmatic whereas per protocol analysis=very explanatory). By having anchor criteria, it was simpler to assess a domain and provide a score. For other domains, such as organization, the criteria were less clear since it included multiple components (resources, provider expertise and organization of care delivery). For this domain, anchor criteria could not be easily provided as they may have been too restrictive or not applicable resulting in differences in scoring between reviewers. Fourth, due to the length of PRECIS-2 and the number of primary studies each reviewer scored, it is possible that information related to a domain was simply missed by one reviewer. This was particularly true for domains that had multiple components where if one of the components was missed, the study could have been scored as more or less pragmatic than it actually was. For this reason alone, it is advisable that PRECIS-2 is scored in duplicate to ensure that primary study information pertaining to each domain is adequately captured and assessed.

Lastly, although there may be value in applying PRECIS-2 retrospectively to determine pragmatism of published trials, we believe the greatest benefit of PRECIS-2 is in its intended purpose which is use in the development and design of RCTs. Due to the comprehensive nature of PRECIS-2, we found some of the domains were simply not reported in trial publications thereby limiting its utility in retrospective application. In the review by Song et al., recruitment methods were not reported in any of the 10 included studies. It could be that word count restrictions were the primary reason for why this information was not reported however, it should be mentioned that recruitment methods are not a checklist item of the CONSORT or the CONSORT Extension for Pragmatic Trials Statements.[5 38] To increase the utility of PRECIS-2, alignment with CONSORT checklist items should be considered in order to streamline information that is reported to information that is being collected through the individual domains.

5.4 Application and Reliability of RITES

RITES was applied to 59 RCTs across a subset of systematic reviews resulting in a moderate diversity of trials along the efficacy-effectiveness spectrum. Agreement was assessed among two teams where ICCs of RITES domains ranged from 0.80 to 1.0 for the first team (KA, TA) and from 0.54 to 0.98 for the second team (TA, DL). Interestingly, these results were much higher than those reported by Wieland et al. in their pilot assessment of inter-rater reliability which ranged from 0.25 to 0.66 across the 4 domains.[17] However, it is important to note that the application of RITES criteria took place after PRECIS-2 and even though there was a minimum of one week between scoring PRECIS-2 and RITES, there may have been some degree of recall that occurred. Wieland et al. attributed lower agreement to difficulties in rating due to lack of available information in a primary RCT or lack of clinical expertise.[17] While lack of clinical expertise was likely the reason for only moderate agreement in the clinical relevance domain in this research, missing information was not a concern with only a single domain in a single RCT having missing data pertaining to trial setting. Inter-rater reliability of the summary score was almost perfect with an ICC of 0.94 for the first team and 0.88 for the second team. Although a summary score was not pursued by Wieland et al., there may be utility for one in order to assess the overall emphasis on efficacy-effectiveness, similar to that of PRECIS-2.

5.5 Comparing and Contrasting PRECIS-2 and RITES

The PRECIS-2 and RITES tools were developed with different yet complementary intentions in that PRECIS-2 was devised for use during the trial design stage whereas RITES was designed for use at the systematic review stage however, the authors of PRECIS-2 note that it may have a role in critical appraisal and systematic reviews.[17 34] To date, several researchers have applied adapted versions of PRECIS in a systematic review setting.[11 12] A key difference between PRECIS-2 and RITES is in the length of the tool and extent of study related information which each tool collects. PRECIS-2 is comprised of 9 domains related to eligibility, recruitment, setting, organization, flexibility, follow-up, primary outcome, and primary analysis while RITES is comprised of 4 domains related to participants characteristics, trial setting, flexibility of intervention(s), and clinical relevance.[17 34] Although PRECIS-2 may be more comprehensive than RITES, a drawback is that not all domains may be available in an RCT publication. Conversely, although RITES may be more convenient to use than PRECIS-2 with information that is more readily available in published articles, there could be important information related to efficacy-effectiveness that is not captured.

Feedback from the reviewers post-scoring was that PRECIS-2 criteria took a significant amount of time to apply, potentially limiting its utility. All reviewers claimed RITES on the whole was easier and quicker to apply. There was substantial agreement of PRECIS-2 summary scores, and almost perfect agreement of RITES summary scores among the reviewers providing support for the practicality of an overall score though less support for individual domain scores with PRECIS-2 due to variability in agreement among the domains. However, improved scoring guidance for both tools with clearer definitions and restriction of domains to single, rather than multiple components, could improve agreement of individual domains. Correlation between the tools was

moderately positive and interestingly, RCTs that were scored as having an emphasis on effectiveness with RITES were scored as equally pragmatic-explanatory with PRECIS-2 likely due to the differences in the detail of information each tool is based on. Since there are only limited results for the utility of RITES, more research is needed to determine the value of the tool for both systematic review authors and decision makers.

Lack of clinical expertise was cited as a possible reason for diverse agreement in domains of both PRECIS-2 and RITES,[16 17] which we have also discussed as a reason for low agreement in this review. Additionally, this key point has been consistent among systematic reviews that have applied PRECIS post-hoc.[11 12] If considering either or both of these tools for retrospective application, scoring with clinicians and methodologists through open discussion may be an additional consideration for obtaining scores, which takes the emphasis away from agreement and places it on collaboration and consensus.

5.6 Limitations

There are some limitations of this research, the first being that there may be individual patient level factors that could explain heterogeneity however they were not explored by authors of the Cochrane review or not included in this review. A second limitation of the research is that only Cochrane systematic reviews were considered and they represent only a portion of all systematic reviews published. It is possible that there are reviews of important interventions that we did not consider in this research. However, we regarded Cochrane reviews as ideal since they have consistent methodology, reporting standards, and are widely accepted as the gold standard of systematic reviews.[39] Third, we advise caution in the use of exploratory results since we used a number of covariates potentially resulting in spurious findings including heterogeneity being falsely reduced such as with individual PRECIS-2 domains. Finally, none of the included SRs had any cluster randomized trials. Cluster RCTs tend to be on the pragmatic end of the spectrum. It is unclear how the inclusion of cluster RCTs would influence our results. Current literature suggests the need to explore pragmatism in cluster RCTs on the individual and groups levels and may be a good avenue for further research.

Chapter 6: Future Directions and Conclusions

6.1 Future Directions

Now that PRECIS-2 has been explored as a means of explaining heterogeneity within and among systematic reviews, it may be relevant to apply the same methodology using the RITES tool. Early work in our review has shown that RITES has strong inter-rater reliability and is quick to apply to primary RCTs. Using RITES as a way to explore heterogeneity could complement its potential utility and might provide important information regarding the characterization of trials on the efficacy-effectiveness spectrum. Additionally, it would be beneficial to replicate this research when more pragmatic trials are published in an effort to cover a broader scope of interventions and pragmatic trial designs.

6.2 Conclusions

In summary, this methodological review is one of the first to evaluate the application of PRECIS-2 tool in a systematic review setting as a way of exploring heterogeneity through meta-regression. For the most part, the PRECIS-2 summary score did not explain heterogeneity of primary outcomes within and among systematic reviews. When assessing PRECIS-2 domains as individual covariates, there were differences among the reviews for which domain explained the most heterogeneity however, for the majority of systematic reviews, individual domains did not reduce heterogeneity enough so that it might not be important ($I^2 \leq 40\%$). It appears pragmatism as measured by PRECIS-2 does not explain heterogeneity therefore there is probably little or no need to perform subgroup analyses or meta-regression based on degree of pragmatism. As such, pooling of pragmatic and explanatory RCTs will unlikely be detrimental to meta-analyses.

References:

1. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 2009;62(5):464-75. doi: 10.1016/j.jclinepi.2008.12.011
2. Haynes RB, Sackett DL, Guyatt GH, et al. Clinical epidemiology : how to do clinical practice research. LWW medical book collection. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2006:xv, 496 p.
3. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis* 1967;20(8):637-48.
4. Sox HC, Lewis RJ. Pragmatic Trials: Practical Answers to "Real World" Questions. *JAMA* 2016;316(11):1205-06. doi: 10.1001/jama.2016.11409
5. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 2008;337:a2390. doi: 10.1136/bmj.a2390
6. Patsopoulos NA. A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci* 2011;13(2):217-24.
7. Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. *J Clin Epidemiol* 2006;59(10):1040-8. doi: 10.1016/j.jclinepi.2006.01.011
8. Tosh G, Soares-Weiser K, Adams CE. Pragmatic vs explanatory trials: the pragmascope tool to help measure differences in protocols of mental health randomized controlled trials. *Dialogues Clin Neurosci* 2011;13(2):209-15.
9. El Dib R, Jorge EC, Kamegasawa A, et al. Differences between the real and the desired worlds in the results of clinical trials. *Clinics (Sao Paulo)* 2015;70(9):618-22. doi: 10.6061/clinics/2015(09)04
10. Bossie CA, Alphs LD, Williamson D, et al. Inter-rater Reliability Assessment of ASPECT-R: (A Study Pragmatic-Explanatory Characterization Tool-Rating). *Innov Clin Neurosci* 2016;13(3-4):27-31.
11. Koppenaal T, Linmans J, Knottnerus JA, et al. Pragmatic vs. explanatory: an adaptation of the PRECIS tool helps to judge the applicability of systematic reviews for daily practice. *J Clin Epidemiol* 2011;64(10):1095-101. doi: 10.1016/j.jclinepi.2010.11.020
12. Yoong SL, Wolfenden L, Clinton-McHarg T, et al. Exploring the pragmatic and explanatory study design on outcomes of systematic reviews of public health interventions: a case study on obesity prevention trials. *J Public Health (Oxf)* 2014;36(1):170-6. doi: 10.1093/pubmed/fdu006
13. Waters E, de Silva-Sanigorski A, Hall BJ, et al. Interventions for preventing obesity in children. *Cochrane Database Syst Rev* 2011(12):CD001871. doi: 10.1002/14651858.CD001871.pub3
14. Witt CM, Manheimer E, Hammerschlag R, et al. How well do randomized trials inform decision making: systematic review using comparative effectiveness research measures on acupuncture for back pain. *PLoS One* 2012;7(2):e32399. doi: 10.1371/journal.pone.0032399
15. Loudon K, Zwarenstein M, Sullivan F, et al. Making clinical trials more relevant: improving and validating the PRECIS tool for matching trial design decisions to trial purpose. *Trials* 2013;14:115. doi: 10.1186/1745-6215-14-115

16. Loudon K, Zwarenstein M, Sullivan F, et al. The PRECIS - 2 tool has good inter - rater reliability and reasonable discriminant validity. *J Clin Epidemiol* 2017 doi: 10.1016/j.jclinepi.2017.06.001
17. Wieland LS, Berman BM, Altman DG, et al. Rating of Included Trials on the Efficacy-Effectiveness Spectrum: development of a new tool for systematic reviews. *J Clin Epidemiol* 2017 doi: 10.1016/j.jclinepi.2017.01.010
18. Higgins JPT, Green S, Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. Chichester, England ; Hoboken, NJ: Wiley-Blackwell 2008.
19. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials* 2009;10:37. doi: 10.1186/1745-6215-10-37
20. Petticrew M, Wilson P, Wright K, et al. Quality of Cochrane reviews. Quality of Cochrane reviews is better than that of non-Cochrane reviews. *BMJ* 2002;324(7336):545.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.
22. Kleinbaum DG. Applied regression analysis and other multivariable methods. 4th ed. Australia ; Belmont, CA: Thomson Brooks/Cole 2008.
23. Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. 5th ed. Boston: Houghton Mifflin 2003.
24. Hafner S, Zolk K, Radaelli F, et al. Water infusion versus air insufflation for colonoscopy. *Cochrane Database Syst Rev* 2015(5):CD009863. doi: 10.1002/14651858.CD009863.pub2
25. Birch DW, Dang JT, Switzer NJ, et al. Heated insufflation with or without humidification for laparoscopic abdominal surgery. *Cochrane Database Syst Rev* 2016;10:CD007821. doi: 10.1002/14651858.CD007821.pub3
26. Hofstede SN, Nouta KA, Jacobs W, et al. Mobile bearing vs fixed bearing prostheses for posterior cruciate retaining total knee arthroplasty for postoperative functional status in patients with osteoarthritis and rheumatoid arthritis. *Cochrane Database Syst Rev* 2015(2):CD003130. doi: 10.1002/14651858.CD003130.pub3
27. Marti-Carvajal AJ, Gluud C, Nicola S, et al. Growth factors for treating diabetic foot ulcers. *Cochrane Database Syst Rev* 2015(10):CD008548. doi: 10.1002/14651858.CD008548.pub2
28. Akl EA, Kahale LA, Ballout RA, et al. Parenteral anticoagulation in ambulatory patients with cancer. *Cochrane Database Syst Rev* 2014(12):CD006652. doi: 10.1002/14651858.CD006652.pub4
29. Buppasiri P, Lumbiganon P, Thinkhamrop J, et al. Calcium supplementation (other than for preventing or treating hypertension) for improving pregnancy and infant outcomes. *Cochrane Database Syst Rev* 2015(2):CD007079. doi: 10.1002/14651858.CD007079.pub3
30. Hnin K, Nguyen C, Carson KV, et al. Prolonged antibiotics for non-cystic fibrosis bronchiectasis in children and adults. *Cochrane Database Syst Rev* 2015(8):CD001392. doi: 10.1002/14651858.CD001392.pub3

31. Lane R, Ellis B, Watson L, et al. Exercise for intermittent claudication. *Cochrane Database Syst Rev* 2014(7):CD000990. doi: 10.1002/14651858.CD000990.pub3
32. Bennett S, Pigott A, Beller EM, et al. Educational interventions for the management of cancer-related fatigue in adults. *Cochrane Database Syst Rev* 2016;11:CD008144. doi: 10.1002/14651858.CD008144.pub2
33. Song H, Zhu J, Lu D. Molecular-targeted first-line therapy for advanced gastric cancer. *Cochrane Database Syst Rev* 2016;7:CD011461. doi: 10.1002/14651858.CD011461.pub2
34. Loudon K, Treweek S, Sullivan F, et al. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;350:h2147. doi: 10.1136/bmj.h2147
35. Zwarenstein M, Treweek S, Loudon K. PRECIS-2 helps researchers design more applicable RCTs while CONSORT Extension for Pragmatic Trials helps knowledge users decide whether to apply them. *J Clin Epidemiol* 2017 doi: 10.1016/j.jclinepi.2016.10.010
36. Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309(6954):597-9.
37. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol* 2011;64(12):1294-302. doi: 10.1016/j.jclinepi.2011.03.017
38. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010;63(8):834-40. doi: 10.1016/j.jclinepi.2010.02.005
39. Smith R. The Cochrane collaboration at 20. *BMJ* 2013;347:f7383. doi: 10.1136/bmj.f7383

Appendix 1: Characteristics of primary RCTs according to systematic review

Characteristics of primary RCTs in the systematic review by Hafner S, et al. 2015 (N=16 articles)

Title of Systematic Review: Water infusion versus air sufflation for colonoscopy Authors: Hafner S, Zolk K, Radaelli F, Otte J, Rabenstein T, Zolk O Year of Publication: 2015						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Water Infusion**	**Air Insufflation**	**Risk Ratio with 95% CI**	**Risk of Bias**
Pohl J, et al.	2011	1	58	58	0.86 (0.75, 0.97)	High
Hsieh YH, et al.	2011	1	102	51	0.88 (0.78, 1.00)	Unclear
Ramirez FC, et al.	2011	1	177	191	0.92 (0.88, 0.96)	High
Cadoni S, et al.	2014	2	410	406	0.98 (0.95, 1.00)	High
Amato A, et al.	2013	1	113	113	0.98 (0.95, 1.02)	Low
Radaelli F, et al.	2010	1	116	114	0.98 (0.92, 1.04)	Low
Leung CW, et al.	2010	1	114	115	1.00 (0.85, 1.17)	Unclear
Leung JW, et al.	2009	1	28	28	1.00 (0.93, 1.07)	Unclear
Portocarrero DJ, et al.	2012	1	11	12	1.00 (0.96, 1.04)	High
Leung J, et al.	2011	1	50	50	1.00 (0.98, 1.02)	Unclear
Bayupurnama P, et al.	2013	3	53	57	1.03 (0.92, 1.16)	Unclear
Hsieh YH, et al.	2013	1	90	90	1.08 (0.96, 1.22)	Unclear
Leung JW, et al.	2013	1	50	50	1.09 (0.97, 1.23)	Unclear
Luo H, et al.	2013	1	55	55	1.21 (1.03, 1.43)	Low
Falt P, et al.	2012	1	102	107	1.22 (1.05, 1.43)	Low
Leung FW, et al.	2010	1	42	40	1.26 (1.06, 1.50)	Unclear

CI: confidence interval

Characteristics of primary RCTs in the systematic review by Martí-Carvajal AJ, et al. 2015 (N=11 articles*)

Title of Systematic Review: Growth factors for treating diabetic foot ulcers
Authors: Martí-Carvajal AJ, Gluud C, Nicola S, Simancas-Racines D, Oliva P, Cedeño-Taborda J
Year of Publication: 2015

Primary Author, et al.	Year of Publication	Number of Sites	Growth Factor	Usual Care Alone or with Placebo	Risk Ratio with 95% CI	Risk of Bias
d'Hemecourt PA, et al.	1998	10	34	68	2.00 (1.11, 3.59)	High
Hanft JR, et al.	2008	9	29	26	1.49 (0.79, 2.82)	High
Hardiker JV, et al.	2005	8	55	58	2.28 (1.50, 3.48)	High
Holloway GA, et al.	1993	4	49	21	2.21 (1.09, 4.50)	High
Jaiswal SS, et al.	2010	1	25	25	0.83 (0.56, 1.25)	High
Richard JL, et al.	1995	2	9	8	0.53 (0.18, 1.55)	Unclear
Saldalamacchia G, et al.	2004	1	7	7	2.00 (0.23, 17.34)	High
Steed D, et al.	1992	2	7	6	4.29 (0.67, 27.24)	High
Steed D, et al.	1995	10	61	57	1.94 (1.14, 3.27)	High
Viswanathan V, et al.	2006	3	29	28	1.72 (1.16, 2.57)	High
Wieman TJ, et al.	1998	23	255	127	1.23 (0.93, 1.63)	High

Characteristics of primary RCTs in the systematic review by Akl EA, et al. 2014 (N=11 articles)

Title of Systematic Review: Parenteral anticoagulation in ambulatory patients with cancer **Authors: Akl E, Kahale LA, Ballout RA, Barba M, Yosuico VED, van Doormaal FF, Middeldorp S, Bryant A, Schünemann H** **Year of Publication: 2014**						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Parenteral Anticoagulant**	**Usual Care or Placebo**	**Hazard Ratio with 95% CI**	**Risk of Bias**
Agnelli G, et al.	2012	395	1608	1604	0.96 (0.87, 1.06)	Low
Altinbas M, et al.	2004	1	42	42	0.52 (0.33, 0.82)	High
Kakker AK, et al.	2004	10	196	189	0.79 (0.63, 0.98)	Low
Klerk CP, et al.	2005	9	148	154	0.75 (0.60, 0.94)	Low
Lebeau B, et al.	1994	27	138	139	0.72 (0.56, 0.91)	High
Lecumberri R, et al.	2013	10	20	18	0.34 (0.14, 0.81)	High
Maraveyas A, et al.	2012	7	60	63	1.08 (0.74, 1.57)	Low
Perry JR, et al.	2010	15	99	87	1.20 (0.72, 2.0)	High
Sideras K, et al.	2006	7	71	70	1.15 (0.79, 1.68)	High
van Doormaal FF, et al.	2011	3	244	259	0.94 (0.75, 1.18)	High
Weber C, et al.	2008	1	10	10	0.64 (0.31, 1.33)	Low

CI: Confidence Interval

Characteristics of primary RCTs in the systematic review by Buppasiri P, et al. 2015 (N=12 articles*)

Title of Systematic Review: Calcium supplementation (other than for preventing or treating hypertension) for improving pregnancy and infant outcomes* Authors: Buppasiri P, Lumbiganon P, Thinkhamrop J, Ngamjarus C, Laopaiboon M, Medley N Year of Publication: 2015						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Calcium Supplement**	**Usual Care or Placebo**	**Risk Ratio with 95% CI**	**Risk of Bias**
Purwar M, et al.	1996	1	103	98	0.32 (0.07, 1.54)	Low
Lopez-Jaramillo P, et al.	1989	1	55	51	Not estimable	Low
Boggess KA, et al.	1997	1	12	11	0.13 (0.01, 2.30)	Low
Belizan JM, et al.	1991	3	593	601	0.91 (0.58, 1.43)	Low
Villar J, et al.	2006	7	4157	4168	0.91 (0.80, 1.04)	Low
Sanchez-Ramos L, et al.	1994	1	33	34	0.73 (0.27, 1.99)	Low
Sanchez-Ramos L, et al.	1995	1	36	39	1.03 (0.65, 1.62)	Low
Levine RJ, et al.	1997	5	2295	2294	1.08 (0.91, 1.28)	Low
Villar J, et al.	1990	1	95	95	0.35 (0.16, 0.80)	Low
Crowther C, et al.	1999	5	227	229	0.44 (0.21, 0.90)	High
Kumar A, et al.	2009	1	290	262	0.55 (0.32, 0.94)	Low
Taherian AA, et al.	2002	1	330	330	1.55 (1.00, 2.41)	High

CI: Confidence Interval; *missing 1 primary RCT

Characteristics of primary RCTs in the systematic review by Hnin K, et al. 2015 (N=13 articles)

Title of Systematic Review: Prolonged antibiotics for non-cystic fibrosis bronchiectasis in children and adults **Authors: Hnin K, Nguyen C, Carson KV, Evans DJ, Greenstone M, Smith BJ** **Year of Publication: 2015**						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Prolonged Antibiotics**	**Usual Care or Placebo**	**Odds Ratio with 95% CI**	**Risk of Bias**
Altenburg J, et al.	2013	14	45	44	0.20 (0.09, 0.46)	Low
Barker AF, et al.	2000	16	37	37	5.62 (0.62, 50.7)	Unclear
Currie DC, et al.	1990	2	19	19	0.18 (0.01, 4.0)	Low
De Diego A, et al.	2013	1	16	14	0.08 (0.02, 0.33)	High
Koh YY, et al.	1997	1	13	12	0.16 (0.01, 3.6)	Unclear
Liu JF, et al.	2012	10	25	25	0.28 (0.03, 2.87)	Unclear
Murray M, et al.	2011	Not reported	32	33	0.13 (0.04, 0.42)	High
Serisier D, et al. (BLESS)	2013	1	59	58	0.49 (0.25, 0.95)	Low
Serisier D, et al. (ORBIT)	2013	11	20	22	0.24 (0.05, 1.12)	High
Tsang KW, et al.	1999	Not reported	14	10	0.15 (0.01, 3.5)	Unclear
Valery P, et al.	2013	Not reported	45	44	0.40 (0.11, 1.41)	Low
Wilson R, et al.	2013	35	60	64	0.90 (0.44, 1.87)	High
Wong C, et al.	2012	3	71	70	0.23 (0.12, 0.47)	Low

CI: Confidence Interval

Characteristics of primary RCTs in the systematic review by Birch DW, et al. 2016 (N=18 articles*)

Title of Systematic Review: Heated insufflation with or without humidification for laparoscopic abdominal surgery* Authors: Birch DW, Dang JT, Switzer NJ, Manouchehri N, Shi X, Hadi G, Karmali S Year of Publication: 2016						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Heated Insufflation**	**Cold Insufflation**	**Mean Difference with 95% CI**	**Risk of Bias**
Backlund M, et al.	1998	1	13	13	0.30 (-0.18, 0.78)	Unclear
Champion JK, et al.	2006	1	25	25	0.00 (-0.35, 0.35)	Low
Davis SS, et al. (A)	2006	1	11	11	0.00 (-0.53, 0.53)	Low
Davis SS, et al. (B)	2006	1	11	11	0.20 (-0.16, 0.56)	Low
Farley DR, et al.	2004	1	49	52	0.32 (0.13, 0.51)	Low
Hamza MA, et al.	2005	1	23	23	1.00 (0.63, 1.37)	Low
Kissler S, et al. (A)	2004	1	17	19	-0.10 (-0.51, 0.31)	Unclear
Kissler S, et al. (B)	2004	1	17	19	-0.20 (-0.48, 0.08)	Unclear
Manwaring JM, et al.	2008	1	30	30	-0.07 (-0.36, 0.22)	Low
Mouton WG, et al.	1999	1	20	20	0.05 (-0.39, 0.49)	Unclear
Nguyen NT, et al.	2002	1	10	10	0.10 (-0.45, 0.65)	Low
Ott DE, et al.	1998	7	25	25	0.18 (0.02, 0.34)	High
Sammour T, et al.	2010	3	41	41	0.16 (-0.10, 0.42)	Low
Savel RH, et al.	2005	1	15	15	0.70 (0.25, 1.15)	Low
Yu TC, et al.	2013	1	97	98	-0.04 (-0.14, 0.06)	Low
Lee KC, et al.	2011	1	15	15	0.30 (-0.01, 0.61)	Low
Nelskyla K, et al.	1999	1	18	19	-0.20 (-0.33, -0.07)	Unclear
Puttick MI, et al.	1999	3	15	15	0.18 (0.02, 0.34)	Unclear
Saad S, et al.	2000	2	10	10	0.10 (-0.28, 0.48)	Unclear
Willis VL, et al.	2001	1	19	21	0.20 (-0.06, 0.46)	Low

CI: Confidence Interval; *missing 1 primary RCT

Characteristics of primary RCTs in the systematic review by Lane R, et al. 2014 (N=12 articles)

Title of Systematic Review: Exercise for intermittent claudication **Authors: Lane R, Ellis B, Watson L, Leng GC** **Year of Publication: 2014**						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Randomized to Exercise**	**Usual Care or Placebo**	**Mean Difference with 95% CI**	**Risk of Bias**
Collins EG, et al.	2005	1	27	25	12.33 (-0.97, 25.6)	Low
Crowther RG, et al.	2012	1	11	11	6.77 (3.05, 10.5)	High
Hiatt WR, et al.	1990	1	10	9	6.80 (3.92, 9.68)	High
Hiatt WR, et al.	1994	1	10	8	7.40 (2.50, 12.3)	Unclear
McDermott MM, et al.	2008	1	51	53	4.40 (2.43, 6.37)	Low
McDermott MM, et al.,	2013	1	97	97	1.35 (-0.31, 3.01)	Low
Mika P, et al.	2006	1	30	30	3.41 (2.72, 4.10)	Unclear
Mika P, et al.	2011	1	34	34	6.08 (5.46, 6.70)	Low
Sanderson B, et al.	2006	3	13	14	-3.43 (-9.99, 3.13)	Low
Tsai JC, et al.	2002	2	32	32	4.90 (2.88, 6.92)	Low
Wood RE, et al.	2006	1	7	6	-2.55 (-15.3, 10.2)	Unclear
Larsen OA, et al.	1966	1	7	7	5.20 (0.88, 9.52)	High

CI: Confidence Interval

Characteristics of primary RCTs in the systematic review by Bennett S, et al. 2016 (N=12 articles)

Title of Systematic Review: Educational interventions for the management of cancer- related fatigue in adults **Authors: Bennett S, Pigott A, Beller EM, Haines T, Meredith P, Delaney C** **Year of Publication: 2016**						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Educational Intervention**	**Usual Care, Control or Alternative Fatigue Intervention**	**Standardized Mean Difference with 95% CI**	**Risk of Bias**
Reif K, et al.	2012	10	129	132	-1.05 (-1.32, -0.77)	High
Wangnum K, et al.	2013	1	30	30	-0.55 (-1.07, -0.04)	High
Yun YH, et al.	2012	4	136	137	-0.46 (-0.72, -0.21)	High
Yates P, et al.	2005	5	53	57	-0.39 (-0.77, 0.0)	High
Ream E, et al.	2006	2	48	55	-0.38 (-0.81, 0.05)	High
Purcell A, et al.	2011	1	82	28	-0.28 (-0.68, 0.13)	High
Godino C, et al.	2006	1	23	17	-0.12 (-1.03, 0.80)	High
Schjolberg T, et al.	2014	1	79	81	-0.06 (-0.50, 0.38)	High
Yuen HK, et al.	2006	1	6	6	-0.03 (-1.17, 1.10)	High
Barsevick AM, et al.	2004	2	200	196	0.0 (-0.20, 0.20)	High
Barsevick AM, et al.	2010	4	153	139	0.04 (-0.23, 0.31)	High
Foster C, et al.	2015	12	85	78	0.20 (-0.15, 0.55)	High

CI: Confidence Interval

Characteristics of primary RCTs in the systematic review by Song H, et al. 2016 (N=10 articles)

Title of Systematic Review: Molecular-targeted first-line therapy for advanced gastric cancer **Authors: Song H, Zhu J, Lu D** **Year of Publication: 2016**						
Primary Author, et al.	**Year of Publication**	**Number of Sites**	**Molecular Targeted Therapy**	**Usual Care**	**Hazard Ratio with 95% CI**	**Risk of Bias**
Bang YJ, et al.	2010	122	298	296	0.71 (0.59, 0.85)	High
Hecht JR, et al.	2013	186	272	273	0.91 (0.73, 1.13)	Low
Iveson T, et al.	2014	43	82	39	0.70 (0.45, 1.09)	Low
Koizumi W, et al.	2013	14	46	47	0.74 (0.46, 1.19)	High
Lordick F, et al.	2013	164	455	449	1.0 (0.86, 1.16)	High
Ohtsu A, et al.	2011	93	387	387	0.87 (0.73, 1.03)	Low
Rao S, et al.	2010	22	36	36	1.02 (0.61, 1.70)	High
Shen L, et al.	2015	14	100	102	1.11 (0.79, 1.56)	Low
Waddell T, et al.	2013	63	278	275	1.37 (1.07, 1.76)	High
Zhang ZD, et al.	2014	1	30	26	0.74 (0.42, 1.30)	High

CI: Confidence Interval

Characteristics of primary RCTs in the systematic review by Hofstede SN, et al. 2015 (N=13 articles*)

Title of Systematic Review: Mobile bearing vs fixed bearing prostheses for posterior cruciate retaining total knee arthroplasty for postoperative functional status in patients with osteoarthritis and rheumatoid arthritis
Authors: Hofstede SN, Nouta KA, Jacobs W, van Hooff ML, Wymenga AB, Pijls BG, Nelissen RGHH, Marang- van de Mheen PJ
Year of Publication: 2015

Primary Author, et al.	Year of Publication	Number of Sites	Mobile Bearing	Fixed Bearing	Mean Difference with 95% CI	Risk of Bias
Bailey O, et al.	2014	4	171	173	1.90 (-1.36, 5.16)	Low
Hanusch B, et al.	2010	Not reported	55	50	0.20 (-5.92, 6.32)	High
Henricson A, et al.	2006	1	26	26	-2.0 (-6.42, 2.42)	High
Jacobs WCH, et al.	2011	2	67	63	2.20 (-2.57, 6.97)	High
Kim YH, et al.	2001	1	120	120	-1.10 (-2.72, 0.51)	High
Kim YH, et al.	2007	1	194	194	1.00 (-0.79, 2.79)	High
Kim YH, et al. (A)	2009	1	92	92	-7.00 (-8.84, -5.16)	High
Kim YH, et al. (B)	2009	1	69	69	1.00 (-2.47, 4.47)	Unclear
Kim TK, et al.	2010	1	71	71	-2.90 (-6.27, 0.47)	Unclear
Lampe F, et al.	2011	1	52	48	-3.00 (-8.60, 2.60)	High
Munro JT, et al.	2010	2	26	25	1.00 (-2.43, 4.43)	Unclear
Price AJ, et al.	2003	4	19	21	-5.80 (-12.09, 0.49)	High
Watanabe T, et al.	2005	1	22	22	-0.70 (-4.78, 3.38)	Low

CI: Confidence Interval; *missing 1 primary RCT

Appendix 2: Mean scores and frequency of missing data by PRECIS-2 domain according to systematic review

Systematic Review Primary Author, et al.	Recruitment	Frequency (%) Missing Recruitment	Setting	Frequency (%) Missing Setting	Flexibility Adherence	Frequency (%) Missing Adherence	Follow Up	Frequency (%) Missing Follow Up	Primary Outcome	Frequency (%) Missing Outcome
Hafner S (16 articles)	3.6 (0.8)	2 (13)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Martí-Carvajal AJ (11 articles*)	3.0 (1.4)	9 (82)	0 (0)	0 (0)	0 (0)	0 (0)	3.5 (1.0)	1 (9)	0 (0)	0 (0)
Akl EA (11 articles)	3.0 (0.8)	7 (64)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4.5 (0.8)	1 (9)
Buppasiri P (12 articles*)	2.5 (1.4)	2 (17)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3.2 (0.8)	1 (8)
Hnin K (13 articles)	3.6 (1.0)	4 (31)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2.1 (1.4)	2 (15)
Birch DW (18 articles*)	3.8 (0.3)	9 (50)	0 (0)	0 (0)	3.7 (1.2)	3 (17)	0 (0)	0 (0)	0 (0)	0 (0)
Lane R (12 articles)	2.6 (1.0)	5 (42)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Bennett S (12 articles)	2.3 (1.0)	3 (25)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Song H (10 articles)	0 (0)	10 (100)	5 (0)	1 (10)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hofstede SN (13 articles*)	3.0 (1.1)	3 (23)	1.8 (1.3)	1 (8)	2.8 (0.9)	1 (8)	0 (0)	0 (0)	2.4 (1.2)	1 (8)

All data are described as mean (standard deviation) unless otherwise indicated; *1 primary RCT missing

Appendix 3: PRECIS-2 domain scores for primary RCTS according to systematic review (N=10)

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Hafner et al. 2015 (N=16 articles)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Pohl J, et al.	3	4	4	2	3	5	4	1	3	29
Hsieh YH, et al.	3	3	4	3	4	5	4	5	2	33
Ramirez FC, et al.	2	5	2	4	3	5	5	2	5	33
Cadoni S, et al.	3	4	4	2	3	5	4	4	3	32
Amato A, et al.	2	4	4	2	3	2	4	1	4	26
Radaelli F, et al.	3	4	4	2	4	5	4	3	5	34
Leung CW, et al.	2	3	1	3	4	5	2	3	4	27
Leung JW, et al.	2	2	1	2	4	5	4	4	3	27
Portocarrero DJ, et al.	4	4	3	2	4	5	4	3	3	32
Leung J, et al.	2	2	1	2	4	3	4	1	5	24
Bayupurnama P, et al.	4	4	4	3	4	5	4	5	3	36
Hsieh YH, et al.	4	4	4	3	4	5	4	5	3	36
Leung JW, et al.	2	3	1	2	3	3	4	1	5	24
Luo H, et al.	2	4	2	1	3	5	4	1	5	27
Falt P, et al.	4	4	1	1	4	5	4	3	2	28
Leung FW, et al.	2	3	1	2	3	5	4	5	5	30

PRECIS-2 scores by domain for primary RCTs included in the systematic review by by Martí-Carvajal AJ, et al. 2015 (N=11 articles*)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
d'Hemecourt PA, et al.	2	3	5	2	2	4	4	5	5	32
Hanft JR, et al.	1	3	5	2	3	4	2	1	3	24
Hardiker JV, et al.	2	3	5	4	3	1	4	5	2	29
Holloway GA, et al.	2	3	3	2	2	1	4	5	1	23
Jaiswal SS, et al.	2	3	1	4	2	3	4	5	3	27
Richard JL, et al.	1	3	2	1	1	2	1	3	3	17
Saldalamacchia G, et al.	3	4	1	4	3	4	4	5	3	31
Steed D, et al.	2	2	2	2	2	3	4	5	3	25
Steed D, et al.	1	3	5	2	2	4	4	3	5	29
Viswanathan V, et al.	1	3	4	2	1	2	3	5	1	22
Wieman TJ, et al.	1	3	5	2	3	3	4	3	5	29

*1 primary RCT missing

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Akl EA, et al. 2014 (N=11 articles)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Agnelli G, et al.	4	3	5	1	2	4	4	4	5	32
Altinbas M, et al.	2	3	1	3	3	3	4	5	4	28
Kakker AK, et al.	4	3	5	1	2	3	5	5	3	31
Klerk CP, et al.	3	3	5	1	2	4	3	5	5	31
Lebeau B, et al.	4	2	5	2	1	3	1	5	5	28
Lecumberri R, et al.	2	3	5	3	3	3	4	5	4	32
Maraveyas A, et al.	2	3	5	3	2	2	2	3	3	25
Perry JR, et al.	2	3	5	1	1	2	3	3	5	25
Sideras K, et al.	4	3	5	1	1	3	2	5	3	27
van Doormaal FF, et al.	3	3	4	2	3	4	4	5	5	33
Weber C, et al.	2	4	1	3	3	3	3	3	3	25

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Buppasiri P, et al. 2015 (N=12 articles*)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Purwar M, et al.	2	3	1	1	3	1	4	4	2	21
Lopez-Jaramillo P, et al.	1	2	1	1	2	1	3	3	1	15
Boggess KA, et al.	3	4	1	1	1	1	2	2	3	18
Belizan JM, et al.	3	3	4	1	1	1	2	3	2	20
Villar J, et al.	4	4	2	1	2	2	4	3	5	27
Sanchez-Ramos L, et al.	3	1	1	1	3	2	4	3	3	21
Sanchez-Ramos L, et al.	1	2	1	1	2	2	2	3	3	17
Levine RJ, et al.	1	1	5	1	2	3	3	3	2	21
Villar J, et al.	2	1	1	1	1	1	4	5	3	19
Crowther C, et al.	4	4	5	1	2	2	4	3	5	30
Kumar A, et al.	4	4	1	1	3	1	3	3	1	21
Taherian AA, et al.	4	2	3	2	3	4	5	3	3	29

*1 primary RCT missing

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Hnin K, et al. 2015 (N=13 articles)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Altenburg J, et al.	2	4	5	1	1	2	1	4	4	24
Barker AF, et al.	2	3	5	1	1	2	1	1	3	19
Currie DC, et al.	1	4	3	1	1	2	2	2	2	18
De Diego A, et al.	2	4	1	2	2	3	3	1	1	19
Koh YY, et al.	2	4	1	1	2	1	2	1	1	15
Liu JF, et al.	4	3	5	2	2	3	3	3	3	28
Murray M, et al.	2	4	3	2	1	1	1	1	1	16
Serisier D, et al. (BLESS)	2	1	1	1	1	2	1	3	5	17
Serisier D, et al. (ORBIT)	2	3	5	1	1	3	2	1	4	22
Tsang KW, et al.	1	4	1	1	2	1	5	2	1	18
Valery P, et al.	1	4	1	1	1	2	2	4	5	21
Wilson R, et al.	2	3	5	1	2	3	1	1	4	22
Wong C, et al.	4	3	4	1	2	2	1	4	5	26
Altenburg J, et al.	2	4	5	1	1	2	1	4	4	24
Barker AF, et al.	2	3	5	1	1	2	1	1	3	19
Currie DC, et al.	1	4	3	1	1	2	2	2	2	18

*1 primary RCT missing

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Birch DW, et al. 2016 (N=18 articles*)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Backlund M, et al.	3	3	1	2	2	4	3	1	3	22
Champion JK, et al.	2	4	1	2	3	4	4	3	3	26
Davis SS, et al. (A)	4	4	1	2	2	1	2	1	3	20
Davis SS, et al. (B)	4	4	1	1	2	3	3	1	2	21
Farley DR, et al.	3	3	1	2	3	4	3	3	2	24
Hamza MA, et al.	4	4	1	3	2	3	2	4	3	26
Kissler S, et al. (A)	2	3	1	3	3	2	4	4	5	27
Kissler S, et al. (B)	4	4	1	2	3	5	3	3	2	27
Manwaring JM, et al.	2	4	1	3	3	5	4	3	3	28
Mouton WG, et al.	4	3	5	4	5	4	4	5	2	36
Nguyen NT, et al.	2	3	4	3	3	4	3	2	4	28
Ott DE, et al.	4	4	1	1	3	5	4	5	4	31
Sammour T, et al.	1	3	1	1	3	4	3	2	3	21
Savel RH, et al.	3	3	1	2	3	5	4	1	3	25
Yu TC, et al.	1	3	1	2	2	4	4	1	2	20
Lee KC, et al.	2	3	2	3	3	3	4	2	3	25
Nelskyla K, et al.	5	4	2	4	4	3	4	3	3	32
Puttick MI, et al.	2	3	1	2	2	2	2	3	2	19

*1 primary RCT missing

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Lane R, et al. 2014 (N=12 articles)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Collins EG, et al.	1	3	1	2	1	2	1	5	3	19
Crowther RG, et al.	3	3	1	2	1	2	1	3	2	18
Hiatt WR, et al.	1	3	1	2	1	2	1	4	2	17
Hiatt WR, et al.	1	3	1	2	1	2	1	5	2	18
McDermott MM, et al.	1	2	4	1	1	2	1	5	4	21
McDermott MM, et al.,	1	1	4	1	2	2	3	5	4	23
Mika P, et al.	4	3	1	1	1	3	2	2	2	19
Mika P, et al.	2	3	1	1	1	3	2	2	2	17
Sanderson B, et al.	4	4	2	3	1	3	1	5	3	26
Tsai JC, et al.	4	3	1	1	1	3	2	5	4	24
Wood RE, et al.	3	3	1	2	1	2	1	1	3	17
Larsen OA, et al.	2	2	1	4	3	3	2	3	3	23

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Bennett S, et al. 2016 (N=12 articles)

Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Reif K, et al.	3	2	5	1	1	2	2	5	3	24
Wangnum K, et al.	1	3	1	2	1	2	2	5	3	20
Yun YH, et al.	1	2	4	2	4	4	4	5	5	31
Yates P, et al.	3	4	5	1	1	3	4	5	2	28
Ream E, et al.	3	4	4	1	2	1	2	5	3	25
Purcell A, et al.	3	2	1	1	1	2	4	5	2	21
Godino C, et al.	2	3	1	1	1	2	3	4	2	19
Schjolberg T, et al.	3	2	1	2	2	1	3	5	2	21
Yuen HK, et al.	3	1	1	1	2	1	2	5	3	19
Barsevick AM, et al.	3	3	4	1	1	1	2	5	3	23
Barsevick AM, et al.	3	2	4	1	1	1	2	3	3	20
Foster C, et al.	2	2	5	2	3	3	4	3	4	28

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Song H, et al. 2016 (N=10 articles)

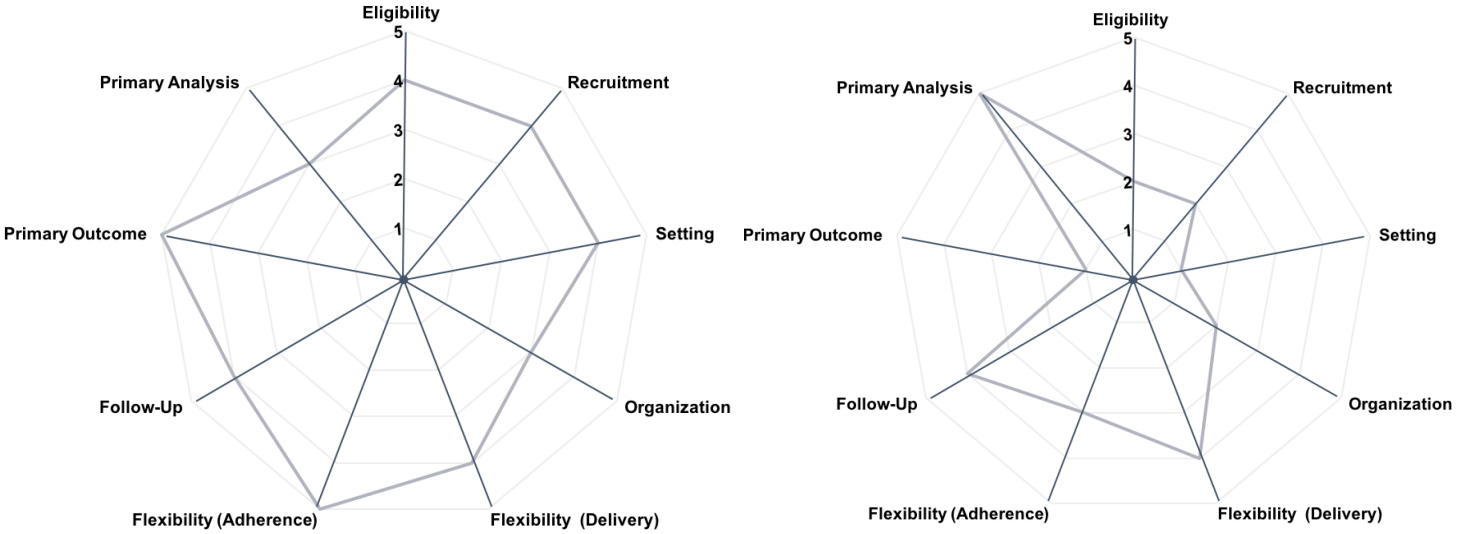
Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Bang YJ, et al.	2	3	5	3	1	2	2	5	3	26
Hecht JR, et al.	2	3	5	1	1	1	2	5	4	24
Iveson T, et al.	2	3	5	1	1	4	1	5	5	27
Koizumi W, et al.	1	3	5	2	2	4	2	5	4	28
Lordick F, et al.	1	3	5	2	3	4	2	5	5	30
Ohtsu A, et al.	2	3	5	1	1	4	1	5	5	27
Rao S, et al.	1	3	5	2	1	4	1	1	5	23
Shen L, et al.	2	3	5	1	1	4	1	5	5	27
Waddell T, et al.	2	3	5	3	1	3	1	5	4	27
Zhang ZD, et al.	2	3	3	3	3	3	2	5	3	27

PRECIS-2 scores by domain for primary RCTs included in the systematic review by Hofstede SN, et al. 2015 (N=13 articles*)

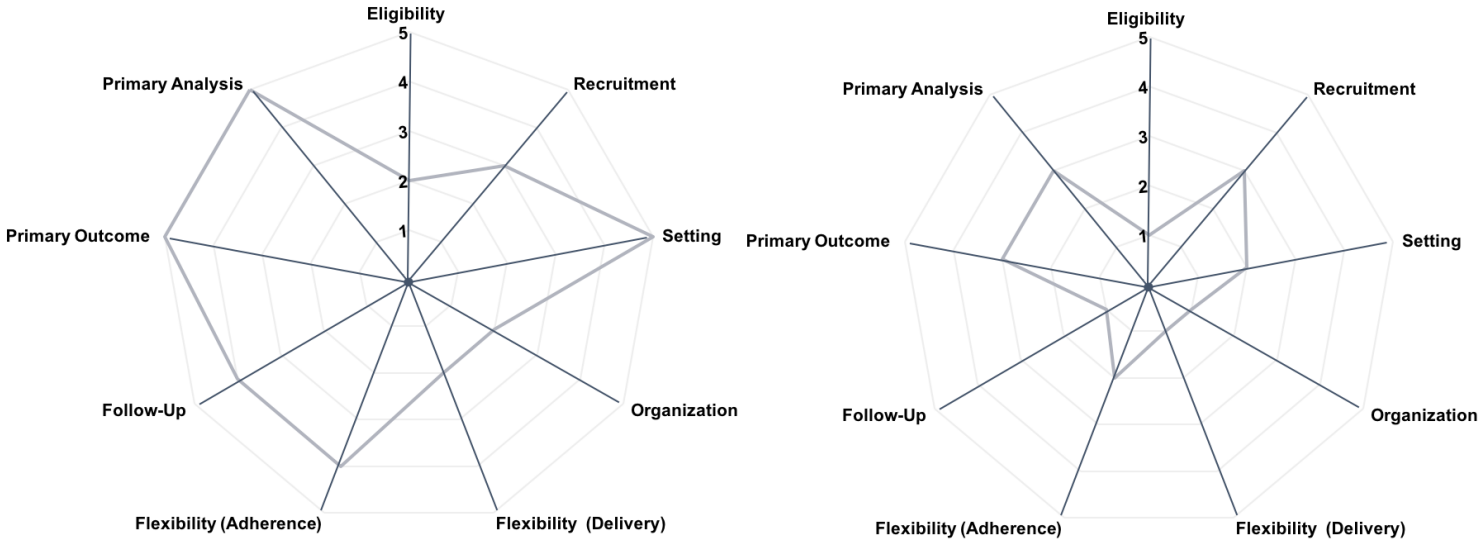
Primary Author, et al.	Eligibility	Recruitment	Setting	Organization	Flexibility Delivery	Flexibility Adherence	Follow Up	Primary Outcome	Primary Analysis	PRECIS-2 Summary Score
Bailey O, et al.	4	2	4	2	4	3	4	3	1	27
Hanusch B, et al.	3	2	3	2	2	2	4	4	1	23
Henricson A, et al.	2	4	1	2	2	4	3	1	2	21
Jacobs WCH, et al.	1	2	3	2	3	2	3	1	1	18
Kim YH, et al.	4	4	1	1	3	3	1	3	3	23
Kim YH, et al.	4	4	1	1	3	2	1	4	2	22
Kim YH, et al. (A)	2	4	1	1	3	2	1	3	2	19
Kim YH, et al. (B)	3	4	1	1	3	2	1	1	2	18
Kim TK, et al.	3	3	1	1	3	2	3	3	2	21
Lampe F, et al.	3	2	1	1	4	4	4	2	3	24
Munro JT, et al.	1	3	3	1	1	3	3	1	3	19
Price AJ, et al.	4	2	4	1	3	4	4	3	5	30
Watanabe T, et al.	4	3	1	3	4	3	3	3	3	27

*1 primary RCT missing

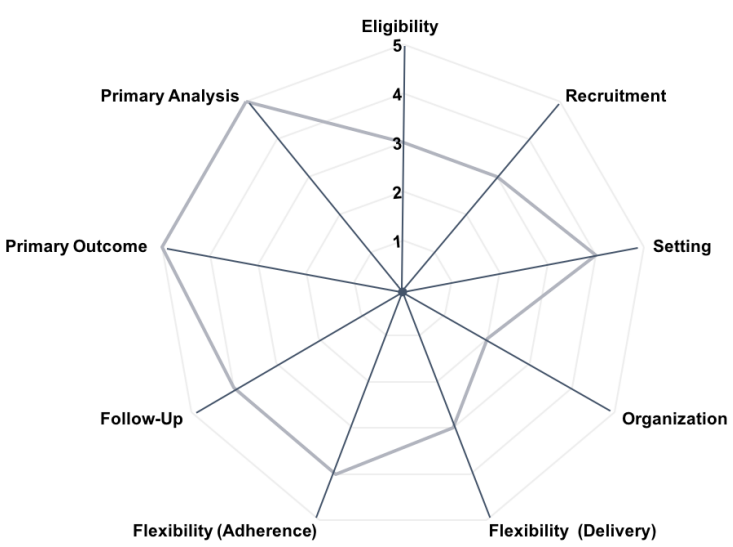
Appendix 4: PRECIS-2 Wheels of the Maximum Pragmatic and Explanatory Trials According to Systematic Review

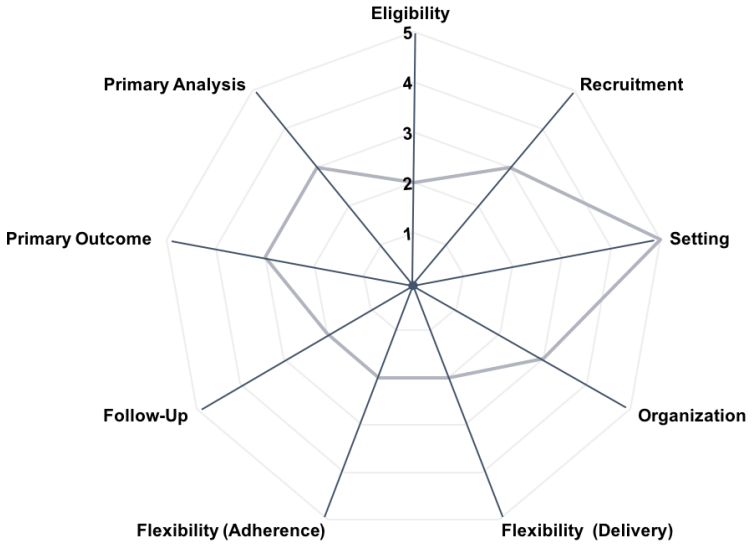


The maximum pragmatic primary study (score=36; left) and explanatory primary study (score=24; right) included in the systematic review by Hafner et al.

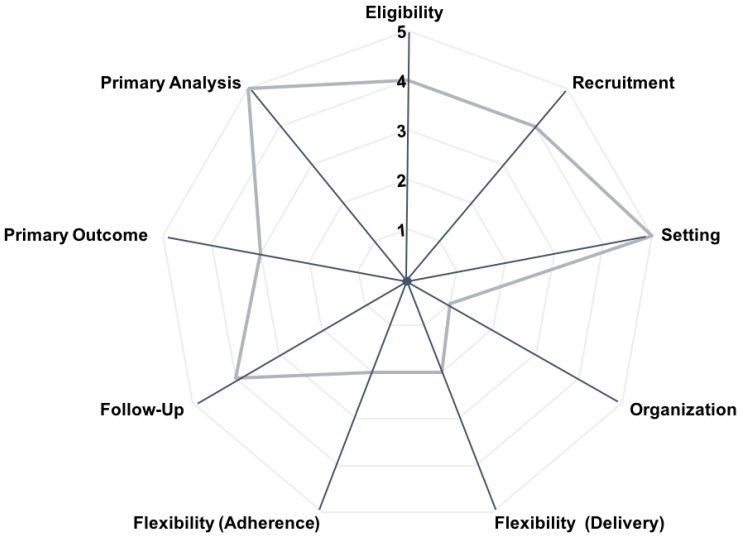


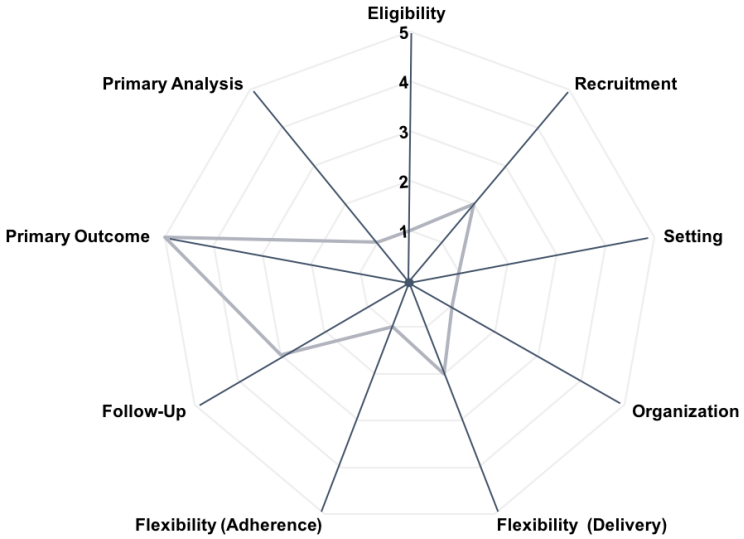
The maximum pragmatic primary study (score=32; left) and explanatory primary study (score=17; right) included in the systematic review by Martí-Carvajal et al.



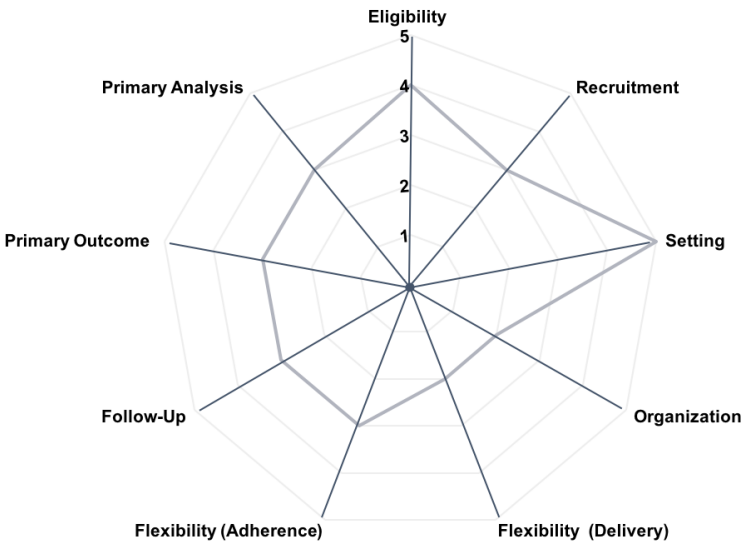


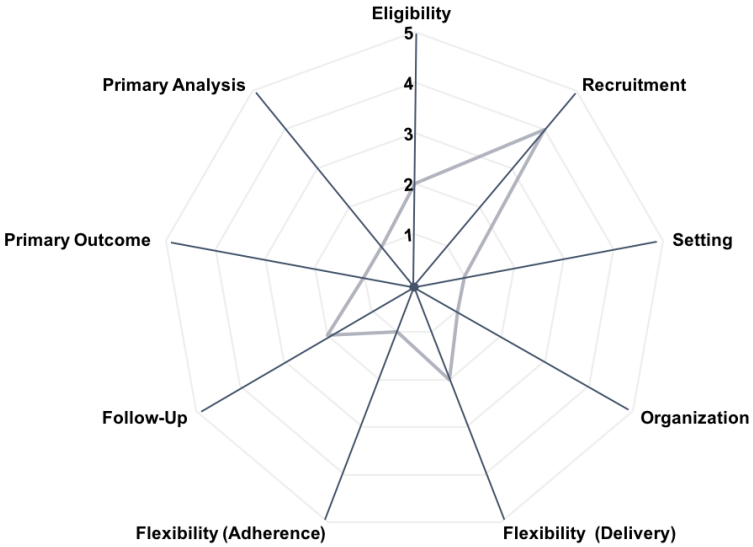
The maximum pragmatic primary study (score=33; left) and explanatory primary study (score=25; right) included in the systematic review by Akl et al.



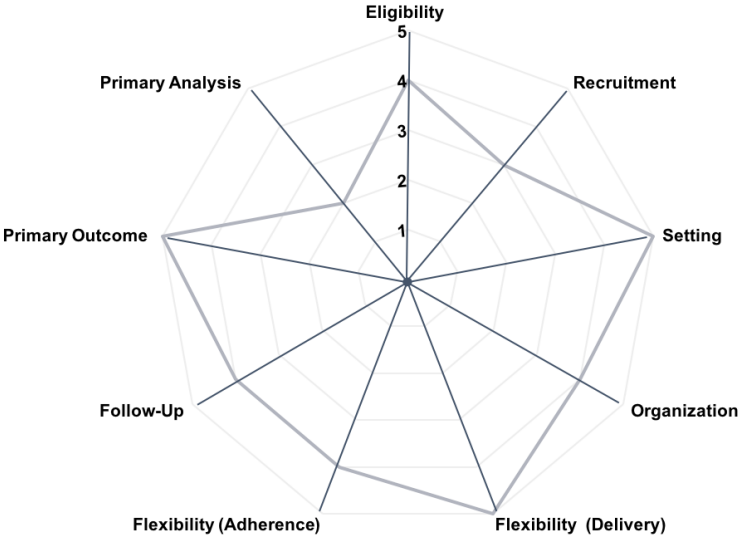


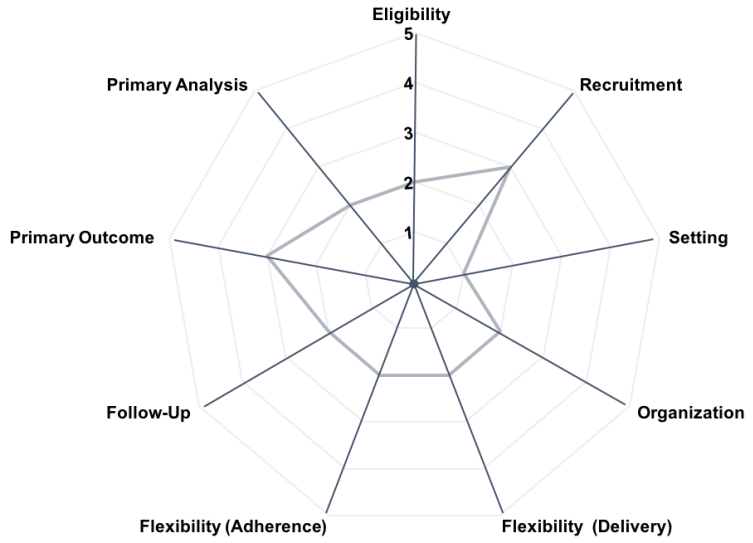
The maximum pragmatic primary study (score=30; left) and explanatory primary study (score=15; right) included in the systematic review by Buppasiri et al.



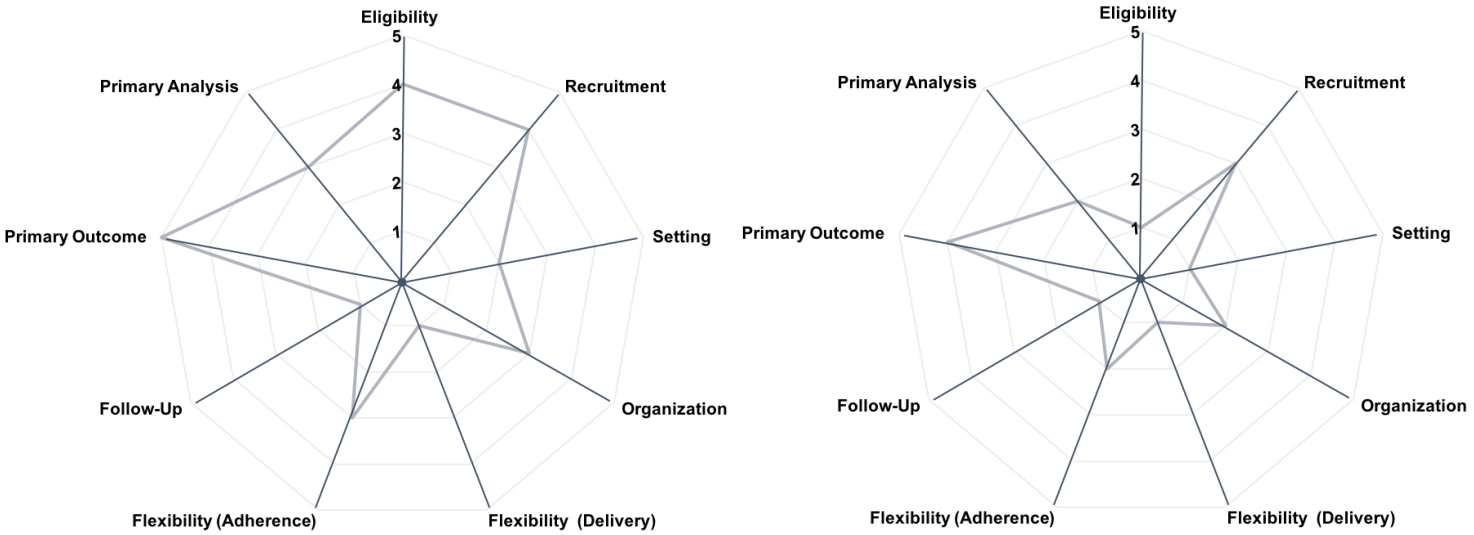


The maximum pragmatic primary study (score=28; left) and explanatory primary study (score=15; right) included in the systematic review by Hnin et al.

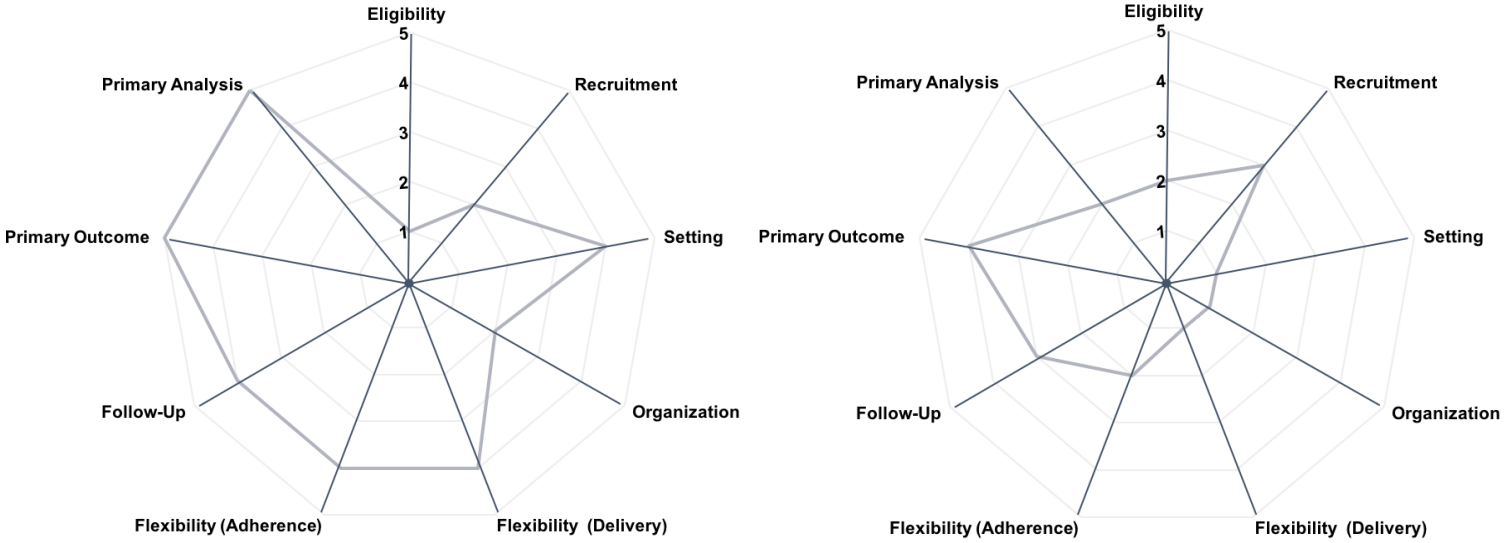




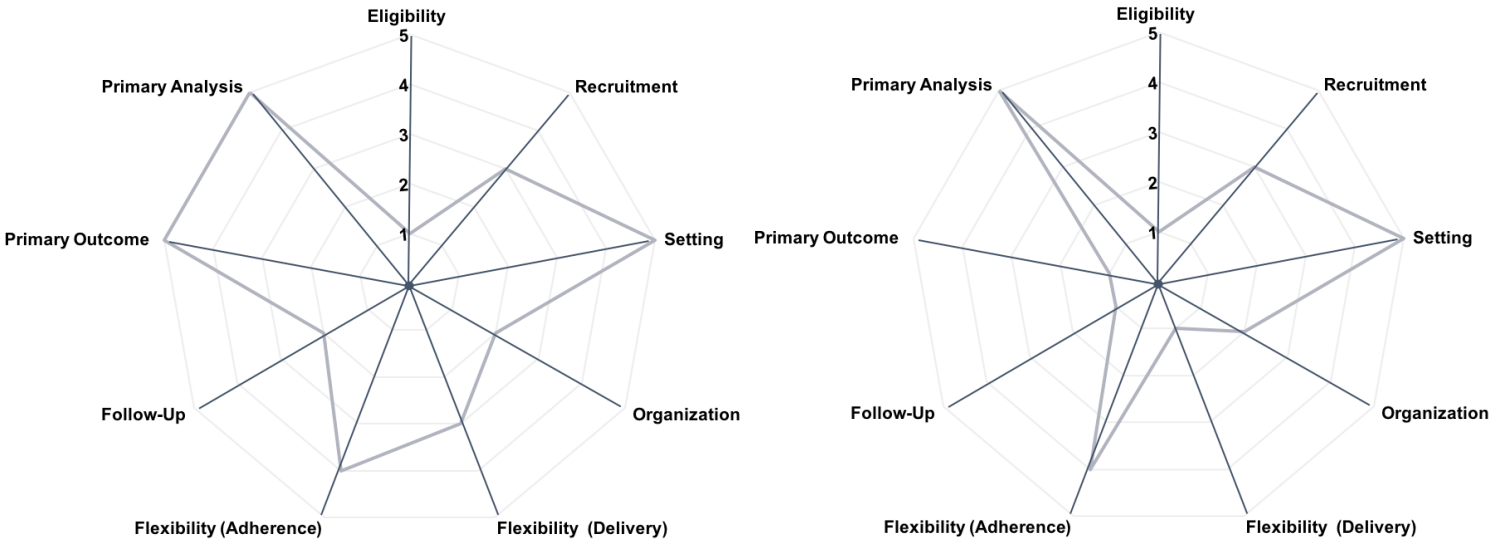
The maximum pragmatic primary study (score=36; left) and explanatory primary study (score=19; right) included in the systematic review by Birch et al.



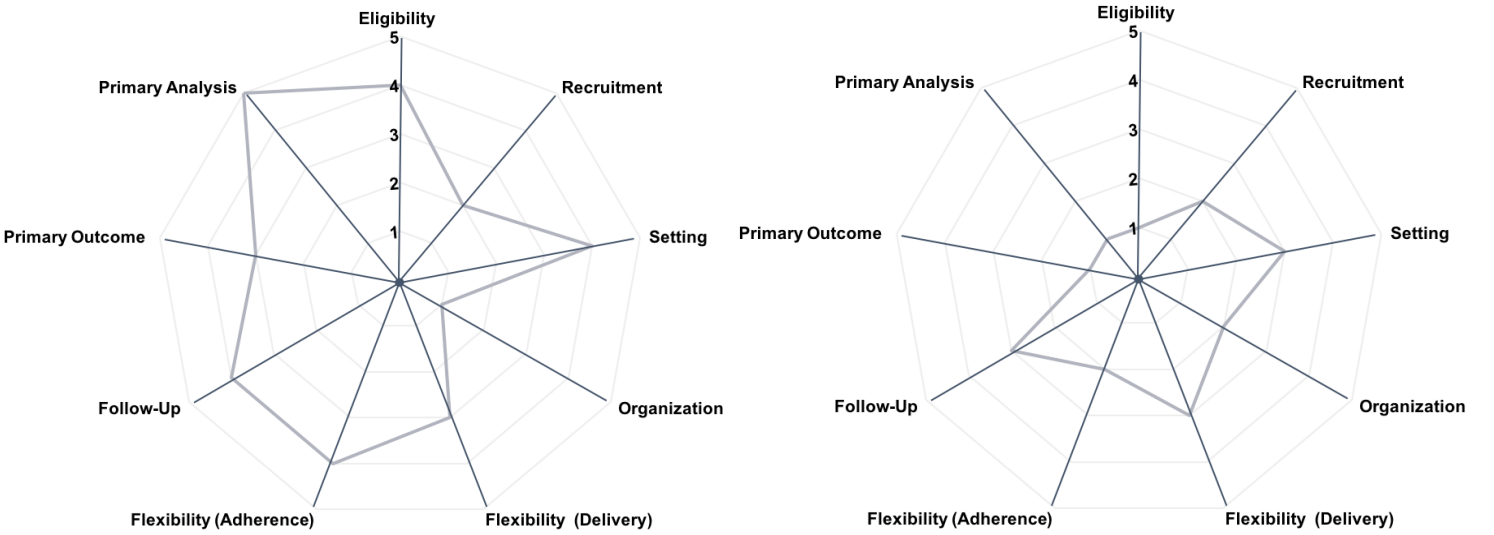
The maximum pragmatic primary study (score=26; left) and explanatory primary study (score=17; right) included in the systematic review by Lane et al.



The maximum pragmatic primary study (score=31; left) and explanatory primary study (score=19; right) included in the systematic review by Bennett et al.



The maximum pragmatic primary study (score=30; left) and explanatory primary study (score=23; right) included in the systematic review by Song et al.



The maximum pragmatic primary study (score=30; left) and explanatory primary study (score=18; right) included in the systematic review by Hofstede et al.

Appendix 5: Inter-Rater Reliability of PRECIS-2 by Systematic Review

Inter-rater reliability of raters 1 and 2 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Hafner S, et al. 2015 (N=16 articles[†])

Domain	ICC	95% CI
Eligibility	0.42*	-0.70, 0.80
Recruitment	0.68	0.08, 0.89
Setting	0.94	0.82, 0.98
Organization	0.15*	-1.61, 0.71
Flexibility: Delivery	0.48*	-0.30, 0.81
Flexibility: Adherence	0*	-1.87, 0.65
Follow-Up	0.66	0.08, 0.88
Primary Outcome	0.75	0.21, 0.92
Primary Analysis	0.47	-0.63, 0.82
Summary Score	0.42*	-0.75, 0.80

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 1: KA, Rater 2: TA; [†]Missing 1 primary RCT; *p>0.05 not statistically significant

Inter-rater reliability of raters 1 and 2 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Martí-Carvajal AJ, et al. 2015 (N=11 articles[†])

Domain	ICC	95% CI
Eligibility	0.49*	-0.35, 0.84
Recruitment	0*	-3.23, 0.67
Setting	0.94	0.81, 0.99
Organization	0.27*	-1.84, 0.81
Flexibility: Delivery	-1.05*	-5.53, 0.43
Flexibility: Adherence	0.53*	-0.56, 0.87
Follow-Up	0.07*	-0.14, 0.44
Primary Outcome	0.89	0.57, 0.97
Primary Analysis	0.72	0.07, 0.92
Summary Score	0.21*	-0.78, 0.74

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 1: KA, Rater 2: TA; [†]Missing 1 primary RCT; *p>0.05 not statistically significant

Inter-rater reliability of raters 1 and 2 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Akl EA, et al. 2014 (N=11 articles)

Domain	ICC	95% CI
Eligibility	0.50*	-0.36, 0.85
Recruitment	0.17*	-1.47, 0.76
Setting	0.95	0.83, 0.99
Organization	-0.08*	-0.61, 0.52
Flexibility: Delivery	-0.29*	-1.14, 0.50
Flexibility: Adherence	-0.58*	-0.63, 0.60
Follow-Up	0.24	-0.82, 0.76
Primary Outcome	0.55	-0.86, 0.88
Primary Analysis	0.37	-0.73, 0.81
Summary Score	0.47*	-0.41, 0.84

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability of raters 1 and 2 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Buppasiri P, et al. 2015 (N=12 articles[†])

Domain	ICC	95% CI
Eligibility	0.16*	-0.28, 0.62
Recruitment	0.15*	-0.15, 0.56
Setting	0.94	0.79, 0.98
Organization	-0.10	-0.56, 0.45
Flexibility: Delivery	0.13*	-1.67, 0.74
Flexibility: Adherence	0.53*	-0.26, 0.85
Follow-Up	0.07*	-1.81, 0.72
Primary Outcome	0.20*	-0.16, 0.63
Primary Analysis	0.53*	-0.37, 0.86
Summary Score	0.24*	-0.64, 0.74

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 1: KA, Rater 2: TA; [†]Missing 1 primary RCT; *p>0.05 not statistically significant

Inter-rater reliability of raters 1 and 2 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Hnin K, et al. 2015 (N=13 articles)

Domain	ICC	95% CI
Eligibility	0.10*	-0.55, 0.62
Recruitment	0.43*	-0.40, 0.81
Setting	0.74	0.10, 0.92
Organization	0*	-0.05, 0.12
Flexibility: Delivery	0*	-0.21, -0.36
Flexibility: Adherence	0.49*	-0.74, 0.85
Follow-Up	0.30*	-0.29, 0.72
Primary Outcome	0.86	0.55, 0.96
Primary Analysis	0.94	0.81, 0.98
Summary Score	0.62	-0.19, 0.89

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability of raters 2 and 3 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Birch DW, et al. 2016 (N=18 articles[†])

Domain	ICC	95% CI
Eligibility	0.54	-0.10, 0.82
Recruitment	0.81	0.49, 0.93
Setting	0.87	0.50, 0.95
Organization	-0.12*	-1.17, 0.51
Flexibility: Delivery	-0.10*	-0.91, 0.47
Flexibility: Adherence	0.06*	-0.49, 0.53
Follow-Up	0.32*	-0.66, 0.74
Primary Outcome	0.79	0.43, 0.92
Primary Analysis	0.87	0.64, 0.95
Summary Score	0.58	-0.11, 0.84

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials l; Rater 2: TA, Rater 3: DL; [†]Missing 1 primary RCT; *p>0.05 not statistically significant

Inter-rater reliability of raters 2 and 3 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Lane R, et al. 2014 (N=12 articles)

Domain	ICC	95% CI
Eligibility	0.62	-0.18, 0.89
Recruitment	0.42	-1.10, 0.83
Setting	0*	-1.70, 0.69
Organization	-1.01*	-2.84, 0.40
Flexibility: Delivery	0.12*	-0.29, 0.58
Flexibility: Adherence	0.10*	-0.97, 0.69
Follow-Up	-0.71*	-3.50, 0.46
Primary Outcome	0.84	0.39, 0.95
Primary Analysis	0.92	0.71, 0.98
Summary Score	0.41	-0.22, 0.81

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 2: TA, Rater 3: DL; *p>0.05 not statistically significant

Inter-rater reliability of raters 2 and 3 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Bennett S, et al. 2016 (N=12 articles)

Domain	ICC	95% CI
Eligibility	-0.17*	-1.63, 0.60
Recruitment	0.62*	-0.24, 0.89
Setting	0.88	0.24, 0.97
Organization	0.55*	-0.25, 0.86
Flexibility: Delivery	0.79	0.26, 0.94
Flexibility: Adherence	0.51*	-0.85, 0.86
Follow-Up	0.54	-0.28, 0.86
Primary Outcome	0.49	-0.79, 0.85
Primary Analysis	0.81	0.34, 0.95
Summary Score	0.83	0.44, 0.95

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 2: TA, Rater 3: DL; *p>0.05 not statistically significant

Inter-rater reliability of raters 2 and 3 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Song H, et al. 2016 (N=10 articles)

Domain	ICC	95% CI
Eligibility	0.34*	-2.28, 0.84
Recruitment	Not estimable	-
Setting	0.89	0.58, 0.97
Organization	0.71	-0.23, 0.93
Flexibility: Delivery	0.42	-0.26, 0.83
Flexibility: Adherence	0.63*	-0.73, 0.91
Follow-Up	-0.10*	-0.26, 0.35
Primary Outcome	0.89	0.58, 0.97
Primary Analysis	0.89	0.56, 0.97
Summary Score	0.40*	-0.28, 0.81

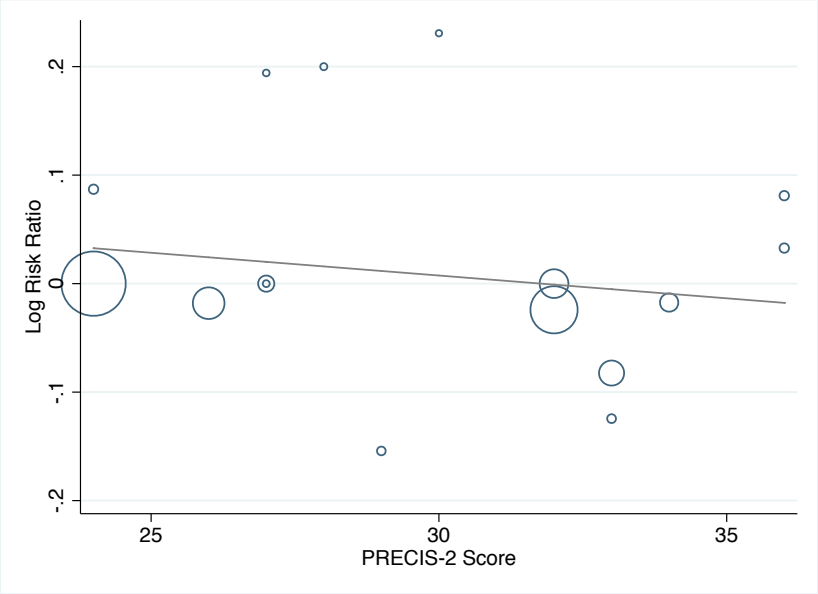
CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 2: TA, Rater 3: DL; *p>0.05 not statistically significant

Inter-rater reliability of raters 2 and 3 for 9 PRECIS-2 domains and overall sum score for primary RCTs in the systematic review by Hofstede SN, et al. 2015 (N=13 articles[†])

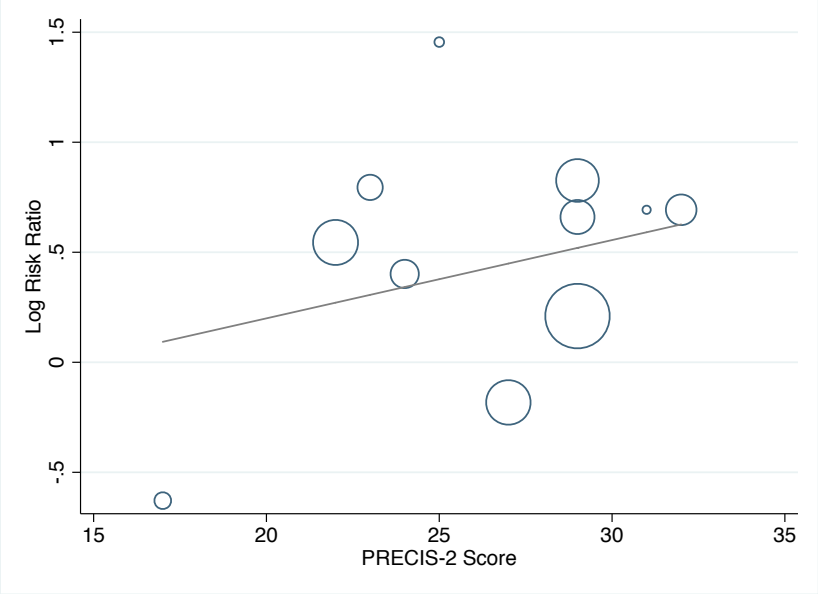
Domain	ICC	95% CI
Eligibility	0.75	0.21, 0.92
Recruitment	0.95	0.84, 0.99
Setting	0.82	0.41, 0.95
Organization	0.43	-0.75, 0.82
Flexibility: Delivery	0*	-0.28, 0.42
Flexibility: Adherence	0.67	-0.01, 0.90
Follow-Up	0.55*	-0.23, 0.85
Primary Outcome	0.64	-0.03, 0.89
Primary Analysis	0.70	0.11, 0.91
Overall Score	0.92	0.76, 0.98

CI: confidence interval; ICC: intraclass correlation coefficient; RCT: randomized controlled trials; Rater 2: TA, Rater 3: DL; [†]Missing 1 primary RCT; *p>0.05 not statistically significant

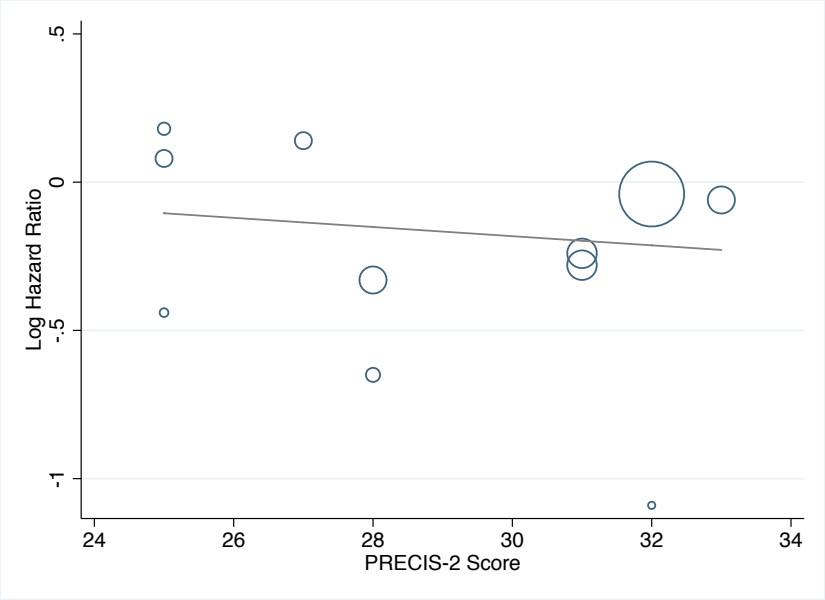
Appendix 6: Graphs of Random Effects Meta-Regression Adjusting for PRECIS-2 by Systematic Review



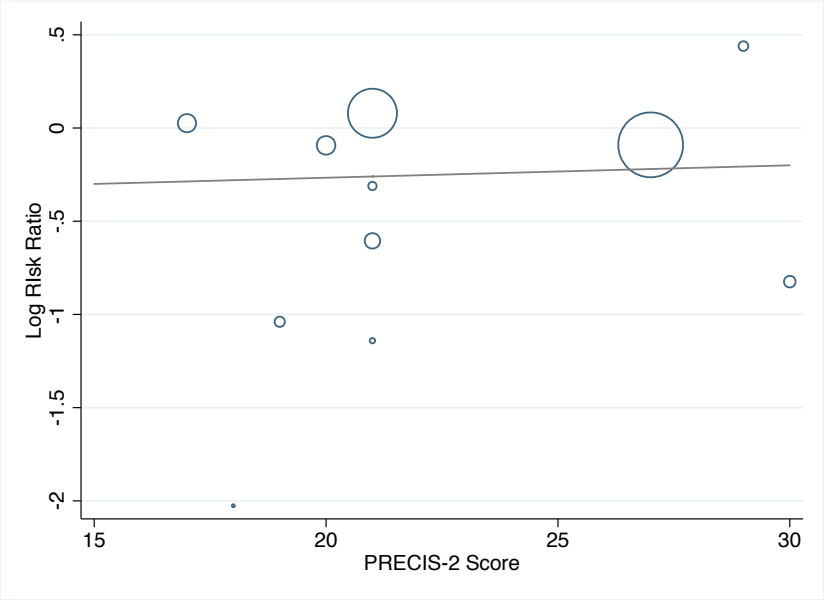
Graph of random effects meta-regression for the systematic review by Hafner et al., adjusting for PRECIS-2. The systematic review compared the effects of colonoscopy with water infusion to colonoscopy with air insufflation on cecal intubation rate (risk ratio: 1.00; 95% CI: 0.98, 1.01; I^2=68%; N=16 articles).



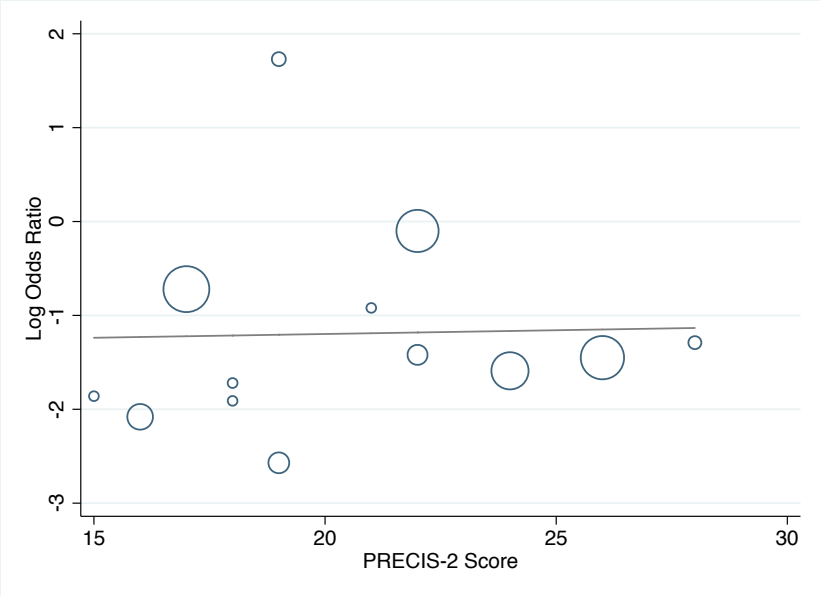
Graph of random effects meta-regression for the systematic review by Martí-Carvajal et al., adjusting for PRECIS-2. The systematic review compared the effects of growth factor treatment to placebo or usual care on complete wound healing in patient with diabetic foot ulcers (risk ratio: 1.03; 95% CI: 0.95, 1.13; I^2=59%; N=12 articles, 1 article missing).



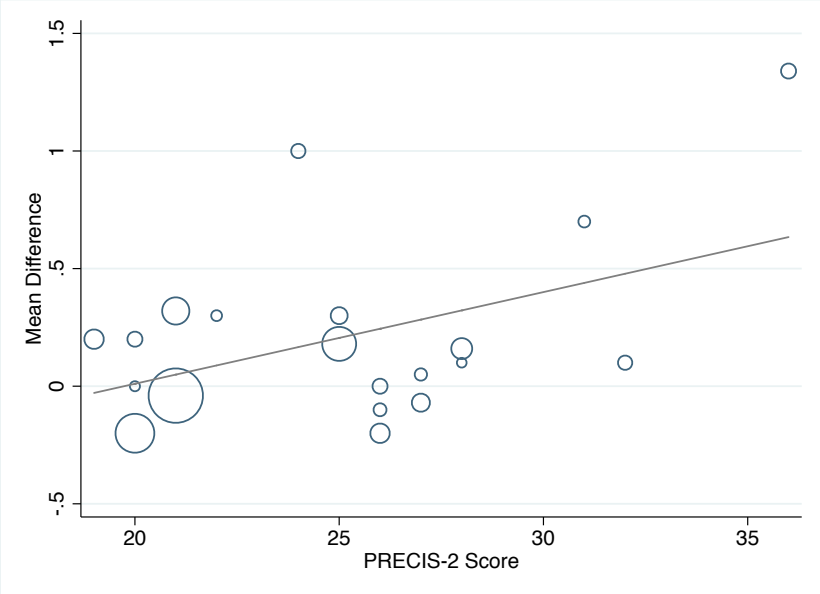
Graph of random effects meta-regression for the systematic review by Akl et al., adjusting for PRECIS-2. The systematic review compared the effects of heparin to placebo or usual care on all-cause mortality over the duration of the trial in patients with cancer and no indication for anticoagulation (hazard ratio: 0.98; 95% CI: 0.92, 1.06; I^2=62%; N=11 articles).



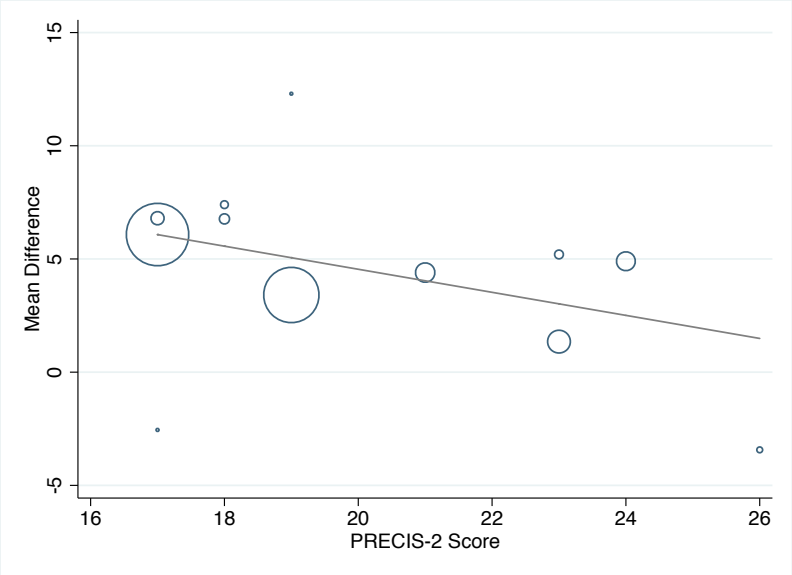
Graph of random effects meta-regression for the systematic review by Buppasiri et al., adjusting for PRECIS-2. The systematic review compared the effects of calcium supplementation to placebo or usual care on preterm birth less than 37 weeks' gestation in pregnant women (risk ratio: 1.01; 95% CI: 0.92, 1.09; I^2=58%; N=13 articles, 1 article missing).



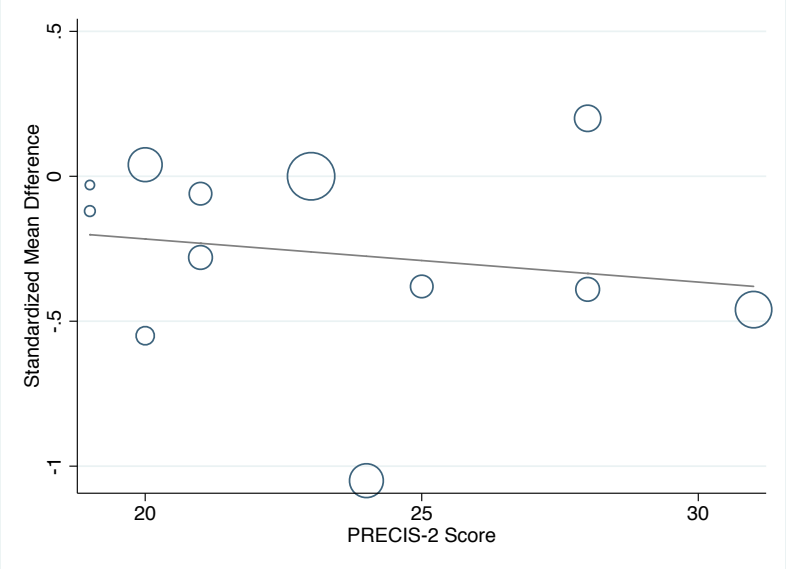
Graph of random effects meta-regression for the systematic review by Hnin et al., adjusting for PRECIS-2. The systematic review compared the effects of prolonged antibiotic therapy to placebo or usual care on exacerbations in adults and children diagnosed with bronchiectasis (odds ratio: 1.01; 95% CI: 0.84, 1.20; I^2=55%; N=13 articles).



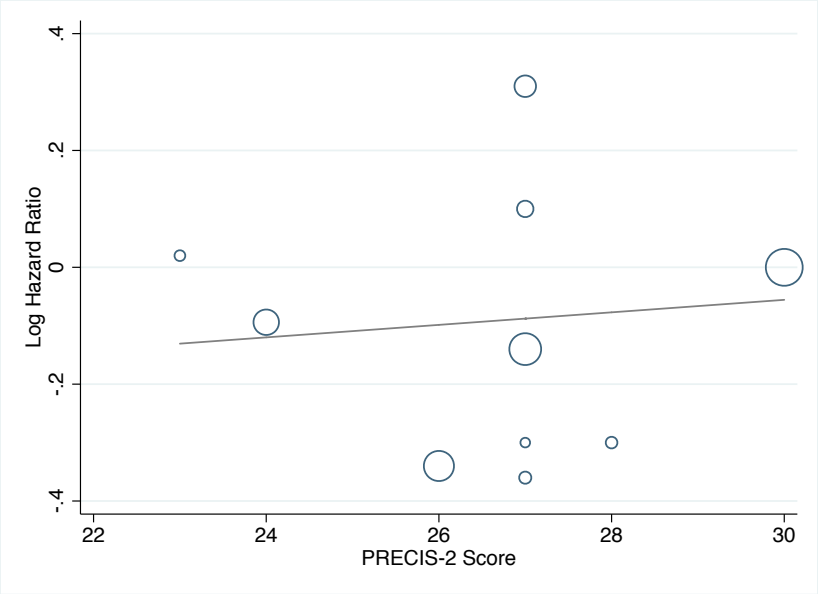
Graph of random effects meta-regression for the systematic review by Birch et al., adjusting for PRECIS-2. The systematic review compared the effects of heated gas insufflation to cold gas insufflation on change in intra-operative core temperature in adults and children undergoing laparoscopic abdominal surgery (mean difference: 0.04; 95% CI: 0.001, 0.08; I^2=79%; N=19 articles, 1 article missing).



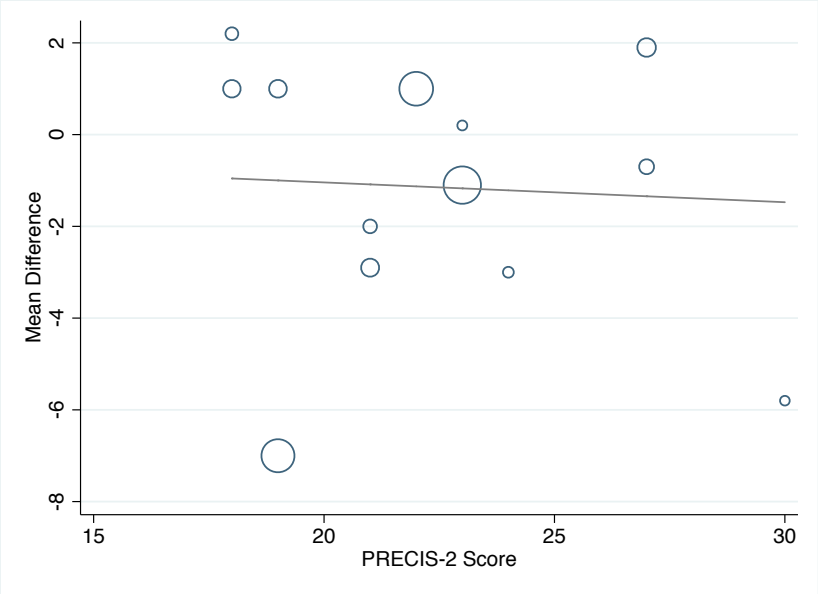
Graph of random effects meta-regression for the systematic review by Lane et al., adjusting for PRECIS-2. The systematic review compared the effects of any exercise program compared to placebo or usual care on maximal walking time in patients with symptomatic intermittent claudication due to atherosclerotic disease (mean difference: -0.51; 95% CI: -1.00, -0.02; I^2=69%; N=12 articles).



Graph of random effects meta-regression for the systematic review by Bennett et al., adjusting for PRECIS-2. The systematic review compared the effects of education interventions designed for the management of cancer-related fatigue to usual care, control, or an alternative intervention for cancer fatigue on general fatigue assessed by self-evaluation (standardized mean difference: -0.01, -0.08, 0.05; I^2=81%; N=12 articles).



Graph of random effects meta-regression for the systematic review by Song et al., adjusting for PRECIS-2. The systematic review compared the effects of molecular target agents plus chemotherapy to usual care or no treatment on overall survival in patients with adenocarcinoma or the stomach or gastro-esophageal junction (hazard ratio: 1.01; 95% CI: 0.92, 1.11; I^2=62%; N=10 articles).



Graph of random effects meta-regression for the systematic review by Hofstede et al., adjusting for PRECIS-2. The systematic review compared the effects of mobile bearing implants to fixed bearing implants on the Knee Society clinical score in patients undergoing total knee arthroplasty for osteoarthritis or rheumatoid arthritis (mean difference: -0.04; 95% CI: -0.63, 0.55; I^2=78%; N=14 articles, 1 article missing).

Appendix 7: Random Effects Meta-Regression Adjusting for PRECIS-2 Tertiles

Exploring heterogeneity through random effects meta-regression methods using PRECIS-2 tertiles as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with PRECIS-2 Tertile (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	67.7	-0.6
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	54.7	-0.3
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	57.1	-1.6
Buppasiri P, 2015 (12*)	RR: 0.80 0.59, 1.08)	54.1	61.7	7.6
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	59.0	8.0
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	81.1	-2.9
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	80.2	-2.0
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	82.5	2.4
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	62.8	2.8
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	73.9	-4.5

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Appendix 8: Random Effects Meta-Regression Adjusting for Individual PRECIS-2 Domains

Exploring heterogeneity through random effects meta-regression methods using the eligibility domains as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Eligibility Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	70.4	2.1
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	58.6	3.6
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	60.2	1.5
Buppasiri P, 2015 (12*)	RR: 0.80 0.59, 1.08)	54.1	52.2	-1.9
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	53.1	2.1
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	79.2	-4.8
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	80.1	-2.1
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.5	1.4
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	62.4	2.4
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	76.0	-2.4

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Exploring heterogeneity through random effects meta-regression methods using the recruitment domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Recruitment Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	66.3	-2.0
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	58.7	3.7
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	59.4	0.7
Buppasiri P, 2015 (12*)	RR: 0.80 0.59, 1.08)	54.1	52.4	0.3
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	43.2	-7.8
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	84.8	0.8
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	79.7	-2.5
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.8	1.7
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	-	-
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	78.8	0.4

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Exploring heterogeneity through random effects meta-regression methods using the setting domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Setting Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	67.5	-0.8
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	36.7	-18.3
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	54.4	-4.3
Buppasiri P, 2015 (12*)	RR: 0.80 0.59, 1.08)	54.1	49.9	-4.2
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	53.7	2.7
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	79.3	-4.7
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	79.7	-2.5
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.4	1.3
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	63.6	3.6
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	77.9	0.5

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing; [†]Statistically significant p=0.022

Exploring heterogeneity through random effects meta-regression methods using the organization domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Organization Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	59.1	-9.2
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	59.4	4.4
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	56.0	2.7
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	55.9	1.8
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	38.2	-12.8
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	83.9	-0.1
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	83.8	1.6
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.9	1.8
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	64.5	4.5
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	78.8	0.4

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Exploring heterogeneity through random effects meta-regression methods using the flexibility (delivery) domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Delivery Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	66.5	-1.8
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	59.5	4.5
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	62.1	3.4
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	58.7	4.6
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	54.9	3.9
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	80.3	-3.7
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	80.9	-1.3
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.9	1.8
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	62.9	2.9
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	80.2	1.8

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing; [†]Statistically significant p=0.008

Exploring heterogeneity through random effects meta-regression methods using the flexibility (adherence) domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Adherence Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	69.4	1.1
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	55.5	0.5
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	61.7	-3.0
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	23.0	-31.1
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	46.7	-4.3
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	84.9	0.9
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	83.0	0.8
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	79.5	-0.6
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	61.4	1.4
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	79.9	1.5

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing; †Statistically significant p=0.023

Exploring heterogeneity through random effects meta-regression methods using the follow-up domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Follow-Up Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	64.1	-4.2
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	58.6	3.6
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	62.2	3.5
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	57.2	3.1
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	47.0	-4.0
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	84.8	0.8
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	81.1	-1.1
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.9	1.8
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	59.9	-0.1
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	79.2	0.8

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Exploring heterogeneity through random effects meta-regression methods using the primary outcome domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Primary Outcome Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	70.3	2.0
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	58.6	3.6
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	44.3	-14.4
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	44.1	-10.0
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	51.1	0.1
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	82.7	-1.3
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	82.4	0.2
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.9	1.8
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	64.2	4.2
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	80.1	1.7

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Exploring heterogeneity through random effects meta-regression methods using the primary analysis domain as a covariate for each systematic review (N=10)

Primary Author, Year (number of articles)	Effect Size (95% CI)	I^2 without PRECIS-2 (%)	I^2 with the Primary Analysis Domain (%)	I^2 Difference (%)
Hafner S, 2015 (16*)	RR: 1.01 (0.96, 1.06)	68.3	70.2	1.9
Martí-Carvajal AJ, 2015 (11*)	RR: 1.54 (1.14, 2.06)	55.0	56.3	1.3
Akl EA, 2014 (11)	HR: 0.84 0.71, 1.00	58.7	62.2	3.5
Buppasiri P, 2015 (12*)	RR: 0.80 (0.59, 1.08)	54.1	57.2	3.1
Hnin K, 2015 (13)	OR: 0.31 (0.17, 0.56)	51.0	46.0	-4.0
Birch DW, 2016 (18*)	MD: 0.21 (0.03, 0.39)	84.0	84.7	0.7
Lane R, 2014 (12)	MD: 4.51 (2.83, 6.20)	82.2	81.1	-1.1
Bennett S, 2016 (12)	SMD: -0.27 (-0.52, -0.03)	80.1	81.8	1.7
Song H, 2016 (10)	HR: 0.91 (0.78, 1.07)	60.0	55.3	-0.7
Hofstede SN, 2015 (13*)	MD: -1.14 (-3.04, 0.76)	78.4	79.9	1.5

RR: Risk ratio; HR: Hazard ratio; OR: Odds ratio; MD: Mean difference; SMD: Standardized mean difference; *1 primary RCT missing

Appendix 9: RITES domain scores for primary RCTS according to systematic review (N=5)

RITES scores by domain for primary RCTs included in the systematic review by Martí-Carvajal et al. 2015 (N=11 articles*)

Systematic Review Primary Author, et al.	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	RITES Summary Score
d'Hemecourt PA, et al.	2	5	4	5	16
Hanft JR, et al.	1	5	2	4	12
Hardiker JV, et al.	2	5	1	4	12
Holloway GA, et al.	2	4	2	4	12
Jaiswal SS, et al.	4	1	2	2	9
Richard JL, et al.	1	3	2	4	10
Saldalamacchia G, et al.	4	1	4	2	11
Steed D, et al.	1	3	2	4	10
Steed D, et al.	1	5	1	4	11
Viswanathan V, et al.	1	4	2	4	11
Wieman TJ, et al.	1	5	1	4	11

*1 primary RCT missing

RITES scores by domain for primary RCTs included in the systematic review by Akl E et al. 2014 (N=11 articles)

Systematic Review Primary Author, et al.	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	RITES Summary Score
Agnelli G, et al.	3	5	2	4	14
Altinbas M, et al.	2	3	2	5	12
Kakker AK, et al.	2	5	4	5	16
Klerk CP, et al.	1	5	2	4	12
Lebeau B, et al.	1	5	2	4	12
Lecumberri R, et al.	1	5	2	5	13
Maraveyas A, et al.	1	5	1	4	11
Perry JR, et al.	1	5	1	4	11
Sideras K, et al.	1	5	2	5	13
van Doormaal FF, et al.	2	3	3	5	13
Weber C, et al.	2	1	3	4	10

RITES scores by domain for primary RCTs included in the systematic review by Lane R, et al. 2014 (N=12 articles)

Systematic Review Primary Author, et al.	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	RITES Summary Score
Collins EG, et al.	2	1	4	5	12
Crowther RG, et al.	4	1	4	5	14
Hiatt WR, et al.	2	1	3	3	9
Hiatt WR, et al.	2	1	2	3	8
McDermott MM, et al.	1	1	1	4	7
McDermott MM, et al.,	1	1	3	4	9
Mika P, et al.	2	1	3	3	9
Mika P, et al.	2	1	4	3	10
Sanderson B, et al.	4	4	4	3	15
Tsai JC, et al.	4	2	4	4	14
Wood RE, et al.	4	1	4	2	11
Larsen OA, et al.	3	1	4	3	11

RITES scores by domain for primary RCTs included in the systematic review by Bennett S, et al. 2016 (N=12 articles)

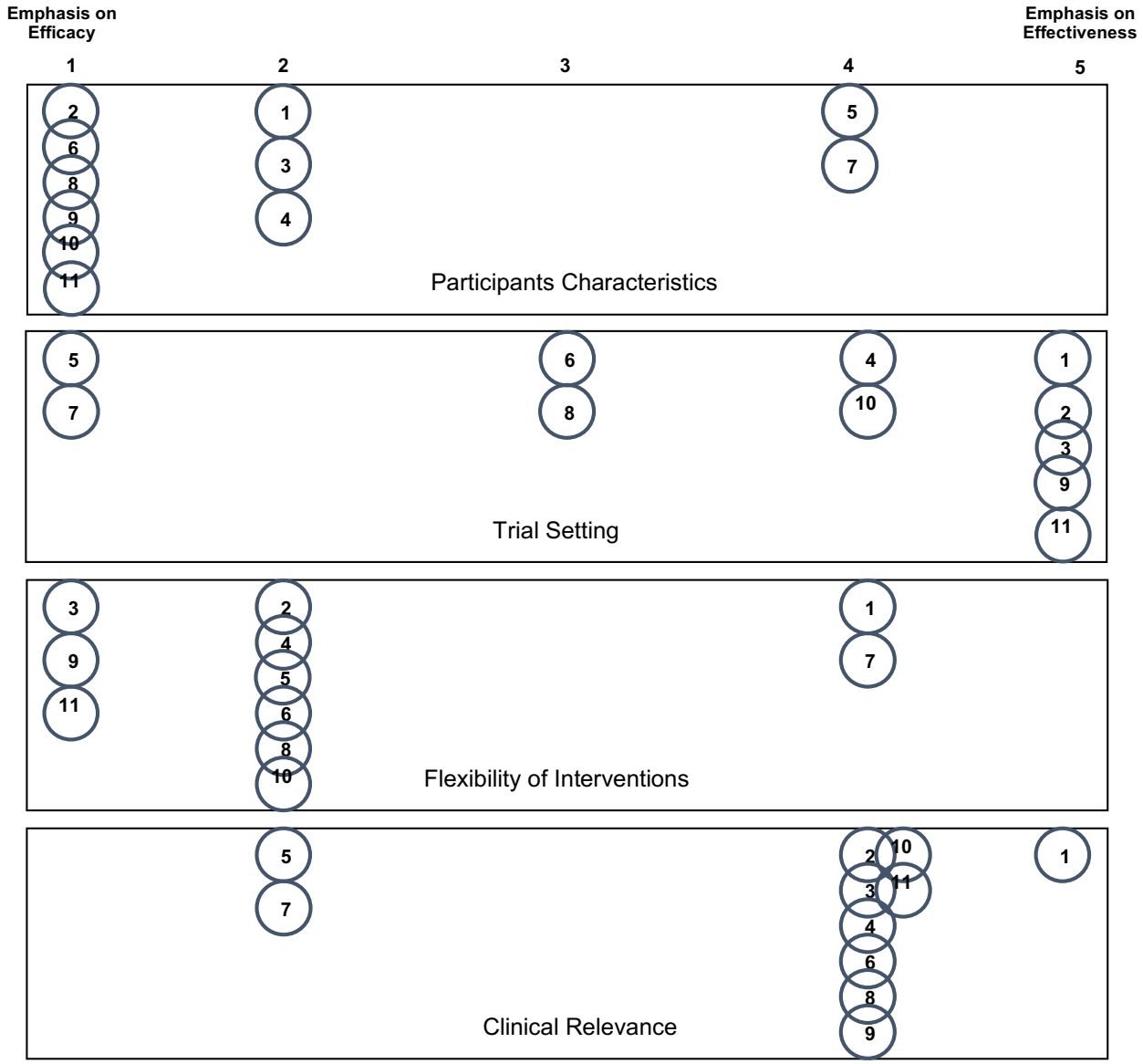
Systematic Review Primary Author, et al.	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	RITES Summary Score
Reif K, et al.	2	5	2	4	13
Wangnum K, et al.	2	1	3	4	10
Yun YH, et al.	2	4	3	5	14
Yates P, et al.	3	5	4	4	16
Ream E, et al.	5	3	3	4	15
Purcell A, et al.	2	1	2	5	10
Godino C, et al.	2	1	4	3	10
Schjolberg T, et al.	3	1	3	4	11
Yuen HK, et al.	2	1	2	4	9
Barsevick AM, et al.	2	4	2	4	12
Barsevick AM, et al.	2	4	2	3	11
Foster C, et al.	3	5	3	3	14

RITES scores by domain for primary RCTs included in the systematic review by Hofstede SN, et al. 2015 (N=13 articles*)

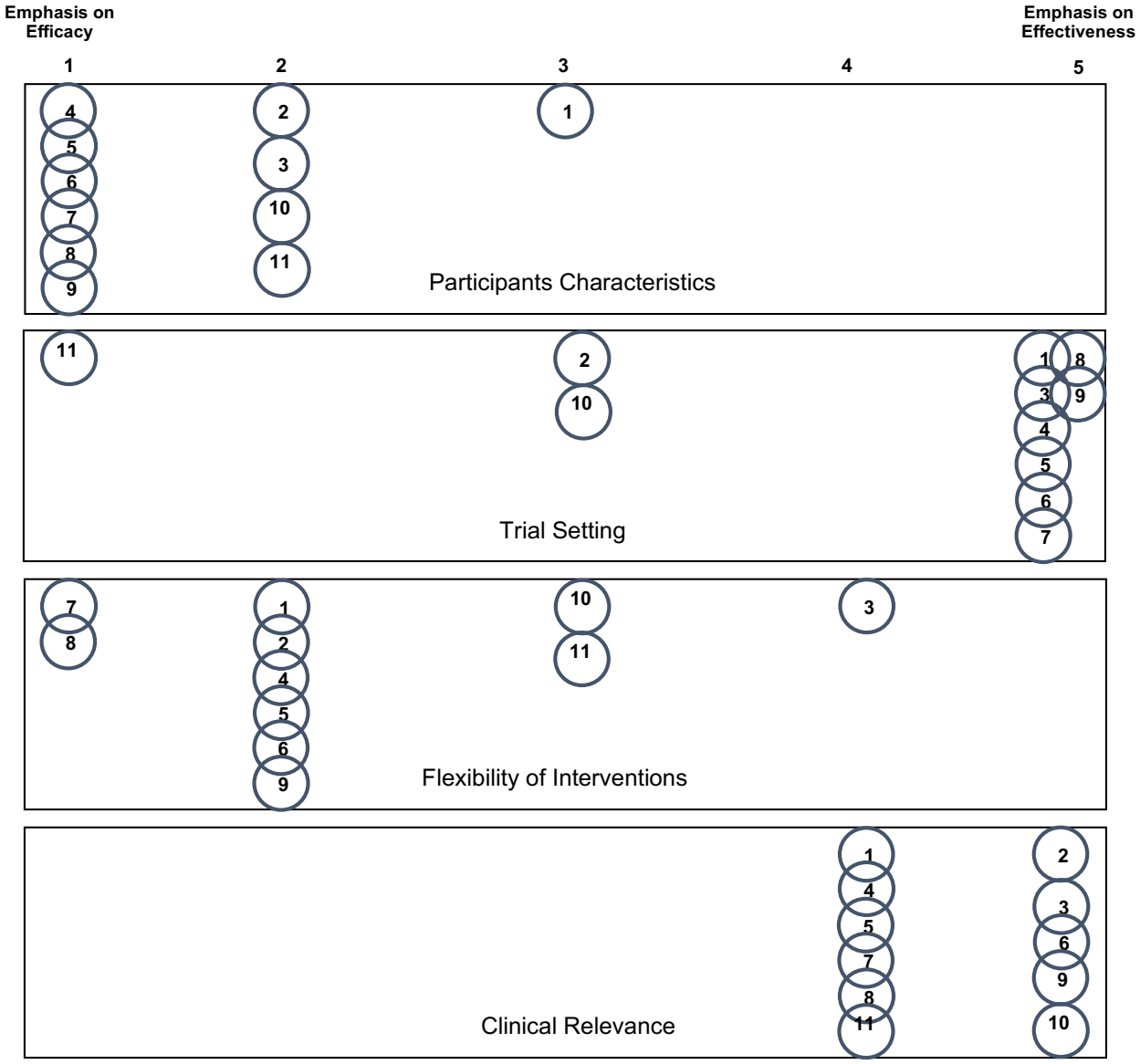
Systematic Review Primary Author, et al.	Participant Characteristics	Trial Setting	Flexibility of Intervention	Clinical Relevance	RITES Summary Score
Bailey O, et al.	3	4	4	5	16
Hanusch B, et al.	3	3	2	5	13
Henricson A, et al.	2	1	3	3	9
Jacobs WCH, et al.	2	3	3	2	10
Kim YH, et al.	5	1	3	4	13
Kim YH, et al.	5	1	3	4	13
Kim YH, et al. (A)	4	1	3	4	12
Kim YH, et al. (B)	3	1	3	3	10
Kim TK, et al.	4	1	4	3	12
Lampe F, et al.	3	1	2	4	10
Munro JT, et al.	2	3	4	3	12
Price AJ, et al.	4	4	3	3	14
Watanabe T, et al.	4	1	3	3	11

*1 primary RCT missing

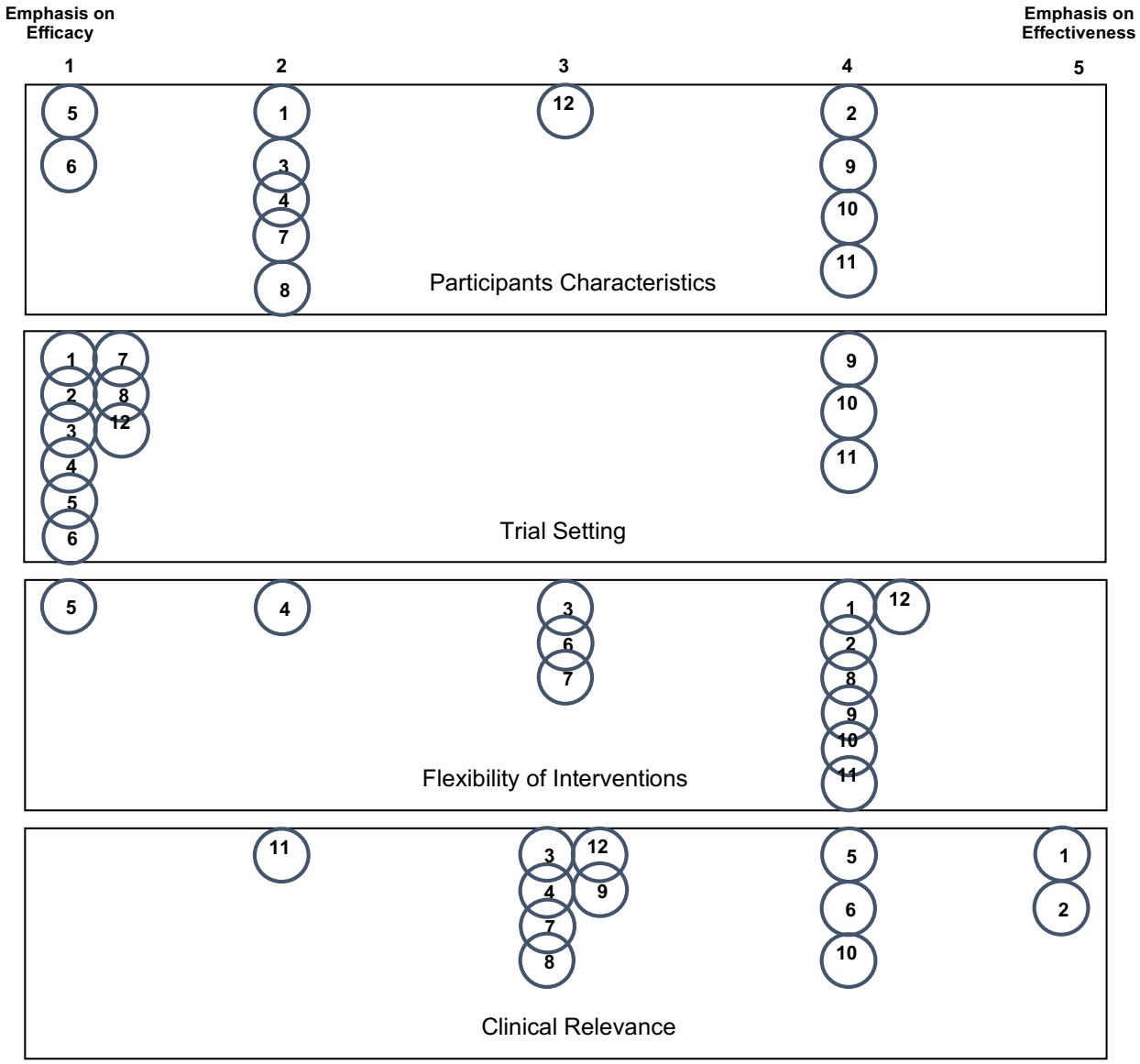
Appendix 10: Visual Description of RITES by Systematic Review



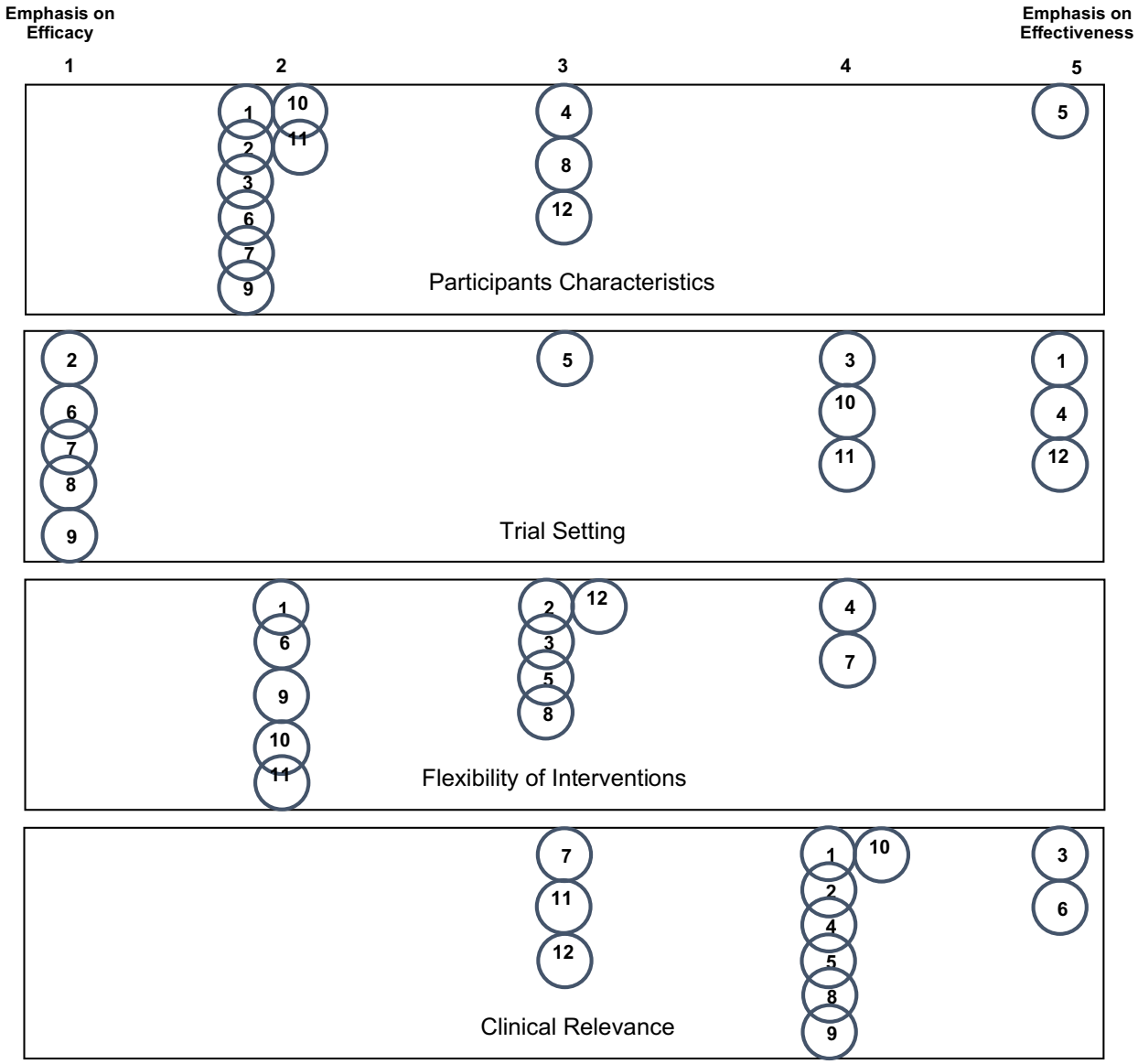
Visual description of primary studies on the efficacy-effectiveness spectrum for the systematic review by Martí-Carvajal et al. Primary trials are represented by numbers (N=11 articles, 1 article missing).



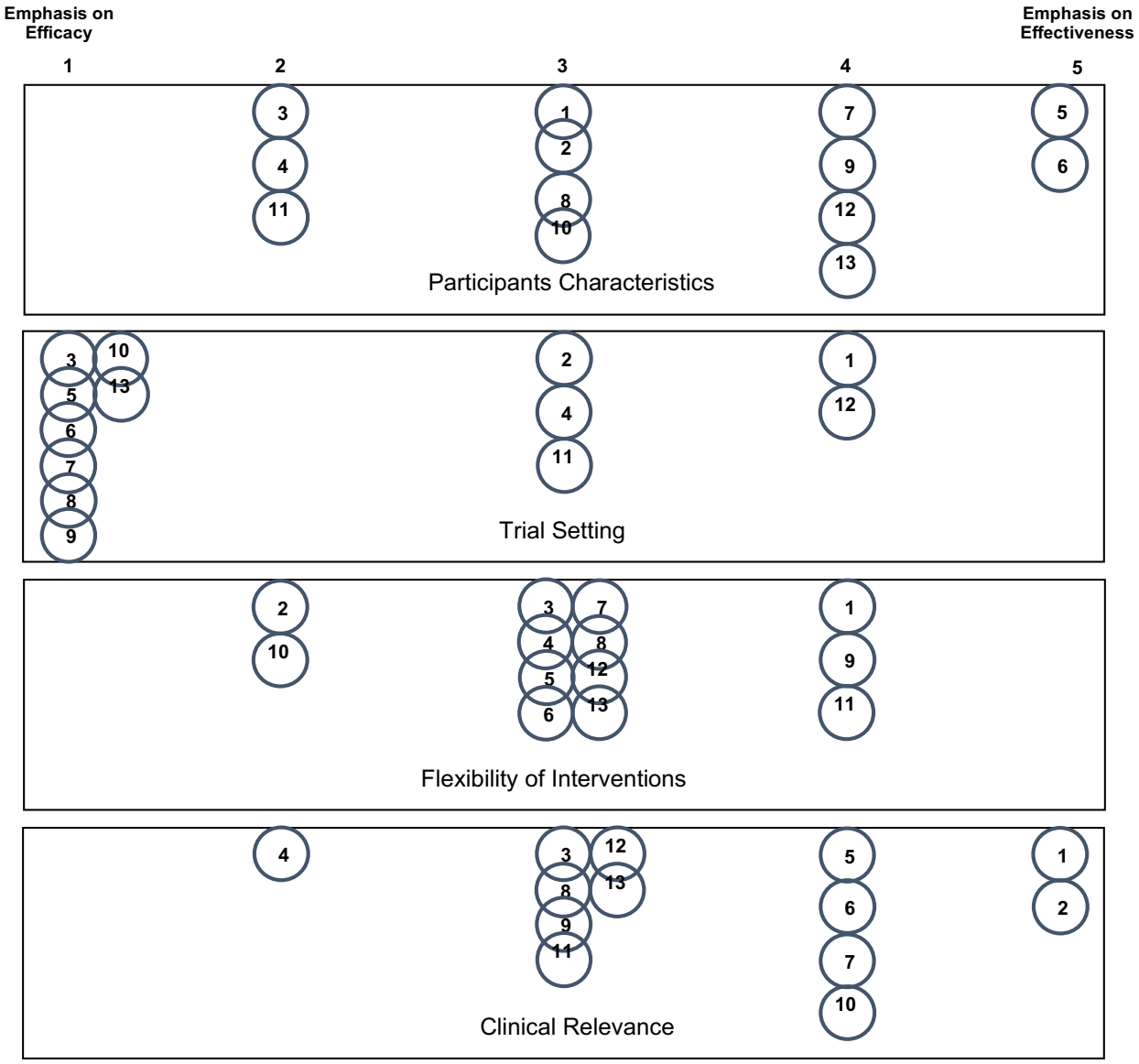
Visual description of primary studies on the efficacy-effectiveness spectrum for the systematic review by Akl et al. Primary trials are represented by numbers (N=11 articles).



Visual description of primary studies on the efficacy-effectiveness spectrum for the systematic review by Lane et al. Primary trials are represented by numbers (N=12 articles).



Visual description of primary studies on the efficacy-effectiveness spectrum for the systematic review by Bennett et al. Primary trials are represented by numbers (N=12 articles).



Visual description of primary studies on the efficacy-effectiveness spectrum for the systematic review by Hofstede et al. Primary trials are represented by numbers (N=13 articles, 1 article missing).

Appendix 11: Inter-Rater Reliability of RITES by Systematic Review

Inter-rater reliability or raters 1 and 2 for all RITES domains and summary score for primary RCTs in the systematic review by Martí-Carvajal AJ, et al. 2015 (N=11 articles[†])

Domain	ICC	95% CI
Participant Characteristics	0.93	0.76, 0.99
Trial Setting	1.0	-
Flexibility of Intervention(s)	0.98	0.91, 0.99
Clinical Relevance of Intervention(s)	0.95	0.81, -.99
Average Score	0.93	0.76, 0.98
Overall Score	0.93	0.76, 0.98

CI: confidence interval; [†]Missing 1 primary RCT; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability or raters 1 and 2 for all RITES domains and summary score for primary RCTs in the systematic review by Akl EA, et al. 2014 (N=11 articles)

Domain	ICC	95% CI
Participant Characteristics	0.77	0.17, 0.94
Trial Setting	1.0	-
Flexibility of Intervention(s)	0.73	-0.21, 0.94
Clinical Relevance of Intervention(s)	0.65	-0.26, 0.91
Average Score	0.93	0.74, 0.98
Overall Score	0.93	0.75, 0.98

CI: confidence interval; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability or raters 2 and 3 for all RITES domains and summary score for primary RCTs in the systematic review by Lane R, et al. 2014 (N=12 articles)

Domain	ICC	95% CI
Participant Characteristics	0.87	0.53, 0.96
Trial Setting	0.93	0.74, 0.98
Flexibility of Intervention(s)	0.78	0.30, 0.94
Clinical Relevance of Intervention(s)	0.66	-0.06, 0.90
Average Score	0.94	0.79, 0.98
Overall Score	0.94	0.79, 0.98

CI: confidence interval; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability or raters 2 and 3 for all RITES domains and summary score for primary RCTs in the systematic review by Bennett S, et al. 2016 (N=12 articles)

Domain	ICC	95% CI
Participant Characteristics	0.52	-0.28, 0.85
Trial Setting	0.97	0.90, 0.99
Flexibility of Intervention(s)	0.39	-1.39, 0.83
Clinical Relevance of Intervention(s)	0*	-2.47, 0.71
Average Score	0.67	-0.04, 0.90
Overall Score	0.67	-0.04, 0.90

CI: confidence interval; Rater 1: KA, Rater 2: TA; *p>0.05 not statistically significant

Inter-rater reliability or raters 2 and 3 for all RITES domains and overall sum score for primary RCTs in the systematic review by Hofstede SN, et al. 2015 (N=13 articles[†])

Domain	ICC	95% CI
Participant Characteristics	0.92	0.78, 0.98
Trial Setting	1.0	-
Flexibility of Intervention(s)	0.47*	-0.81, 0.84
Clinical Relevance of Intervention(s)	0.52*	-0.43, 0.85
Average Score	0.86	0.56, 0.96
Overall Score	0.86	0.56, 0.96

CI: confidence interval; Rater 1: KA, Rater 2: TA; [†]Missing 1 primary RCT; *p>0.05 not statistically significant