BMJ Open

Risk-of-bias assessment of studies reporting prevalence

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037324
Article Type:	Original research
Date Submitted by the Author:	29-Jan-2020
Complete List of Authors:	Glasgow, Matthew; The University of Auckland, Liggins Institute Edlin, Richard; University of Auckland, Health Systems, School of Population Health Harding, Jane; University of Auckland Liggins Institute
Keywords:	HEALTH ECONOMICS, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Risk-of-bias assessment of studies reporting prevalence

Matthew J Glasgow [1], Richard Edlin [2*], Jane E Harding [1]

[1] Liggins Institute, University of Auckland, Auckland, New Zealand

[2] School of Population Health, University of Auckland, Auckland, New Zealand

[*] Corresponding author (e-mail: j.harding@auckland.ac.nz)

Financial support for this study was provided in part by a grant (Douglas Goodfellow Medical Research Fellowship, grant number 1417003) from the Auckland Medical Research Foundation. The funding agreement ensured the authors' independence in study design, data collection, data analysis, data interpretation, writing of the report, and the decision to submit the paper for publication. The authors declare no conflicts of interest. Each author listed on the manuscript has seen and approved the submission of this version of the manuscript and takes full responsibility for the manuscript.

The authors declare no competing interests.

Word Count: 3610

Abstract

Objectives: Within cost-effectiveness models, prevalence figures can inform transition probabilities. The methodological quality of studies can inform the choice of prevalence figures but no single obvious candidate tool exists for assessing quality of the observational epidemiological studies for selecting prevalence estimates. We aimed to compare different tools to assess the risk of bias of studies reporting prevalence, and develop and compare possible numerical scoring systems using these tools to set a threshold for inclusion of reports of prevalence in an economic analysis of neonatal hypoglycaemia.

Design: Assessments of bias from two tools (JBI Checklist for Prevalence Studies and a modified version of ROBINS-I) were compared 18 studies relevant to a single setting (neonatal hypoglycaemia). Inclusion of studies for use in a decision analysis model were considered based on summary scores derived from these tools.

Results: Both tools were considered easy to use, with dispersed scores obtained. The ROBINS-I scores were more skewed than the JBI scores, particularly at higher thresholds. The studies selected for inclusion are generally the same using either tool. However, the JBI tool is shorter and may be easier to interpret and apply to studies that do not involve a control group, while the ROBINS-I tool assesses more methodological detail in studies that include a control group.

Conclusion: Both tools performed well for systematically assessing studies that report on outcome prevalence and provided similar discrimination of risk of bias between. This convergent validity supports both tools for the purpose of assessing risk of bias and selecting studies that report prevalence.

Strengths and limitations of this study

- This study addresses a methodological task for which no single obvious candidate tool exists.
- Assessments of candidate tools and approaches to use of the tools were undertaken independently by the three authors.
- Convergent validity between the tools examined supports the use of either approach, or derivations thereof, to guide the inclusion of prevalence reports in economic modelling.
- Studies were assessed by each researcher using one tool immediately followed by the other in a consistent order. For assessment items that are similar, responses to one tool may therefore have influenced responses using the second tool.

Introduction

The probability of an outcome occurring is a fundamental parameter required in the creation of a decision analytic model. It represents the likelihood that patients in a cohort will move from one health state to another in a decision tree or state transition model (e.g., Markov model), and is thus often referred to as a transition probability. [1] When referring to clinical outcomes, the transition probability is equivalent to the prevalence of that outcome in the population represented in the model.

The evidence base from which model parameters are drawn often involves more than a single data source, and developing the model may involve aggregation of this data. [2] The process of deciding which values to use as the key inputs in a model, including the transition probabilities, should be based on a systematic review of the literature, and a description of this process should accompany the model, [3-5] with the use of a source and any translational steps justified. [1, 4] The use of published studies as a source for transition probabilities should have their validity transparently assessed by applying critical appraisal criteria. [4]

In 2016, Sterne et al observed that, in terms of assessing study validity, there has been a shift in focus away from analysis of methodological quality to assessments of risk of bias, often in a domain-oriented manner i.e., considering different domains of bias in turn. [6] The potential for bias, and types of bias, in non-randomised studies may differ from those in randomised studies. [7] A number of instruments for assessing the risk of bias in non-randomised studies have been developed. [7] In 2003, Deeks et al identified six that were considered to have utility for systematic reviews, although they noted that none had been formally validated. [7] In 2007, Sanderson et al concluded that there was a lack of a single obvious candidate tool for assessing quality of observational epidemiological studies. [8] They identified three domains as being fundamental in assessing risk of bias (appropriate selection of patients, appropriate

measurement of variables, and appropriate control of confounding), but noted that these were present in only approximately half of the checklists that they evaluated. [8]

Subsequent to these systematic reviews, Sterne et al developed the ROBINS-I ("Risk Of Bias In Non-randomised Studies - of Interventions") tool to evaluate the risk of bias in studies that do not use randomisation to allocate participants to comparison groups. [6] The ROBINS-I includes a total of 7 bias domains: selection of comparison groups, confounding, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result. These domains can be further compartmentalised into pre-intervention (confounding and participant selection), intervention (classification of interventions), and post-intervention (the remainder) categories. [6] The ROBINS-I assesses risk of bias using an absolute scale, as distinct to the approach commonly used by other similar tools of comparing against a theoretical, perfect observational study or a high quality randomised trial. [9] ROBINS-I was constructed with an objective of allowing the risk-of-bias assessment to determine the degree to which the rating of a study is downgraded. [6] This would facilitate comparison between ratings of randomised trials and ratings of non-randomised studies when using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. If we consider the intervention to be an exposure (e.g., the occurrence of neonatal hypoglycaemia), the ROBINS-I provides a systematic approach that can assess a non-interventional observational study for risk of bias within the seven specified domains.

In 2015, Munn et al observed a lack of guidance for authors undertaking systematic reviews of observational epidemiological studies, including those reporting prevalence or incidence information. [10] That absence of guidance included the lack of a standard method for conducting critical appraisals of the studies used in systematic reviews of prevalence data. [11] The same authors also observed a significant increase in the volume of systematic

reviews being performed and published that focused on questions of prevalence. [11] This combination of factors led to the establishment of a working group, composed of researchers from the Joanna Briggs Institute (JBI, University of Adelaide, Australia), to create guidance for conducting systematic reviews of studies reporting incidence and prevalence parameters. [10] This guidance has been published as a checklist with supporting explanatory information. [12] When applied to prevalence studies, reported risks of bias in the JBI tool cover a similar array of concepts to the ROBINS-I tool.

The ROBINS-I and JBI tools were selected for comparison in this study in light of the conclusions by Sanderson et al in their comprehensive 2007 systematic review that, despite the existence at that time of 86 candidate tools developed to assess the quality of evidence from observational epidemiological studies, none could be recommended as a single ideal candidate. [8] Both the ROBINS-I and JBI tools were developed subsequent to that review, and address a number of the recommendations from Sanderson et al, particularly those relating to rigour in their development, and appropriate coverage of key domains.

The ROBINS-I domain pertaining to bias in ascertainment of exposures is notably lacking from the JBI tool, which was not designed with the explicit intent of assessing reports of prevalence after a nominated exposure. It does not explicitly inquire about such concepts as whether exposure was measured prior to determination of outcome; whether exposure measures were defined, reliable, and consistently applied; whether different levels of exposure were considered; or whether the exposure was assessed more than once over time. The JBI tool also does not explicitly address bias in reporting of results, particularly the implications of performing multiple measurements or analyses of the exposure-outcome relationship.

Conversely, although the ROBINS-I tool does assess a number of concepts related to measurement of the outcomes, it does not explicitly examine the validity of outcome ascertainment, and it does not downgrade on the basis of sample size alone. Further, the ROBINS-I tool contains a series of assessment items examining the appropriateness of methods for selecting a control group; a topic not included in the JBI tool.

Differences in prevalence for the same or similar outcomes vary for a number of reasons, including methodological differences, differences in definitions of the outcomes, and differences in the populations being examined. We wished to select published reports of prevalence of outcomes of neonatal hypoglycaemia for use in a decision analytic model for an economic analysis. A wide range of prevalence figures have been reported for these outcomes, in part because of inconsistencies in the definition of neonatal hypoglycaemia, particularly the blood glucose concentration threshold used to diagnose asymptomatic cases, changes in that definition over time, and differences in approaches to screen for and identify the condition. The blood glucose concentration threshold for diagnosing neonatal hypoglycaemia has ranged from, 20 mg/100mL (1.11 mmol/L) [13] in earlier studies to 2.6 mmol/L [14], and has variably included additional criteria such as a requirement for low results on consecutive measurements.

In economic analyses, each prevalence parameter needs to be informed by the available information even if the underlying quality of information is of poor quality. This means that the question becomes how to decide which sources of information to include and not include for each outcome, rather than determining a single inclusion threshold across all studies. We first undertook this study to examine the use of risk-of-bias assessments to assist with these decisions.

Objective

We aimed to 1) undertake a comparison of different tools to assess the risk of bias of studies reporting prevalence for use as data sources for economic analyses, and 2) develop and compare possible numerical scoring systems using these tools to set a threshold for inclusion of reports of prevalence in an economic analysis, using the example of outcomes of neonatal hypoglycaemia.

Methods

Both the Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I) tool [6] and the Joanna Briggs Institute Checklist for Prevalence Studies [12] were selected for initial assessment. We chose these two tools based on their applicability to observational studies and/or studies reporting prevalence, consistency with the GRADE approach to assessment of uncertainty, and advice from local researchers familiar with candidate instruments. A modified version of the ROBINS-I tool was pre-formatted into a spreadsheet for ease of use by assessors. The ROBINS-I assessment item pertaining to the bias domain of deviations from intended interventions was excluded as the topic of interest was an exposure at a point in time rather than an intervention over time. Instead, we added three assessment items pertaining to the domain of study design (clarity of the statement of objective, inclusion of sample size justification or similar, inclusion of an unexposed group) and three pertaining to external validity (specification of the study population, relevance of the cohort to the target population, and drop-out rate). For each domain the overall bias was summarised as high, low, or uncertain.

From the pool of non-randomised studies that reported, or allowed the calculation of, prevalence of outcomes of neonatal hypoglycaemia, three were initially selected covering a

range of methodologies and study population sizes. [15-17] Each focused on a single outcome.

Both assessment tools were used independently by three researchers to assess all three studies. Completed assessments were tabulated and compared, and differences were discussed and reconciled. In each instance, discussion about the reasons for discrepant assessment elements resulted in a consensus between researchers. A further 18 studies [18-35] reporting prevalence of outcomes after neonatal hypoglycaemia were then each assessed by combinations of two of the three researchers using both tools.

Three summary scores were formed to facilitate further comparison between studies (Supplementary Table 1).

1. Count Score(s): sums of responses in each column. That is, the total number of responses indicating low risk of bias, and the total number of responses indicating high risk of bias. Two separate values are thus generated. For the sum of responses indicating low risk of bias, a higher score represents a low risk of bias; for the sum of responses indicating a high risk of bias, a higher score represents a high risk of bias. These are presented as a percentage of the total value possible on the tool. (Note that the total number of questions, and therefore the maximum total value, is 12 on the modified ROBINS-I tool and 9 on the JBI tool).
2. Composite Score: calculated by subtracting the total number of responses indicating high risk of bias from the total number of responses indicating low risk of bias. A higher score represents a lower risk of bias. Negative values are possible for studies that score a greater number of high risk of bias elements/domains than low risk of bias elements. This is presented as a percentage of the total value possible on the tool.

3. Applicable Score: conversion of the Composite Score into a percentage by dividing the Composite Score by the maximum score possible after subtracting any "not applicable" responses. A higher score represents a lower risk of bias. Negative values are also possible using this approach.

All three scores have a maximum value of 100%. Excluding "not applicable" responses in the Applicable Score is intended to more accurately reflect which elements of the tool are relevant to the study being assessed.

Patient and public involvement

This work is a research methods paper, and as such was undertaken without patient involvement.

Results

Ease of use and assessor agreement for initial three studies

All three researchers reported that the assessment tools and spreadsheets were generally easy to use, and that, because of the structural and content similarities between the two tools, assessment using both tools did not result in a large increase in time required compared to assessment using a single tool. However, since the JBI tool includes fewer assessment items it may have a modest time advantage over the ROBINS-I tool.

For the initial individual assessments there was not unanimous inter-assessor agreement. However, for the ROBINS-I tool, the only field for which a discrepancy was consistently reported across all three studies was that relating to bias due to missing data. For the JBI tool, no fields were found for which a discrepancy was consistently reported across all studies (Supplementary Table 2).

Disagreements were predominantly where one or more researchers responded with "unknown" or "not applicable" responses while other researchers allocated a high or low risk of bias in that field ("applicability classification conflict"). There were few discrepancies where a single researcher classified a field as high risk of bias and others classified it as low risk of bias, or vice versa ("high-low conflict"). The number of high-low conflicts with the JBI tool were fewer than with the ROBINS-I tool (Supplementary Table 2).

Assessment tool scores and agreement

When used by combinations of two researchers to assess 40 study-outcome combinations (hereafter "assessments") from the 18 studies, both the ROBINS-I and JBI tools resulted in a wide distribution of scores for each outcome (Figures 1 and 2), potentially allowing selection of studies for inclusion at a wide range of thresholds. The distribution of scores with the ROBINS-I tool was generally skewed slightly higher than the distribution of scores with the JBI tool.

Using the Count Scores, the difference between the two tools in the number of studies selected for inclusion or exclusion varies with the threshold in a non-linear manner (Table 1). For lower thresholds (e.g., 25%), there is greater difference between the two tools than for higher thresholds (e.g., 50%, 75%), with more studies being included using ROBINS-I than with JBI at the lower thresholds.

	Scoring system:	1. Count Score, Positive			2. Composite Score			3. Applicable Score		
Studies included	Threshold:	25%	50%	75%	25%	50%	75%	25%	50%	75%
All assessment	ROBINS-I	33	22	12	21	18	11	21	18	11

types (of 40 assessments)	JBI	23	19	14	20	19	13	20	19	13
Learning disabilities (of 13 assessments)	ROBINS-I	12	9	4	8	6	3	8	6	3
	JBI	9	7	5	7	7	4	7	7	4
Severe learning disabilities (of 4 assessments)	ROBINS-I	4	4	2	4	4	2	4	4	2
	JBI	4	4	3	4	4	3	4	4	3
Cerebral palsy (of 7 assessments)	ROBINS-I	6	4	2	4	4	2	4	4	2
	JBI	4	4	2	4	4	2	4	4	2
Epilepsy; seizures (of 8 assessments)	ROBINS-I	7	4	3	4	3	3	4	3	3
	JBI	4	3	3	3	3	3	3	3	3
Vision disorders; blindness (of 8 assessments)	ROBINS-I	4	1	1	1	1	1	1	1	1
	JBI	2	1	1	2	1	1	2	1	1

Table 1: Number of studies selected for inclusion when assessing different outcomes using three different scoring systems of the ROBINS-I and JBI tools at different thresholds

Using the Composite or the Applicable Scores, the ROBINS-I and JBI tool resulted in very similar numbers of studies included (Table 1). If 50% was used as the cut-off threshold for inclusion or exclusion of studies based on their risk of bias, both tools would give the same results using the Applicable Score (Figure 3). The level of agreement fell (i.e., some studies would be included using one tool but not the other) with either higher or lower cut-off thresholds.

Notable outliers where the scores were very different using the two tools (Figure 3) were one study(30) on the outcomes of learning disabilities and epilepsy (ROBINS-I Applicable Score 42%, JBI Applicable Score 0%) and one study(32) on epilepsy and vision disorders (ROBINS-I Applicable Score 18%, JBI Applicable Score -56%). Both of these studies have low numbers of subjects (39 and 45 cases respectively). For both studies, items relating to bias due to confounding were graded as being at high risk of bias when using the ROBINS-I tool, but a low risk of bias using the JBI tool. These differences related to scoring of bias in selection of comparison groups and in measurement of outcomes. For the selection of comparison groups, the ROBINS-I tool items were scored as uncertain or not applicable, but the JBI tool items were scored as high risk of bias. For the measurement of outcomes, the ROBINS-I tool items were scored as low risk of bias, while the JBI tool items were scored as uncertain.

Discussion

Both of the domain-based assessment tools we considered performed well for systematically assessing studies that report on outcome prevalence and provided similar discrimination between studies with higher and lower risk of bias. Although the selection of a threshold for inclusion or exclusion of prevalence studies is subjective, the application of a standardised

risk of bias assessment before selecting a threshold does allow discrimination between the upper and lower ranges of risk of bias among the candidate studies.

Although presented with different wording, the component questions of the ROBINS-I and JBI assessment tools include variations on the same concepts, and overlap in a number of domains. Both tools address overall applicability, selection, and description of the study population(s), reporting and appropriateness of sample size and statistical analyses, risk of bias due to measurement of outcomes, and the response rate/missing data.

Both assessment tools were perceived as being simple to use, with a minimal learning curve. Numerically, the ROBINS-I tool (both original and modified) includes more components that need to be considered to complete the domain-level assessments and covers greater breadth of potential bias domains. The JBI tool, however, was designed to specifically critique studies including reports of prevalence, and its component items may be more focussed on this goal.

Although neither tool is designed to output a numeric score, both tools gave similar results using any of the three different scoring systems that we devised to determine whether particular studies should be included or excluded from use in estimating prevalence of an outcome, particularly at higher thresholds. The distributions of scores were wide enough with both tools to allow selection for inclusion at a number of different thresholds. This convergent validity supports both tools for the purpose of assessing risk of bias and selecting studies that report prevalence. The selection of a specific threshold may be based on the number of applicable studies available or the relative or absolute number needed for inclusion.

The two studies assessed as having very different scores using the two tools had low population numbers, which the JBI tool penalises to a greater extent that ROBINS-I, and both included “unclear”/”unknown” responses in their JBI assessments, which reduces the

denominator in the Applicable Score calculation, thus increasing the impact of the remaining assessment items on the score calculation. The specific differences between the ROBINS-I and JBI tools that accounted for the different scores were i) those related to data being gathered from both a sample and control and the potential for confounding due to patient characteristics (covered in ROBINS-I) as compared to measuring outcomes in a sample population only (JBI), and ii) those related to blinding of outcome assessors (covered only in ROBINS-I). Scores are therefore lower using the ROBINS-I tool for reports of prevalence in studies that measure outcomes in an exposure population, but not a non-exposed control group, and studies in which the assessor is not blind to the exposure. Such blinding may not be practical in many of the studies in which outcome prevalences are reported.

These differences between tools are likely to be more important where a lower threshold for inclusion is used, either because most available studies are at higher risk of bias, or there are few studies available reporting a particular outcome. The JBI tool may be easier to interpret and apply to studies where a control group is not present, whereas the ROBINS-I tool addresses a slightly wider range of parameters related to overall methodological quality when assessing studies that compare the rates of outcomes between an exposure group and a control population.

Limitations

Conversion of the tools to numeric scores does not apply any differential weighting to the assessment domains. Arguably, this may result in inclusion of some studies with severe bias in a critical domain or exclusion of studies with bias in domains that the researcher considers less critical for the purposes of the planned economic analysis. However, forming a numeric score does not preclude researchers also using qualitative assessments before making a final

decision. Where many potential data sources exist, risk of bias tools may supplement such judgements by suggesting an initial ordering of candidate studies.

We utilised published risk-of bias assessment tools; one modified and one unmodified. Although we did not assess the impact of our modifications of the ROBINS-I tool, addition of methodological domains not present in the original versions may be useful for other researchers to including domains deemed relevant for the purpose of a planned economic analysis.

In assessing these tools, studies were assessed by each researcher using one tool immediately followed by the other in a consistent order. For assessment items that are similar, responses to one tool may therefore have influenced responses using the second tool.

Summary

Either the ROBINS-I or JBI risk-of-bias assessment tools can be used to select observational studies reporting prevalence for inclusion in an economic analysis. The results of the risk-of-bias assessments can be converted into numerical scores, and thresholds for inclusion can be selected at an appropriate level to include more or fewer studies as required. Particularly at higher thresholds, the studies selected for inclusion are generally the same using either tool. However, the JBI tool is slightly shorter and may be easier to interpret and apply to studies that do not involve a control group, but the ROBINS-I tool assesses more methodological detail particularly in studies that include a control group.

References

1. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC) version 5.02016 [cited 2018 31 July]. Epub September 2016. Available from: https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf

2. Saramago P, Manca A, Sutton AJ. Deriving input parameters for cost-effectiveness modeling: taxonomy of data types and approaches to their statistical synthesis. Value in Health. 2012;15(5):639-49.

3. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices - Modeling Studies. Value Health. 2003;6(1):9-17.

4. Naglie G, Krahn MD, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 3 - Estimating probabilities and utilities. Med Decis Making. 1997;17(2):136-41.

5. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.

6. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

7. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173.

8. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):666-76.

9. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2018.

10. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015;13(3):147-53.

11. Munn Z, Moola S, Riitano D, Lisy K. The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. Int J Health Policy Manag. 2014;3(3):123-8.

12. The Joanna Briggs Institute. Checklist for Prevalence Studies http://joannabriggs.org/: University of Adelaide; 2016 [updated 21 Feb 2018]. Available from: http://joannabriggs-webdev.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Prevalence_Studies.pdf

13. Pildes R, Forbes AE, O'Connor SM, Cornblath M. The incidence of neonatal hypoglycemia - a completed survey. J Pediatr. 1967;70(1):76-80.

14. Harris DL, Weston PJ, Harding JE. Incidence of neonatal hypoglycemia in babies identified as at risk. J Pediatr. 2012;161(5):787-91.

15. Caksen H, Guven AS, Yilmaz C, Unal O, Basaranoglu M, Sal E, et al. Clinical outcome and magnetic resonance imaging findings in infants with hypoglycemia. J Child Neurol. 2011;26(1):25-30.

16. Kaiser JR, Bai S, Gibson N, Holland G, Lin TM, Swearingen CJ, et al. Association between transient newborn hypoglycemia and fourth-grade achievement test proficiency: a population-based study. JAMA Pediatr. 2015;169(10):913-21.

17. Razaz N, Tedroff K, Villamor E, Cnattingius S. Maternal body mass index in early pregnancy and risk of epilepsy in offspring. JAMA Neurol. 2017;74(6):668-76.

18. Burns CM, Rutherford MA, Boardman JP, Cowan FM. Patterns of cerebral injury and neurodevelopmental outcomes after symptomatic neonatal hypoglycemia. Pediatrics. 2008;122(1):65-74.

19. Filan PM, Inder TE, Cameron FJ, Kean MJ, Hunt RW. Neonatal hypoglycemia and occipital cerebral injury. J Pediatr. 2006;148(4):552-5.

20. Goode RH, Rettiganti M, Li J, Lyle RE, Whiteside-Mansell L, Barrett KW, et al. Developmental outcomes of preterm infants with neonatal hypoglycemia. Pediatrics. 2016;138(6).

21. Harris DL, Alsweiler JM, Ansell JM, Gamble GD, Thompson B, Wouldes TA, et al. Outcome at 2 years after dextrose gel treatment for neonatal hypoglycemia: follow-up of a randomized trial. J Pediatr. 2016;170:54-9 e2.

22. Karimzadeh P, Tabarestani S, Ghofrani M. Hypoglycemia-occipital syndrome: a specific neurologic syndrome following neonatal hypoglycemia? J Child Neurol. 2011;26(2):152-9.

23. Kerstjens JM, Bocca-Tjeertes IF, de Winter AF, Reijneveld SA, Bos AF. Neonatal morbidities and developmental delay in moderately preterm-born children. Pediatrics. 2012;130(2):e265-72.

24. Khan MR, Maheshwari PK, Shamim H, Saleem AF, Ahmed S, Ali SR, et al. Neurodevelopmental outcomes of premature infants at a tertiary care center in Pakistan. Pediatr Neurol. 2012;47(2):109-13.

25. Koivisto M, Blanco-Sequeiros M, Krause U. Neonatal symptomatic and asymptomatic hypoglycaemia: a follow-up study of 151 children. Dev Med Child Neurol. 1972;14(5):603-14.

26. Lucas A, Morley R, Cole TJ. Adverse neurodevelopmental outcome of moderate neonatal hypoglycaemia. BMJ. 1988;297(6659):1304-8.

27. McKinlay CJ, Alsweiler JM, Ansell JM, Anstice NS, Chase JG, Gamble GD, et al. Neonatal glycemia and neurodevelopmental outcomes at 2 years. N Engl J Med. 2015;373(16):1507-18.

28. Montassir H, Maegaki Y, Ogura K, Kurozawa Y, Nagata I, Kanzaki S, et al. Associated factors in neonatal hypoglycemic brain injury. Brain Dev. 2009;31(9):649-56.

29. Murakami Y, Yamashita Y, Matsuishi T, Utsunomiya H, Okudera T, Hashimoto T. Cranial MRI of neurologically impaired children suffering from neonatal hypoglycaemia. Pediatr Radiol. 1999;29(1):23-7.

30. Pildes RS, Cornblath M, Warren I, Page-El E, Di Menza S, Merritt DM, et al. A prospective controlled study of neonatal hypoglycemia. Pediatrics. 1974;54(1):5-14.

31. Sujatha R, Jain N. Prediction of neurodevelopmental outcome of preterm babies using risk stratification score. Indian J Pediatr. 2016;83(7):640-4.

32. Tam EW, Widjaja E, Blaser SI, Macgregor DL, Satodia P, Moore AM. Occipital lobe injury and cortical visual outcomes after neonatal hypoglycemia. Pediatrics. 2008;122(3):507-12.

33. Tin W, Brunskill G, Kelly T, Fritz S. 15-year follow-up of recurrent "hypoglycemia" in preterm infants. Pediatrics. 2012;130(6):e1497-503.

34. Tottman AC, Alsweiler JM, Bloomfield FH, Pan M, Harding JE. Relationship between measures of neonatal glycemia, neonatal illness, and 2-year outcomes in very preterm infants. J Pediatr. 2017;188:115-21.

35. Yalnizoglu D, Haliloglu G, Turanli G, Cila A, Topcu M. Neurologic outcome in patients with MRI pattern of damage typical for neonatal hypoglycemia. Brain Dev. 2007;29(5):285-92.

Figure captions

Figure 1. Distribution of applicable scores for different outcomes using the ROBINS-I tool.

Figure 2. Distribution of applicable scores for different outcomes using the JBI tool.

Figure 3. Agreement Between Applicable Scores for Assessment of Different Studies and Outcomes using ROBINS-I and JBI Tools

Supplementary table captions

Supplementary Table 1. Example of scoring system calculations.

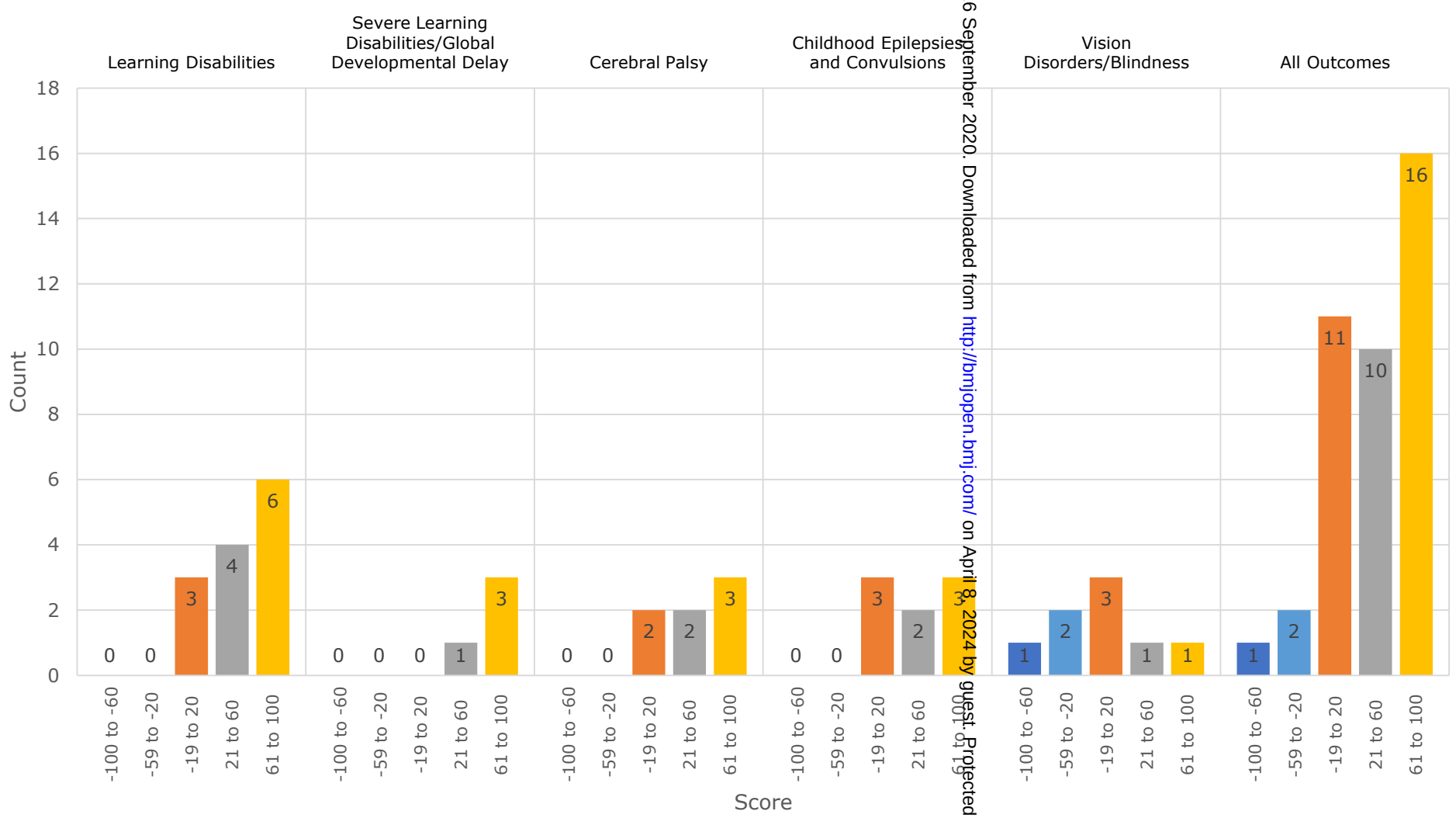
Supplementary Table 2. Initial inter-assessor discrepancies using each assessment tool.

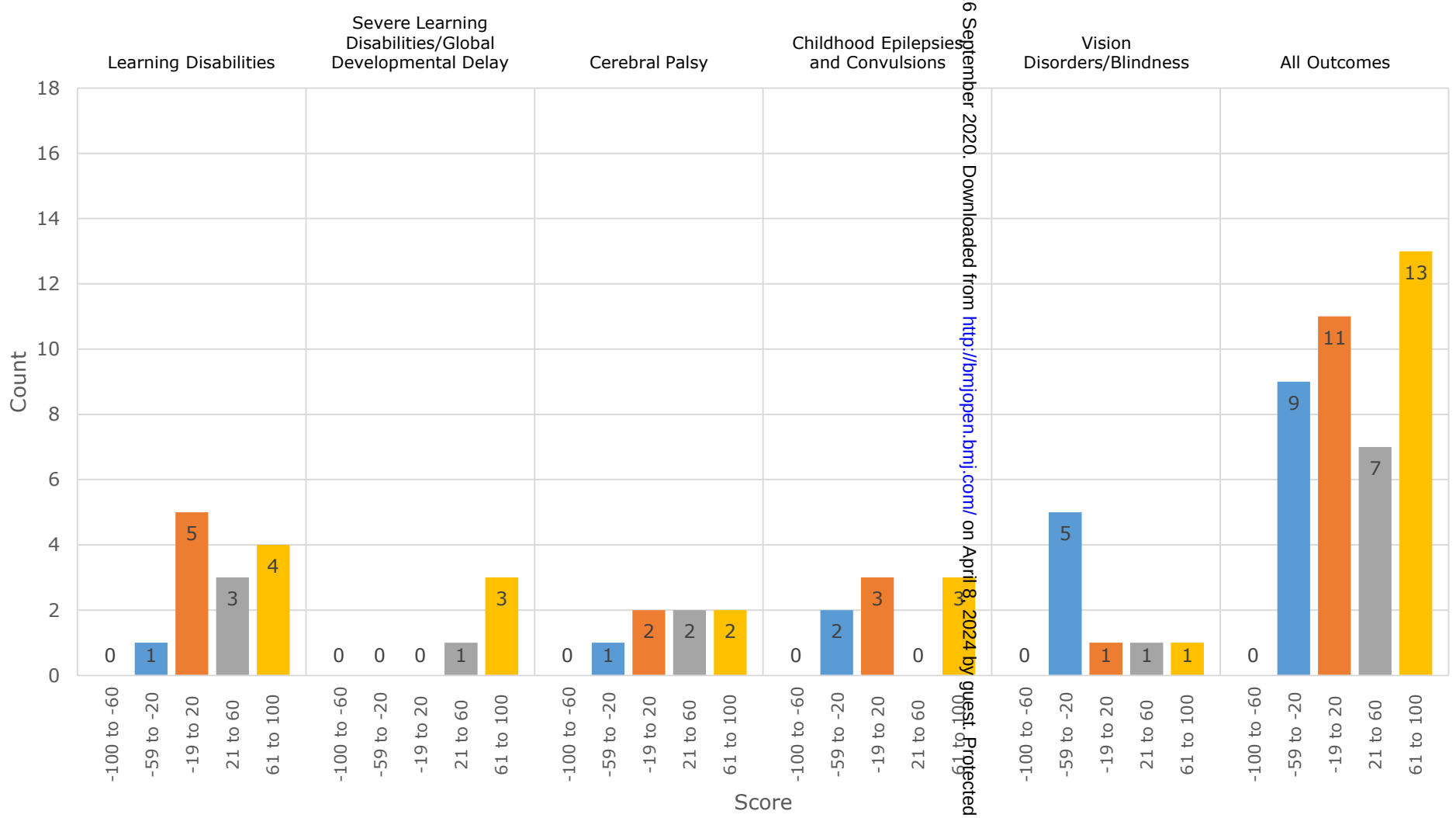
Contributorship statement

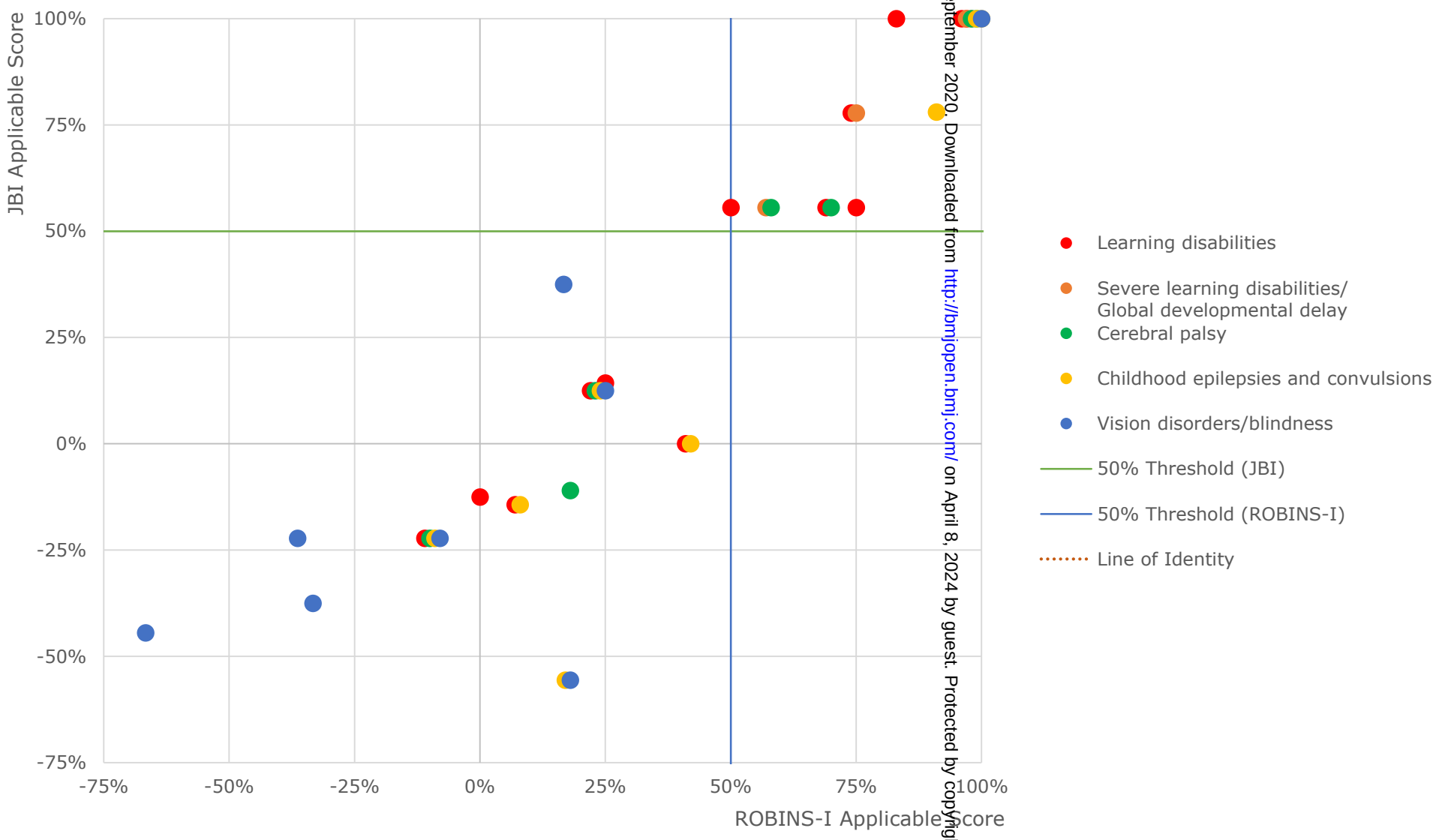
All listed authors contributed to the planning, conduct, and reporting of the accompanying work. MJ Glasgow wrote the first draft, while R Edlin and JE Harding contributed supervision/oversight, and review and editing of the manuscript. Each author listed has seen and approved the submission of this version of the manuscript and takes full responsibility for the manuscript.

Data availability statement

All data relevant to the study are included in the article or uploaded as supplementary information.







Note: overlapping data points have been spread to allow for visibility. The highest data point in an overlapping cluster is the actual value.

Score calculation method	Low risk of bias	High risk of bias	Uncertain	Not applicable
Item 1			x	
Item 2	x			
Item 3	x			
Item 4			x	
Item 5	x			
Item 6	x			
Item 7				x
Item 8		x		
Item 9		x		
Item 10			x	
Item 11	x			
Item 12	x			
TOTALS	6	2	3	1
1. Count Scores	Positive: 6/12 = 50% Negative: 2/12 = 17%			
2. Composite Score	(6-2)/12 = 33%			
3. Applicable Score	(6-2)/(12-1) = 36%			

Supplementary Table 1: Example of scoring system calculations

Assessment tool	ROBINS-I	JBI	Total
Study	Number of conflicts between 2 or more assessors		
Applicability classification conflict			
Caksen et al 2011[15]	10	6	16
Kaiser et al 2015[16]	2	2	2
Razaz et al 2017[17]	5	5	10
High-low conflicts			
Caksen et al 2011[15]	2	1	3
Kaiser et al 2015[16]	1	0	1
Razaz et al 2017[17]	1	0	1

Supplementary Table 1: Initial inter-assessor discrepancies using each assessment tool

BMJ Open

Comparison of risk -of-bias assessment approaches for selection of studies reporting prevalence for economic analyses

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037324.R1
Article Type:	Original research
Date Submitted by the Author:	22-Jul-2020
Complete List of Authors:	Glasgow, Matthew; The University of Auckland, Liggins Institute Edlin, Richard; University of Auckland, Health Systems, School of Population Health Harding, Jane; University of Auckland Liggins Institute
Primary Subject Heading:	Health economics
Secondary Subject Heading:	Paediatrics
Keywords:	HEALTH ECONOMICS, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Comparison of risk -of-bias assessment approaches for selection of studies reporting prevalence for economic analyses

Matthew J Glasgow [1], Richard Edlin [2], Jane E Harding [1*]

[1] Liggins Institute, University of Auckland, Auckland, New Zealand

[2] School of Population Health, University of Auckland, Auckland, New Zealand

[*] Corresponding author (e-mail: j.harding@auckland.ac.nz)

Financial support for this study was provided in part by a grant (Douglas Goodfellow Medical Research Fellowship, grant number 1417003) from the Auckland Medical Research Foundation. The funding agreement ensured the authors' independence in study design, data collection, data analysis, data interpretation, writing of the report, and the decision to submit the paper for publication. The authors declare no conflicts of interest. Each author listed on the manuscript has seen and approved the submission of this version of the manuscript and takes full responsibility for the manuscript.

The authors declare no competing interests.

Word Count: 3550

Abstract

Objectives: Within cost-effectiveness models, prevalence figures can inform transition probabilities. The methodological quality of studies can inform the choice of prevalence figures but no single obvious candidate tool exists for assessing quality of the observational epidemiological studies for selecting prevalence estimates. We aimed to compare different tools to assess the risk of bias of studies reporting prevalence, and develop and compare possible numerical scoring systems using these tools to set a threshold for inclusion of reports of prevalence in an economic analysis of neonatal hypoglycaemia.

Design: Assessments of bias using two tools (JBI Checklist for Prevalence Studies and a modified version of ROBINS-I) were compared for 18 studies relevant to a single setting (neonatal hypoglycaemia). Inclusion of studies for use in a decision analysis model were considered based on summary scores derived from these tools.

Results: Both tools were considered easy to use and produced dispersed scores for each of the 40 study-outcome combinations. The modified ROBINS-I scores were more skewed than the JBI scores, particularly at higher thresholds. The studies selected for inclusion are generally the same using either tool; if 50% was used as the cut-off threshold using the Applicable Score both tools would yield the same results. However, the JBI tool is shorter and may be easier to interpret and apply to studies that do not involve a control group, while the modified ROBINS-I tool assesses more methodological detail in studies that include a control group.

Conclusion: Both tools performed well for systematically assessing studies that report on outcome prevalence and provided similar discrimination between studies for risk of bias. This convergent validity supports use of both tools for the purpose of assessing risk of bias and selecting studies that report prevalence for inclusion in economic analyses.

Strengths and limitations of this study

- This study addresses a methodological task for which no single obvious candidate tool exists.
- Assessments of candidate tools and approaches to use of the tools were undertaken independently by the three authors.
- Convergent validity between the tools examined supports the use of either approach to guide the inclusion of prevalence reports in economic modelling.
- Studies were assessed by each researcher using one tool immediately followed by the other in a consistent order. For assessment items that are similar, responses to one tool may therefore have influenced responses using the second tool.

Introduction

The probability of an outcome occurring is a fundamental parameter required in the creation of a decision analytic model. It represents the likelihood that patients in a cohort will move from one health state to another in a decision tree or state transition model (e.g., Markov model), and is thus often referred to as a transition probability. [1] When referring to clinical outcomes, the transition probability is equivalent to the prevalence of that outcome in the population represented in the model.

The evidence base from which model parameters are drawn often involves more than a single data source, and developing the model may involve aggregation of this data. [2] The process of deciding which values to use as the key inputs in a model, including the transition probabilities, should be based on a systematic review of the literature, and a description of this process should accompany the model, [3-5] with the use of a source and any translational steps justified. [1, 4] The use of published studies as a source for transition probabilities should have their validity transparently assessed by applying critical appraisal criteria. [4]

In 2016, Sterne et al observed that, in terms of assessing study validity, there has been a shift in focus away from analysis of methodological quality to assessments of risk of bias, often in a domain-oriented manner i.e., considering different domains of bias in turn. [6] The potential for bias and types of bias in non-randomised studies may differ from those in randomised studies. [7] A number of instruments have been developed for assessing the risk of bias in non-randomised studies. [7] In 2003, Deeks et al identified six that were considered to have utility for systematic reviews, although they noted that none had been formally validated. [7] In 2007, Sanderson et al concluded that there was a lack of a single obvious candidate tool for assessing quality of observational epidemiological studies. [8] They identified three domains as being fundamental in assessing risk of bias (appropriate selection of patients, appropriate

measurement of variables, and appropriate control of confounding), but noted that these were present in only approximately half of the checklists that they evaluated. [8]

Subsequent to these systematic reviews, Sterne et al developed the ROBINS-I ("Risk Of Bias In Non-randomised Studies - of Interventions") tool to evaluate the risk of bias in studies that do not use randomisation to allocate participants to comparison groups. [6] The ROBINS-I includes a total of 7 bias domains: selection of comparison groups, confounding, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result. These domains can be further compartmentalised into pre-intervention (confounding and participant selection), intervention (classification of interventions), and post-intervention (the remainder) categories. [6] The ROBINS-I assesses risk of bias using an absolute scale, as distinct to the approach commonly used by other similar tools of comparing against a theoretical, perfect observational study or a high quality randomised trial. [9] ROBINS-I was constructed with an objective of allowing the risk-of-bias assessment to determine the degree to which the rating of a study is downgraded. [6] This would facilitate comparison between ratings of randomised trials and ratings of non-randomised studies when using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. If we consider the intervention to be an exposure (e.g., the occurrence of neonatal hypoglycaemia), the ROBINS-I provides a systematic approach that can assess a non-interventional observational study for risk of bias within the seven specified domains.

In 2015, Munn et al observed a lack of guidance for authors undertaking systematic reviews of observational epidemiological studies, including those reporting prevalence or incidence information. [10] That absence of guidance included the lack of a standard method for conducting critical appraisals of the studies used in systematic reviews of prevalence data. [11] The same authors also observed a significant increase in the volume of systematic

reviews being performed and published that focused on questions of prevalence. [11] This combination of factors led to the establishment of a working group, composed of researchers from the Joanna Briggs Institute (JBI, University of Adelaide, Australia), to create guidance for conducting systematic reviews of studies reporting incidence and prevalence parameters. [10] This guidance has been published as a checklist with supporting explanatory information. [12] When applied to prevalence studies, reported risks of bias in the JBI tool cover an array of concepts similar to those in the ROBINS-I tool.

The ROBINS-I and JBI tools were selected for comparison in this study in light of the conclusions by Sanderson et al in their comprehensive 2007 systematic review that, despite the existence at that time of 86 candidate tools developed to assess the quality of evidence from observational epidemiological studies, none could be recommended as a single ideal candidate. [8] Both the ROBINS-I and JBI tools were developed subsequent to that review, and address a number of the recommendations from Sanderson et al, particularly those relating to rigour in their development and appropriate coverage of key domains.

The ROBINS-I domain pertaining to bias in ascertainment of exposures is notably lacking from the JBI tool, which was not designed with the explicit intent of assessing reports of prevalence after a nominated exposure. It does not explicitly inquire about such concepts as whether exposure was measured prior to determination of outcome; whether exposure measures were defined, reliable, and consistently applied; whether different levels of exposure were considered; or whether the exposure was assessed more than once over time. The JBI tool also does not explicitly address bias in reporting of results, particularly the implications of performing multiple measurements or analyses of the exposure-outcome relationship.

Conversely, although the ROBINS-I tool does assess a number of concepts related to measurement of the outcomes, it does not explicitly examine the validity of outcome ascertainment, and it does not downgrade on the basis of sample size alone. Further, the ROBINS-I tool contains a series of assessment items examining the appropriateness of methods for selecting a control group; a topic not included in the JBI tool.

Differences in prevalence for the same or similar outcomes vary for a number of reasons, including methodological differences, differences in definitions of the outcomes, and differences in the populations being examined. We wished to select published reports of prevalence of outcomes of neonatal hypoglycaemia for use in a decision analytic model for an economic analysis. A wide range of prevalence figures have been reported for these outcomes, in part because of inconsistencies in the definition of neonatal hypoglycaemia, particularly the blood glucose concentration threshold used to diagnose asymptomatic cases, changes in that definition over time, and differences in approaches to screen for and identify the condition. The blood glucose concentration threshold for diagnosing neonatal hypoglycaemia has ranged from, 20 mg/100mL (1.11 mmol/L) [13] in earlier studies to 2.6 mmol/L [14], and has variably included additional criteria such as a requirement for low results on consecutive measurements.

In economic analyses, each prevalence parameter needs to be informed by the available information even if the underlying quality of information is of poor quality. This means that the question becomes how to decide which sources of information to include and not include for each outcome, rather than determining a single inclusion threshold across all studies. We first undertook this study to examine the use of risk-of-bias assessments to assist with these decisions.

Objective

We aimed to 1) undertake a comparison of different tools to assess the risk of bias of studies reporting prevalence for use as data sources for economic analyses, and 2) develop and compare possible numerical scoring systems using these tools to set a threshold for inclusion of reports of prevalence in an economic analysis, using the example of outcomes of neonatal hypoglycaemia.

Methods

Both the Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I) tool [6] and the Joanna Briggs Institute Checklist for Prevalence Studies [12] were selected for initial assessment. We chose these two tools based on their applicability to observational studies and/or studies reporting prevalence, consistency with the GRADE approach to assessment of uncertainty, and advice from local researchers familiar with candidate instruments. A modified version of the ROBINS-I tool was pre-formatted into a spreadsheet for ease of use by assessors. The ROBINS-I assessment item pertaining to the bias domain of deviations from intended interventions was excluded as the topic of interest was an exposure at a point in time rather than an intervention over time. Instead, we added three assessment items pertaining to the domain of study design (clarity of the statement of objective, inclusion of sample size justification or similar, inclusion of an unexposed group) and three pertaining to external validity (specification of the study population, relevance of the cohort to the target population, and drop-out rate) (Supplementary Table 1). For each domain the overall bias was summarised as high, low, or uncertain.

From the pool of non-randomised studies that reported, or allowed the calculation of, prevalence of outcomes of neonatal hypoglycaemia, we selected three that covered a range of methodologies and study population sizes and focused on a single outcome. [15-17] All three

researchers assessed these three studies using both assessment tools, discussed discrepancies and reached consensus on how the questions should be interpreted. A further 18 studies [18-35] reporting prevalence of outcomes after neonatal hypoglycaemia were then each assessed by combinations of two of the three researchers using both tools.

Three summary scores were formed to facilitate further comparison between studies (Supplementary Table 2).

1. Count Score(s): sums of responses in each column. That is, the total number of responses indicating low risk of bias, and the total number of responses indicating high risk of bias. Two separate values are thus generated. For the sum of responses indicating low risk of bias, a higher score represents a low risk of bias; for the sum of responses indicating a high risk of bias, a higher score represents a high risk of bias. These are presented as a percentage of the total value possible on the tool. (Note that the total number of questions, and therefore the maximum total value, is 12 on the modified ROBINS-I tool and 9 on the JBI tool).
2. Composite Score: calculated by subtracting the total number of responses indicating high risk of bias from the total number of responses indicating low risk of bias. A higher score represents a lower risk of bias. Negative values are possible for studies that score a greater number of high risk of bias elements/domains than low risk of bias elements. This is presented as a percentage of the total value possible on the tool.
3. Applicable Score: conversion of the Composite Score into a percentage by dividing the Composite Score by the maximum score possible after subtracting any “not applicable” responses. A higher score represents a lower risk of bias. Negative values are also possible using this approach.

All three scores have a maximum value of 100%. Excluding "not applicable" responses in the Applicable Score is intended to more accurately reflect which elements of the tool are relevant to the study being assessed.

Patient and public involvement

This work is a research methods paper, and as such was undertaken without patient involvement.

Results

Ease of use and assessor agreement for initial three studies

All three researchers reported that the assessment tools and spreadsheets were generally easy to use, and that, because of the structural and content similarities between the two tools (Supplementary Table 1), assessment using both tools did not result in a large increase in time required compared to assessment using a single tool. However, since the JBI tool includes fewer assessment items it may have a modest time advantage over the modified ROBINS-I tool.

After the initial training assessment of 3 studies, chance corrected AC1 agreement, a more valid measure of inter-rater reliability than the Kappa statistic[36] between the two assessors ranged from 0.51 (95% CI 0.18, 0.84) to 0.93 (95% CI 0.81, 1.00) for the modified ROBINS-I and 0.39 (95% CI 0.06, 0.71) to 0.79 (95% CI 0.56, 1) for the JBI tool. There were no consistent patterns in the assessment fields for which scores were discrepant across the 12 modified ROBINS-I or 9 JBI domains from 18 studies. All discrepancies were resolved by discussion before inclusion in subsequent evaluation.

Assessment tool scores and agreement

When used by combinations of two researchers to assess 40 study-outcome combinations (hereafter "assessments") from the 18 studies, both the modified ROBINS-I and JBI tools resulted in a wide distribution of scores for each outcome (Figures 1 and 2), potentially allowing selection of studies for inclusion at a wide range of thresholds. The distribution of scores with the modified ROBINS-I tool was generally skewed slightly higher than the distribution of scores with the JBI tool.

Using the Count Scores, the difference between the two tools in the number of studies selected for inclusion or exclusion varies with the threshold in a non-linear manner (Table 1). For lower thresholds (e.g., 25%), there is greater difference between the two tools than for higher thresholds (e.g., 50%, 75%), with more studies being included using the modified ROBINS-I than using the JBI at the lower thresholds.

	Scoring system:	1. Count Score, Positive			2. Composite Score			3. Applicable Score		
Studies included	Threshold:	25%	50%	75%	25%	50%	75%	25%	50%	75%
All assessment types (of 40 assessments)	ROBINS-I	33	22	12	21	18	11	21	18	11
	JBI	23	19	14	20	19	13	20	19	13
Learning disabilities (of 13 assessments)	ROBINS-I	12	9	4	8	6	3	8	6	3
	JBI	9	7	5	7	7	4	7	7	4

Severe learning disabilities (of 4 assessments)	ROBINS-I	4	4	2	4	4	2	4	4	2
	JBI	4	4	3	4	4	3	4	4	3
Cerebral palsy (of 7 assessments)	ROBINS-I	6	4	2	4	4	2	4	4	2
	JBI	4	4	2	4	4	2	4	4	2
Epilepsy; seizures (of 8 assessments)	ROBINS-I	7	4	3	4	3	3	4	3	3
	JBI	4	3	3	3	3	3	3	3	3
Vision disorders; blindness (of 8 assessments)	ROBINS-I	4	1	1	1	1	1	1	1	1
	JBI	2	1	1	2	1	1	2	1	1

Table 1: Number of studies selected for inclusion when assessing different outcomes using three different scoring systems of the ROBINS-I and JBI tools at different thresholds

Using the Composite or the Applicable Scores, the modified ROBINS-I and JBI tool each resulted in very similar numbers of studies included (Table 1). If 50% was used as the cut-off threshold for inclusion or exclusion of studies based on their risk of bias, both tools would give the same results using the Applicable Score (Figure 3). The level of agreement fell (i.e., some studies would be included using one tool but not the other) with either higher or lower cut-off thresholds.

Notable outliers where the scores were very different using the two tools (Figure 3) were one study(30) on the outcomes of learning disabilities and epilepsy (modified ROBINS-I Applicable Score 42%, JBI Applicable Score 0%) and one study(32) on epilepsy and vision disorders (modified ROBINS-I Applicable Score 18%, JBI Applicable Score -56%). Both of these studies have low numbers of subjects (39 and 45 cases respectively). For both studies, items relating to bias due to confounding were graded as being at high risk of bias when using the modified ROBINS-I tool, but a low risk of bias using the JBI tool. These differences related to scoring of bias in selection of comparison groups and in measurement of outcomes. For the selection of comparison groups, the modified ROBINS-I tool items were scored as uncertain or not applicable, but the JBI tool items were scored as high risk of bias. For the measurement of outcomes, the modified ROBINS-I tool items were scored as low risk of bias, while the JBI tool items were scored as uncertain.

Discussion

Both of the domain-based assessment tools we considered performed well for systematically assessing studies that report on outcome prevalence and provided similar discrimination between studies with higher and lower risk of bias. Although the selection of a threshold for inclusion or exclusion of prevalence studies is subjective, the application of a standardised risk of bias assessment before selecting a threshold does allow discrimination between the upper and lower ranges of risk of bias among the candidate studies.

Although presented with different wording, the component questions of the modified ROBINS-I and JBI assessment tools include variations on the same concepts, and overlap in a number of domains. Both tools address overall applicability, selection, and description of the study population(s), reporting and appropriateness of sample size and statistical analyses, risk of bias due to measurement of outcomes, and the response rate/missing data.

Both assessment tools were perceived as being simple to use, with a minimal learning curve; after an initial set of training assessments, agreement between assessors was high and unanimity was readily reached with brief discussion where required. Numerically, the ROBINS-I tool (both original and modified) includes more components that need to be considered to complete the domain-level assessments and covers greater breadth of potential bias domains. The JBI tool, however, was designed to specifically critique studies including reports of prevalence, and its component items may be more focussed on this goal.

Although neither tool is designed to output a numeric score, both tools gave similar results using any of the three different scoring systems that we devised to determine whether particular studies should be included or excluded from use in estimating prevalence of an outcome, particularly at higher thresholds. The distributions of scores were wide enough with both tools to allow selection for inclusion at a number of different thresholds, or to stratify studies into different levels of risk-of-bias as a component of consideration for inclusion. This convergent validity supports both tools for the purpose of assessing risk of bias and selecting studies that report prevalence. The selection of a specific threshold may be based on the number of applicable studies available or the relative or absolute number needed for inclusion.

The two studies assessed as having very different scores using the two tools had low population numbers, which the JBI tool penalises to a greater extent than the modified ROBINS-I, and both included “unclear”/”unknown” responses in their JBI assessments, which reduces the denominator in the Applicable Score calculation, thus increasing the impact of the remaining assessment items on the score calculation. The specific differences between the modified ROBINS-I and JBI tools that accounted for the different scores were i) those related to data being gathered from both a sample and control and the potential for confounding due to patient characteristics (covered in modified ROBINS-I) as compared to

measuring outcomes in a sample population only (JBI), and ii) those related to blinding of outcome assessors (covered only in modified ROBINS-I). Scores are therefore lower using the modified ROBINS-I tool for reports of prevalence in studies that measure outcomes in an exposure population, but not a non-exposed control group, and studies in which the assessor is not blind to the exposure. Such blinding may not be practical in many of the studies in which outcome prevalences are reported.

These differences between tools are likely to be more important where a lower threshold for inclusion is used, either because most available studies are at higher risk of bias, or there are few studies available reporting a particular outcome. The JBI tool may be easier to interpret and apply to studies where a control group is not present, whereas the modified ROBINS-I tool addresses a slightly wider range of parameters related to overall methodological quality when assessing studies that compare the rates of outcomes between an exposure group and a control population.

Limitations

Conversion of the tools to numeric scores does not apply any differential weighting to the assessment domains. Arguably, this may result in inclusion of some studies with severe bias in a critical domain or exclusion of studies with bias in domains that the researcher considers less critical for the purposes of the planned economic analysis. However, forming a numeric score does not preclude researchers also using qualitative assessments before making a final decision. Where many potential data sources exist, risk of bias tools may supplement such judgements by suggesting an initial ordering of candidate studies.

We utilised published risk-of bias assessment tools; one modified and one unmodified. Although we did not assess the impact of our modifications of the ROBINS-I tool, addition of methodological domains not present in the original versions may be useful for other

researchers to include domains deemed relevant for the purpose of a planned economic analysis.

In assessing these tools, studies were assessed by each researcher using one tool immediately followed by the other in a consistent order. For assessment items that are similar, responses to one tool may therefore have influenced responses using the second tool.

Summary

Either the modified ROBINS-I or JBI risk-of-bias assessment tools can be used to select observational studies reporting prevalence for inclusion in an economic analysis. The results of the risk-of-bias assessments can be converted into numerical scores, and thresholds for inclusion can be selected at an appropriate level to include more or fewer studies as required. Particularly at higher thresholds, the studies selected for inclusion are generally the same using either tool. However, the JBI tool is slightly shorter and may be easier to interpret and apply to studies that do not involve a control group, but the modified ROBINS-I tool assesses more methodological detail particularly in studies that include a control group.

References

1. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC) version 5.02016 21 Feb 2018 [cited 2018 31 July]. Available from: https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf.

2. Saramago P, Manca A, Sutton AJ. Deriving input parameters for cost-effectiveness modeling: taxonomy of data types and approaches to their statistical synthesis. Value in Health. 2012;15(5):639-49.

3. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices - Modeling Studies. Value Health. 2003;6(1):9-17.

4. Naglie G, Krahn MD, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 3 - Estimating probabilities and utilities. Med Decis Making. 1997;17(2):136-41.

5. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.

6. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

7. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173.

8. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):666-76.

9. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2018.

10. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015;13(3):147-53.

11. Munn Z, Moola S, Riitano D, Lisy K. The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. Int J Health Policy Manag. 2014;3(3):123-8.

12. The Joanna Briggs Institute. Checklist for Prevalence Studies http://joannabriggs.org/: University of Adelaide; 2016 [updated 21 Feb 2018; cited 2018 20 March]. Available from: http://joannabriggs-webdev.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Prevalence_Studies.pdf.

13. Pildes R, Forbes AE, O'Connor SM, Cornblath M. The incidence of neonatal hypoglycemia - a completed survey. J Pediatr. 1967;70(1):76-80.

14. Harris DL, Weston PJ, Harding JE. Incidence of neonatal hypoglycemia in babies identified as at risk. J Pediatr. 2012;161(5):787-91.

15. Caksen H, Guven AS, Yilmaz C, Unal O, Basaranoglu M, Sal E, et al. Clinical outcome and magnetic resonance imaging findings in infants with hypoglycemia. J Child Neurol. 2011;26(1):25-30.

16. Kaiser JR, Bai S, Gibson N, Holland G, Lin TM, Swearingen CJ, et al. Association between transient newborn hypoglycemia and fourth-grade achievement test proficiency: a population-based study. JAMA Pediatr. 2015;169(10):913-21.

17. Razaz N, Tedroff K, Villamor E, Cnattingius S. Maternal body mass index in early pregnancy and risk of epilepsy in offspring. JAMA Neurol. 2017;74(6):668-76.

18. Burns CM, Rutherford MA, Boardman JP, Cowan FM. Patterns of cerebral injury and neurodevelopmental outcomes after symptomatic neonatal hypoglycemia. Pediatrics. 2008;122(1):65-74.

19. Filan PM, Inder TE, Cameron FJ, Kean MJ, Hunt RW. Neonatal hypoglycemia and occipital cerebral injury. J Pediatr. 2006;148(4):552-5.

20. Goode RH, Rettiganti M, Li J, Lyle RE, Whiteside-Mansell L, Barrett KW, et al. Developmental outcomes of preterm infants with neonatal hypoglycemia. Pediatrics. 2016;138(6).

21. Harris DL, Alsweiler JM, Ansell JM, Gamble GD, Thompson B, Wouldes TA, et al. Outcome at 2 years after dextrose gel treatment for neonatal hypoglycemia: follow-up of a randomized trial. J Pediatr. 2016;170:54-9 e2.

22. Karimzadeh P, Tabarestani S, Ghofrani M. Hypoglycemia-occipital syndrome: a specific neurologic syndrome following neonatal hypoglycemia? J Child Neurol. 2011;26(2):152-9.

23. Kerstjens JM, Bocca-Tjeertes IF, de Winter AF, Reijneveld SA, Bos AF. Neonatal morbidities and developmental delay in moderately preterm-born children. Pediatrics. 2012;130(2):e265-72.

24. Khan MR, Maheshwari PK, Shamim H, Saleem AF, Ahmed S, Ali SR, et al. Neurodevelopmental outcomes of premature infants at a tertiary care center in Pakistan. Pediatr Neurol. 2012;47(2):109-13.

25. Koivisto M, Blanco-Sequeiros M, Krause U. Neonatal symptomatic and asymptomatic hypoglycaemia: a follow-up study of 151 children. Dev Med Child Neurol. 1972;14(5):603-14.

26. Lucas A, Morley R, Cole TJ. Adverse neurodevelopmental outcome of moderate neonatal hypoglycaemia. BMJ. 1988;297(6659):1304-8.

27. McKinlay C, Alsweiler J, Ansell J, Anstice NS, Chase JG, Gamble GD, et al. Neonatal glycemia and neurodevelopmental outcomes at 2 years. N Engl J Med. 2015;373(16):1507-18.

28. Montassir H, Maegaki Y, Ogura K, Kurozawa Y, Nagata I, Kanzaki S, et al. Associated factors in neonatal hypoglycemic brain injury. Brain Dev. 2009;31(9):649-56.

29. Murakami Y, Yamashita Y, Matsuishi T, Utsunomiya H, Okudera T, Hashimoto T. Cranial MRI of neurologically impaired children suffering from neonatal hypoglycaemia. Pediatr Radiol. 1999;29(1):23-7.

30. Pildes R, Cornblath M, Warren I, Page-El E, Di Menza S, Merritt DM, et al. A prospective controlled study of neonatal hypoglycemia. Pediatrics. 1974;54(1):5-14.

31. Sujatha R, Jain N. Prediction of neurodevelopmental outcome of preterm babies using risk stratification score. Indian J Pediatr. 2016;83(7):640-4.

32. Tam EW, Widjaja E, Blaser SI, Macgregor DL, Satodia P, Moore AM. Occipital lobe injury and cortical visual outcomes after neonatal hypoglycemia. Pediatrics. 2008;122(3):507-12.

33. Tin W, Brunskill G, Kelly T, Fritz S. 15-year follow-up of recurrent "hypoglycemia" in preterm infants. Pediatrics. 2012;130(6):e1497-503.

34. Tottman AC, Alsweiler JM, Bloomfield FH, Pan M, Harding JE. Relationship between measures of neonatal glycemia, neonatal illness, and 2-year outcomes in very preterm infants. J Pediatr. 2017;188:115-21.

35. Yalnizoglu D, Haliloglu G, Turanli G, Cila A, Topcu M. Neurologic outcome in patients with MRI pattern of damage typical for neonatal hypoglycemia. Brain Dev. 2007;29(5):285-92.

36. Detsky AS, Naglie G, Krahn MD, Naimark D, Redelmeier DA. Primer on medical decision analysis: Part 1--Getting started. Med Decis Making. 1997;17(2):123-5.

Figure captions

Figure 1. Distribution of applicable scores for different outcomes using the ROBINS-I tool.

Figure 2. Distribution of applicable scores for different outcomes using the JBI tool.

Figure 3. Agreement Between Applicable Scores for Assessment of Different Studies and Outcomes using ROBINS-I and JBI Tools

Supplementary table captions

Supplementary Table 1: Comparison of domain representation in each assessment tool.

Supplementary Table 2. Example of scoring system calculations.

Contributorship statement

All listed authors contributed to the planning, conduct, and reporting of the accompanying work. MJ Glasgow wrote the first draft, while R Edlin and JE Harding contributed supervision/oversight, and review and editing of the manuscript. Each author listed has seen and approved the submission of this version of the manuscript and takes full responsibility for the manuscript.

Data availability statement

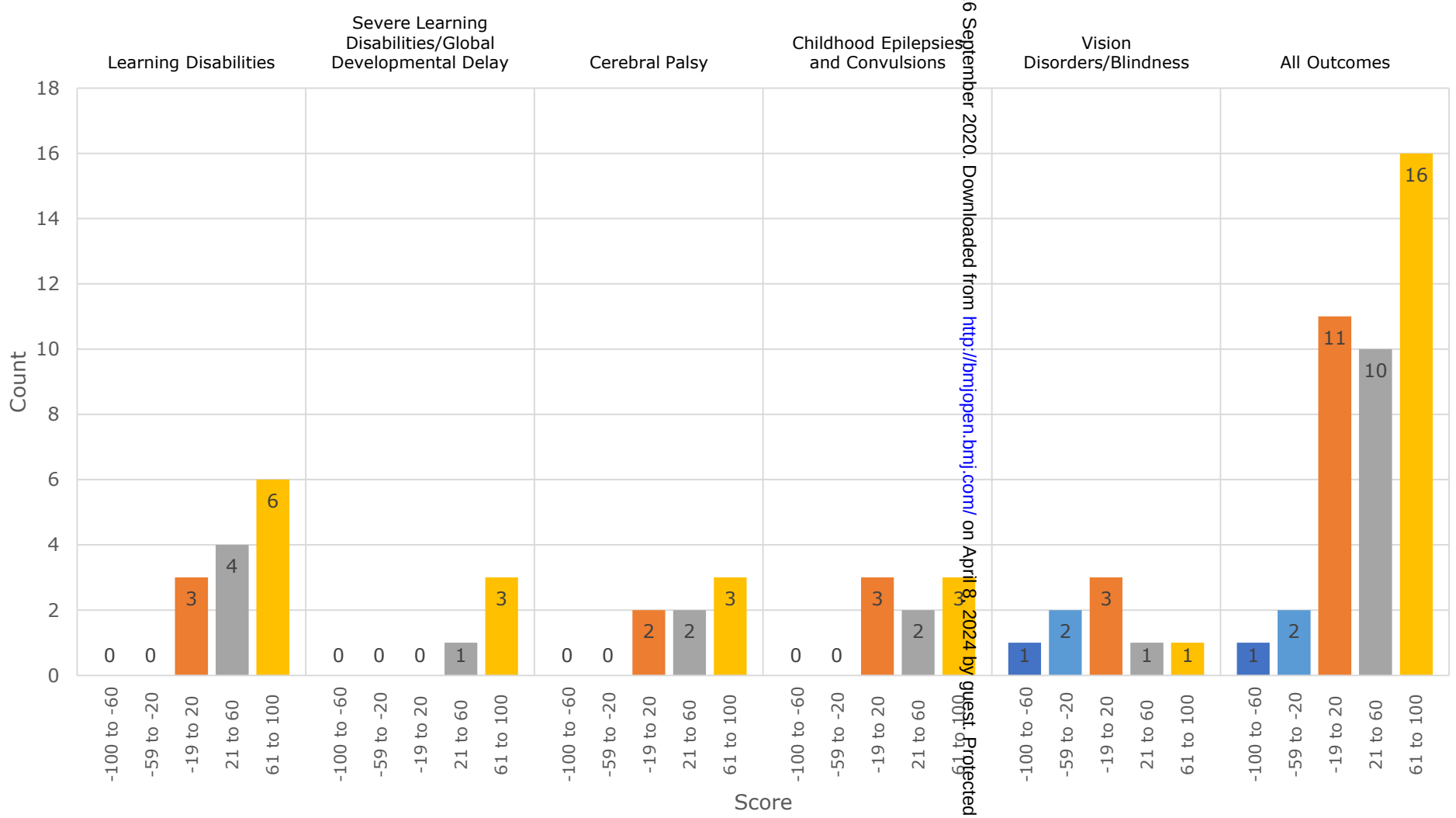
All data relevant to the study are included in the article or uploaded as supplementary information.

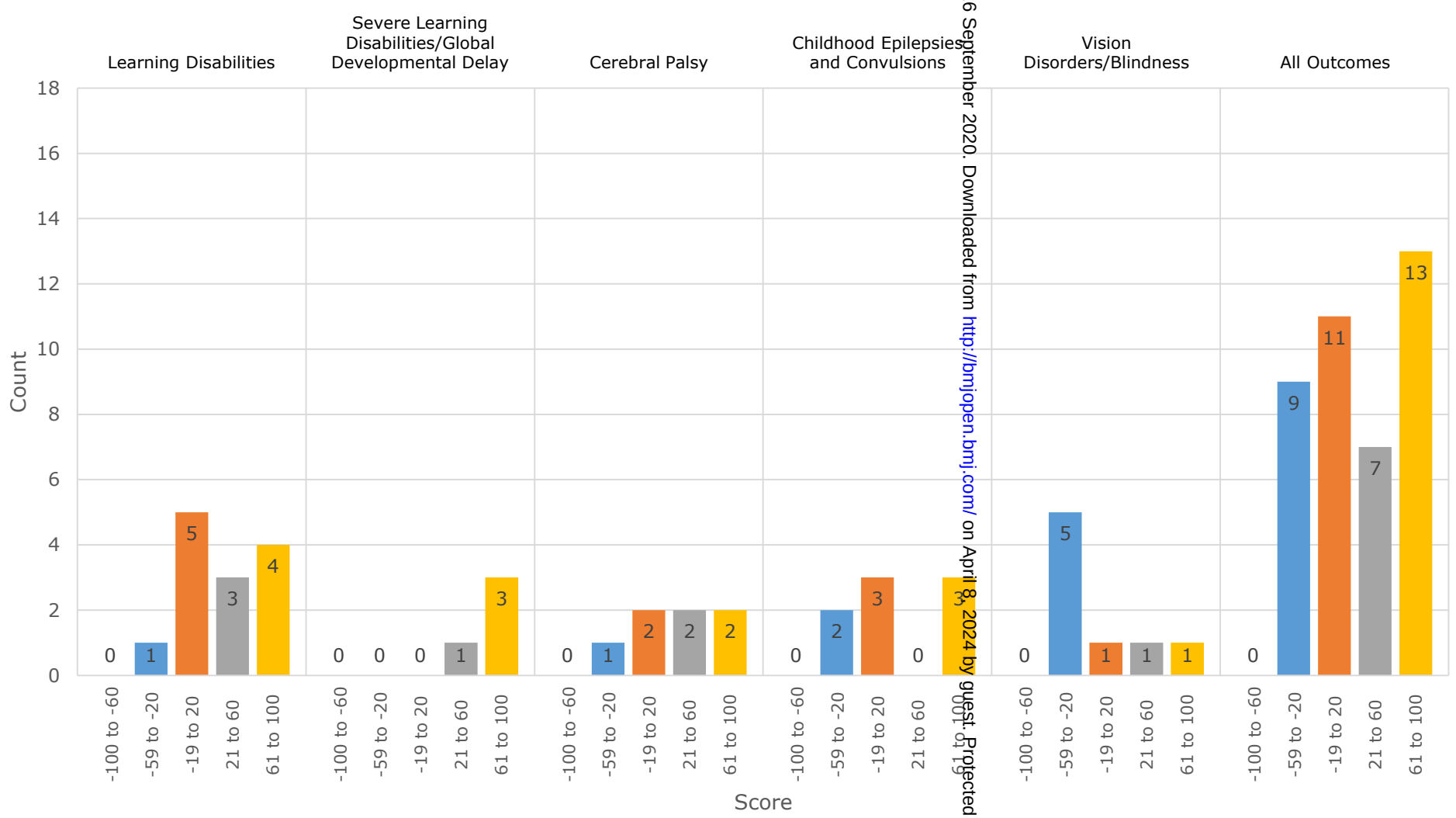
Funding statement

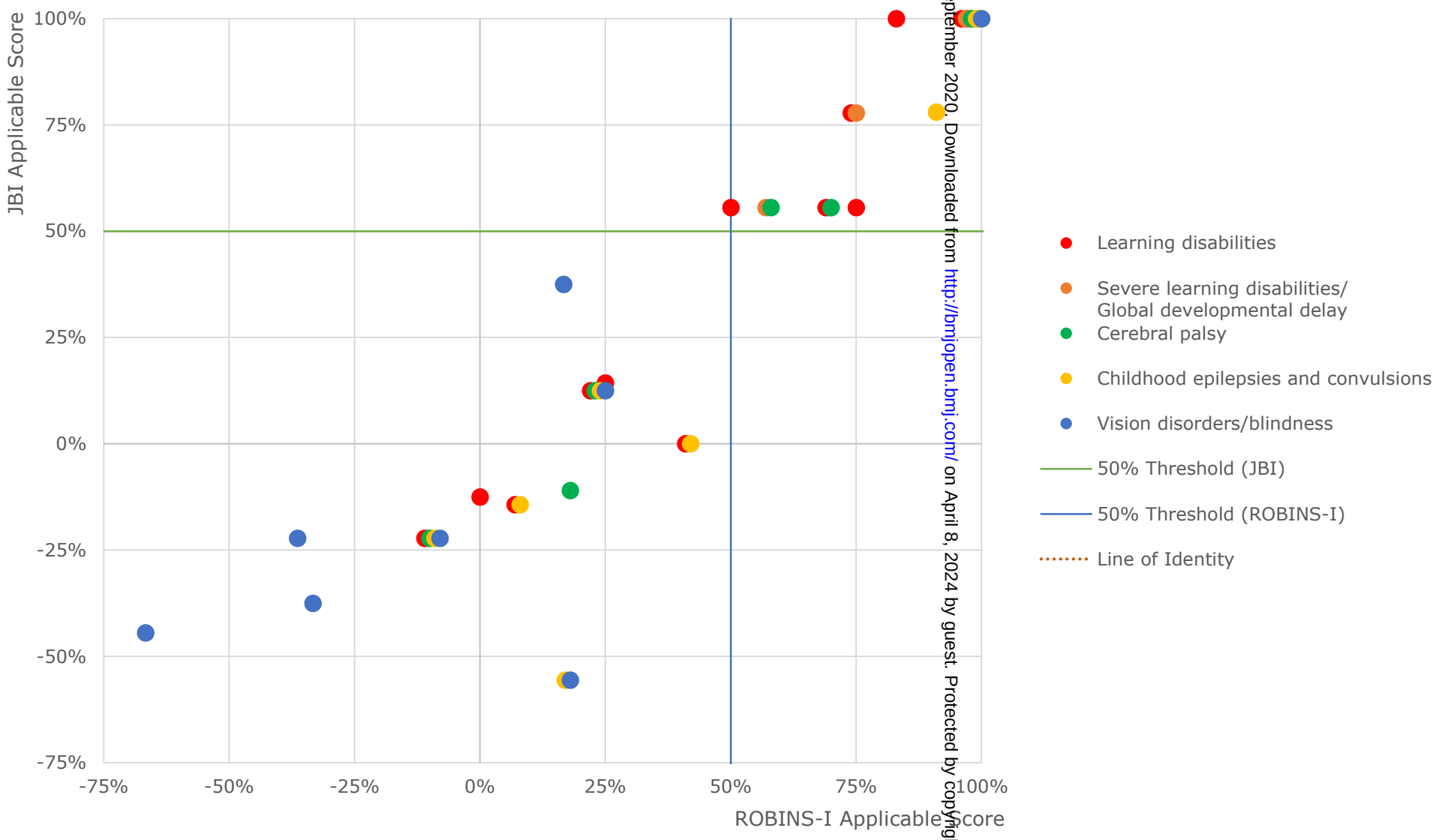
This work was supported in part by the Auckland Medical Research Foundation, grant number 1417003.

Competing interests statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.







Note: overlapping data points have been spread to allow for visibility. The highest data point in an overlapping cluster is the actual value.

Assessment domain	Modified ROBINS-I tool	JBI tool
Selection of comparison groups	Bias in selection of comparison groups (recruitment timing and populations)	Was the sample frame appropriate to address the target population? Were study participants sampled in an appropriate way? Was the sample size adequate?
Confounding	Bias due to confounding (balance of cointerventions; adjustments for confounding)	
Ascertainment of exposures	Bias in ascertainment of exposures (clearly defined, valid, reliable and consistent implementation of exposure measurement)	Were valid methods used for the identification of the condition? Was the condition measured in a standard, reliable way for all participants?
Measurement of outcomes	Bias in measurement of outcomes (Were the outcome assessors blinded to the exposure status of participants? Were the methods of outcome assessment comparable across exposure groups?)	
Missing data	Bias due to missing data (Were outcome data available for all or nearly all participants? If no, was the proportion of participants and reasons for missing data similar across groups?)	Was the response rate adequate, and if not, was the low response rate managed appropriately? Was the data analysis conducted with sufficient coverage of the identified sample?
Reporting of results	Bias in reporting of results (Was analysis of outcomes adequately prespecified? Were there multiple analyses of the exposure-outcome relationships?)	
Design	*Was the research question or objective clearly stated? *Was a sample size justification, power description, or variance and effect estimates provided? *Was there a relevant comparison (two or more groups, or same participants over time)?	Was there appropriate statistical analysis?
Applicability	*Was the study population clearly specified and defined? *Does the study cohort adequately represent the target population (cohort studies)? *Was the participation rate of eligible participants at least 50%?	Were the study subjects and the setting described in detail?
*added to the original ROBINS-I tool		

Supplementary Table 1: Comparison of domains included in each assessment tool.

Score calculation method	Low risk of bias	High risk of bias	Uncertain	Not applicable
Item 1			x	
Item 2	x			
Item 3	x			
Item 4			x	
Item 5	x			
Item 6	x			
Item 7				x
Item 8		x		
Item 9		x		
Item 10			x	
Item 11	x			
Item 12	x			
TOTALS	6	2	3	1
1. Count Scores	Positive: 6/12 = 50% Negative: 2/12 = 17%			
2. Composite Score	(6-2)/12 = 33%			
3. Applicable Score	(6-2)/(12-1) = 36%			

Supplementary Table 2: Example of scoring system calculations