STUDY PROTOCOL

REVISED Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol [version 3; peer review: 2 approved, 1 approved with reservations]

Elizabeth Korevaar[1], Amalia Karahalios[1], Andrew B. Forbes[1], Simon L. Turner[1], Steve McDonald[1], Monica Taljaard[2,3], Jeremy M. Grimshaw[2-4], Allen C. Cheng[1,5], Lisa Bero[6], Joanne E. McKenzie[1]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, 3004, Australia
[2]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, K1Y 4E9, Canada
[3]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada
[4]Department of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada
[5]Infection Prevention and Healthcare Epidemiology Unit, Alfred Health, Melbourne, Victoria, 3004, Australia
[6]Faculty of Medicine and Health and Charles Perkins Centre, University of Sydney, Sydney, New South Wales, 2006, Australia

v3 **First published:** 12 Feb 2020, **9**:110
https://doi.org/10.12688/f1000research.22226.1
Second version: 23 Jul 2020, **9**:110
https://doi.org/10.12688/f1000research.22226.2
Latest published: 11 Dec 2020, **9**:110
https://doi.org/10.12688/f1000research.22226.3

Abstract

Background: Systematic reviews are used to inform healthcare decision making. In reviews that aim to examine the effects of organisational, policy change or public health interventions, or exposures, evidence from interrupted time series (ITS) studies may be included. A core component of many systematic reviews is meta-analysis, which is the statistical synthesis of results across studies. There is currently a lack of guidance informing the choice of meta-analysis methods for combining results from ITS studies, and there have been no studies examining the meta-analysis methods used in practice. This study therefore aims to describe current meta-analysis methods used in a cohort of reviews of ITS studies.

Methods: We will identify the 100 most recent reviews (published between 1 January 2000 and 11 October 2019) that include meta-analyses of ITS studies from a search of eight electronic databases covering several disciplines (public health, psychology, education, economics). Study selection will be undertaken independently by two authors. Data extraction will be undertaken by one author, and for a random sample of the reviews, two authors. From eligible reviews we will extract details at the review level including discipline, type of interruption and any tools used to assess the risk of bias /

Open Peer Review

Reviewer Status ? ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
version 3 (revision) 11 Dec 2020			
version 2 (revision) 23 Jul 2020			✓ report
version 1 12 Feb 2020	? report	✓ report	

1. **Christopher James Rose**, Norwegian Institute of Public Health, Oslo, Norway

2. **Alexandra McAleenan**, University of Bristol, Bristol, UK

methodological quality of included ITS studies; at the meta-analytic level we will extract type of outcome, effect measure(s), meta-analytic methods, and any methods used to re-analyse the individual ITS studies. Descriptive statistics will be used to summarise the data.
Conclusions: This review will describe the methods used to meta-analyse results from ITS studies. Results from this review will inform future methods research examining how different meta-analysis methods perform, and ultimately, the development of guidance.

Keywords
Interrupted time series, meta-analysis, systematic review

3. **Tim Mathes**, Witten/Herdecke University, Cologne, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Joanne E. McKenzie (joanne.mckenzie@monash.edu)

Author roles: Korevaar E: Funding Acquisition, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Karahalios A**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Forbes AB**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Turner SL**: Methodology, Writing – Review & Editing; **McDonald S**: Methodology, Writing – Review & Editing; **Taljaard M**: Funding Acquisition, Methodology, Writing – Review & Editing; **Grimshaw JM**: Funding Acquisition, Writing – Review & Editing; **Cheng AC**: Funding Acquisition, Writing – Review & Editing; **Bero L**: Funding Acquisition, Writing – Review & Editing; **McKenzie JE**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: EK and SLT are supported through an Australian Government Research Training Program (RTP) Scholarship administered by Monash University, Australia. JEM is supported by an NHMRC Career Development Fellowship (1143429). The project is funded by the Australian National Health and Medical Research Council (NHMRC) project grant GNT1145273, "How should we analyse, synthesize, and interpret evidence from interrupted time series studies? Making the best use of available evidence", McKenzie JE, Forbes A, Taljaard M, Cheng A, Grimshaw J, Bero L, Karahalios A.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Korevaar E, Karahalios A, Forbes AB *et al.* **Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol [version 3; peer review: 2 approved, 1 approved with reservations]** F1000Research 2020, **9**:110 https://doi.org/10.12688/f1000research.22226.3

First published: 12 Feb 2020, **9**:110 https://doi.org/10.12688/f1000research.22226.1

REVISED **Amendments from Version 2**

This version of the review protocol contains a minor clarification of the inclusion criteria.

Any further responses from the reviewers can be found at the end of the article

Introduction

Systematic reviews aim to collate and synthesise all available evidence on a particular topic. They are used to inform healthcare decision making, either directly, or through their inclusion in knowledge tools such as clinical practice guidelines. Many reviews examining the effects of clinical interventions are appropriately limited in scope to inclusion of randomised trials. However, in reviews where the aim is to examine the effects of organisational, policy change or public health interventions or exposures (e.g. chemical exposures), evidence from non-randomised studies may offer the only evidence, or provide important additional evidence to that gained from randomised trials[1].

Interrupted time series (ITS) studies are a type of non-randomised design in which measurements on a group of individuals (e.g. a community) are taken repeatedly both before and after an 'interruption'[2]. The interruption may be intended (e.g. a government-implemented policy), although will not necessarily be initiated or designed by the ITS investigators (e.g. by researchers within a university)[3], or may be unintended (e.g. an exposure such as a natural disaster). The key benefit of the ITS design is that the period before the interruption can be used to estimate the underlying time trend. If modelled correctly, this before-trend can be projected into the post-interruption period to provide a counterfactual for what would have occurred in the absence of the interruption[4]. ITS studies with controls (e.g. an internal or external control series, control outcome) may provide more certainty in causally attributing any observed effects to the interruption[5]. Several effect estimates can be obtained from an ITS study to characterise both short and long-terms effects of the interruption (e.g. level change and slope change).

Meta-analysis is the statistical synthesis of results across studies leading to combined effect estimates[6]. Meta-analysis (and its extensions) is a core component of many systematic reviews. The benefits of meta-analysis have long been established, including the ability to more precisely estimate effects, examine and quantify inconsistency of the effects across studies, and identify factors that may potentially modify the size of the effects[7–10]. Two approaches for meta-analysing results from ITS studies include the two-stage or one-stage approach[11]. In the two-stage approach, effect estimates from each series are first computed, and these are then combined across series using a meta-analysis method (e.g. DerSimonian and Laird[12]). In the one-stage approach, a single model including all series is fitted to simultaneously obtain the combined effect estimates[11]. The one-stage approach requires the raw time series data to be available for all series, but has the proposed advantage of being more efficient since the data across all the series are used in estimating the effects[11].

For two-stage meta-analysis, a notable challenge is that many primary ITS studies are analysed incorrectly. For example, ITS studies may be analysed as though the study was a before-after design[13], or analysed as an ITS design, but without taking account of the correlation between observations over time (known as autocorrelation)[14–16]. The former is likely to result in estimates of the effect of the interruption that are biased, while the latter is likely to result in estimates of standard errors that are too small[17]. Both have important implications for a two-stage meta-analysis in terms of bias, the weights that studies receive, and in turn, the precision of the combined estimate. A further challenge is that the effect measures chosen and reported by the primary study authors (e.g. level change) may not match those of interest to the systematic reviewer (e.g. slope change).

In some studies, the raw time series data may be available through extraction of data from graphs or their availability in tables[6,8,11,18]. In this circumstance, it may be possible to overcome some of the above challenges through re-analysis of the ITS studies by appropriately accounting for the design and autocorrelation, or re-analysing the raw data to obtain the desired effect measure for the meta-analysis when it differs from that reported in the primary study. These computed effects may then be combined using two-stage meta-analysis. Alternatively, each study's raw data may be analysed in a single model using a one-stage meta-analysis approach[6,8].

To our knowledge, there have been no reviews examining the approaches and methods used to meta-analyse effect estimates from ITS studies. In this review we therefore aim to: 1) investigate whether reviewers re-analyse primary ITS studies included in reviews, and if so, what re-analysis methods are used; 2) what meta-analysis methods are used; 3) what effect measures are used, and how completely the estimated combined effects are reported; and 4) what tools and domains are used to assess the risks of bias or methodological quality of the included ITS studies. Here, we report the planned design of our review, including the criteria that we will use to identify eligible studies, as well as the information we will extract and describe.

Methods

Overview

This study aims to identify and describe reviews that include meta-analyses of ITS studies. The reviews will be identified by searching several electronic databases including MEDLINE (Ovid), EMBASE (Ovid), Campbell Systematic Reviews, EconLit (EBSCOhost), 3ie, PsycINFO (Ovid), ERIC (ProQuest) and the Cochrane Database of Systematic Reviews (CDSR). Study selection will be undertaken independently by two authors; data extraction will be undertaken by one author, and for a minimum of 20% of randomly selected reviews, two authors. We will extract details at the systematic review level, including: discipline (public health, psychology, education, economics), type of interruption, assessment of risk of bias and methodological

quality; and at the meta-analytic level: type of outcome, effect measure(s), meta-analytic methods, and any methods used to re-analyse the individual ITS studies. These aspects will be analysed and described using summary statistics, tables and figures.

Eligibility criteria

Studies which meet our eligibility criteria (described below) will be included. We will not restrict inclusion of reviews based on discipline or any of the PICO elements (i.e. participants/populations, interventions/interruptions, comparators, or outcomes).

Inclusion criteria. Studies meeting the following criteria will be included:

1. the study is a review that includes at least two ITS studies which meet the review authors' definition of an ITS design; and
2. the review includes at least one meta-analysis of ITS studies.

Our definition of a 'review' is very broad. It includes systematic reviews, reviews of selected studies (i.e. between-study meta-analysis), and studies that combine multiple ITS across sites within the same study (i.e. within-study meta-analysis). We have opted for broad inclusion since our primary interest is in the meta-analysis methods, which apply regardless of the particular study design. We will not restrict the meta-analysis by approach, that is, we will include both one-stage and two-stage meta-analyses. We will only include meta-analyses that provide pooled estimates of model parameters that quantify the effect of the interruption.

Exclusion criteria. Studies will be excluded if they meet one or more of the following criteria. The study is:

1. written in a language other than English;
2. a methodological review that describes or evaluates methods to synthesise results from ITS studies;
3. a review of ITS studies reported in a conference abstract, letter, book, or dissertation;
4. a protocol for a review of ITS studies; or
5. a stepped-wedge randomised trial.

Criterion 1 is included because we are not able to translate studies written in a language other than English due to resource constraints. Criterion 2 excludes methodological reviews that describe or evaluate methods to synthesise data from ITS studies, as our aim is to describe current statistical methods applied in practice.

Search methods

Several databases will be searched to capture the broad range of disciplines that use the ITS study design. To capture reviews in health, we will search MEDLINE (Ovid), EMBASE (Ovid), Campbell Systematic Reviews, the CDSR and 3ie. For CDSR, we will directly search the 'Characteristics of included studies' table included in each systematic review for ITS studies. This will allow more specific identification of eligible reviews. The search of MEDLINE (Ovid) will also capture systematic reviews from the Joanna Briggs Institute Database of Systematic Reviews and Implementation Reports. To capture reviews in economics, we will search EconLit (EBSCOhost), and for psychology and education disciplines, we will search PsycINFO (Ovid) and ERIC (ProQuest).

Our search strategy has been informed by previous publications that have reviewed ITS studies[14,15,19]. Reviews of ITS studies will be identified using terms adapted from the search strategies of these publications and then combined with terms to identify meta-analyses and systematic reviews. As there is little consistency in the terminology used to describe ITS studies[15,16], our search terms are intentionally broad to achieve greater search sensitivity. Terms will be searched both as free text in the titles, abstracts and keywords fields, and as MeSH terms (or equivalent) where applicable. The MEDLINE (Ovid) strategy is presented in Table 1, and the search strategies for the remaining databases are presented in Appendix 1 (see *Extended data*)[20]. The search is limited to the period 1 Jan 2000 to 11 Oct 2019 for all databases except CDSR which is limited to the period 1 Jan 2000 to 9 Aug 2019.

Study selection

Citations identified from the searches will be imported into Endnote X8 (Clarivate Analytics, Philadelphia) to remove duplicates. Titles and abstracts will be sorted by year in descending order and will be screened against the eligibility criteria, with each abstract assessed as: 1) 'Yes/Maybe includes two or more ITS studies' and 'Yes/Maybe a meta-analysis of ITS studies has been undertaken', 2) 'Yes/Maybe includes two or more ITS studies' and 'No meta-analysis of ITS studies has been undertaken', or 3) 'No, does not include two or more ITS studies'. This process will be piloted on 20 studies by EK, SLT, AK and JEM. The remaining abstracts will be screened independently by at least two members of the review team (EK, and any of SLT, AK and JEM). The full-text articles of the titles and abstracts assessed as potentially meeting the eligibility criteria (i.e. group 1 above) will be retrieved, sorted by most recent first and screened against the eligibility criteria until all reviews (if less than 100), or the 100 most recently published reviews are identified. Conflicts in screening decisions at the abstract and full-text stages will be resolved via discussion between the screeners or through consultation with the broader team.

Sample size

Our sample size of 100 reviews was primarily selected for reasons of feasibility. A sample of this size will allow estimation of the percentage of reviews with a particular element (e.g. the prevalence of the reviews that re-analyse the primary study data) to within a maximum margin of error of 10% (assuming a prevalence of 50%). This margin of error represents the worst-case scenario and will decrease if the prevalence varies from 50%.

Table 1. MEDLINE (Ovid) search strategy.

#	Searches
1	Interrupted Time Series Analysis/
2	interrupted time series.mp.
3	(time series or time trend$ or trend analys?s).mp.
4	(change point or repeated measures or phase design or multiple baseline$ or difference-in-difference$ or single case research or single case experimental).mp.
5	(ARIMA or autoregressive integrated moving average or integrated moving average or piecewise regression or segmented regression).mp.
6	or/1–5
7	(Systematic Review or Meta-Analysis).pt. or (meta-analys?s or pooled analys?s).ti,ab,kw.
8	and/6–7
9	Limit 8 to (abstracts and English language and yr="2000-Current")

Selection of outcomes

Reviews may include several meta-analyses of ITS studies for *different* outcomes. We plan to examine the meta-analysis methods for only *one* outcome per review. The following set of rules will be applied hierarchically until a unique outcome is identified (for which there could be multiple meta-analyses of different effect estimates):

1) The outcome that has the largest number of effect measures (e.g. the outcome that has meta-analyses of level change and slope change estimates would be selected ahead of an outcome with only a meta-analysis of level change estimates);

2) The outcome with the largest number of ITS studies; or

3) The outcome that is first reported in the abstract, then the methods section, then the results section of the manuscript.

A single outcome is chosen as it is likely that the meta-analysis methods are consistent across outcomes within a review. Criterion 1 has been included so that we can capture the range of effect measures used. Uncertainty in the selection of the outcome will be resolved through discussion with the review team.

Data extraction and management

The data extraction form will be designed using the Research Electronic Data Capture (REDCap) online designer[21,22]. The review team (EK, AK, SLT, ABF, and JEM) will pilot the data extraction form by independently extracting data from 10 reviews. This pilot testing will be used to revise the form if we uncover ambiguity or a lack of clarity in any items, identify missing items and test the logic of the form. Following piloting, data extraction will be undertaken by EK for all eligible studies and independently by at least two members of the review team (one of AK, SLT, ABF and JEM) for a further 20% of randomly selected reviews. Any inconsistencies in data extraction will be resolved via discussion between the data extractors or through consultation with the broader team. For any items where a large percentage of inconsistency is found, the percentage of studies with double data extraction will be increased.

A summary of the data extraction items is presented in Table 2, while version 1 of the data extraction form can be viewed in Appendix 2 (see *Extended data*)[23]. In brief, we will extract details of the review's aims, meta-analysis methods (including the reviewer's rationale) and methods used to assess the methodological quality and/or risk of bias of the included ITS studies. For the selected outcome, we will extract the type of effect measure(s), methods of synthesis, adjustment for autocorrelation and/or seasonality.

Analysis

We will summarise the characteristics of included systematic reviews with descriptive statistics. For categorical data (e.g. the meta-analysis approach used, the risk of bias tool used) we will present frequencies (with percentages), and for numerical data (e.g. the number of meta-analysed ITS studies, the number of pooled estimates) we will present means (with standard deviations) or medians (with interquartile range). Statistical analyses will be undertaken using Stata version 15.0[24].

Table 2. Summary of data extraction items for the selected meta-analysis.

Domain	Items
Review characteristics	Author name; publication year; journal; target population or content area; number of ITS studies included in the review; reviewers' definition of ITS studies
Outcome and studies included	Description and type of outcome (e.g. continuous, count, rate); number of ITS studies measuring the chosen outcome
Methods for synthesising ITS results	Number of ITS studies meta-analysed; use of primary study data (i.e. re-analysis); pairwise or network meta-analysis; fixed/random effects model; methods to quantify between-study variation
Results/Estimates	Description and type of effect measures (e.g. change in level, change in slope, combination of change in level and slope (i.e., counterfactual)); completeness of reporting estimates (e.g. combined effect estimate, confidence interval, measure of heterogeneity)
Risk of bias and/or assessment of study quality	Description of assessment (if performed) of primary study risk of bias / methodological quality; tool or domains used for assessment

Abbreviations: ITS, interrupted time series

Discussion

To our knowledge, this will be the first review to examine methods for meta-analysis of ITS studies that are used in practice. Specifically, the choice of meta-analysis approach, effect measures, completeness of reporting, and tools for assessing quality or risk of bias (if undertaken). The results of this review will inform our broader research program which aims to examine how different meta-analysis methods of ITS studies perform, and provide guidance on the methods. Specifically, the review will identify the range of statistical methods that are used in practice, characteristics of the times series studies included in the meta-analyses (e.g. number of series, length of series) and the types of effect measures used (e.g. level change, slope change). These characteristics will inform a statistical simulation study that will examine the performance of different methods for meta-analysis of ITS studies. In addition, we will identify reviews for which the raw time series data of the included ITS studies are reported. These reviews will be used in an empirical evaluation to examine the impact of using different meta-analysis methods of real ITS studies.

Strengths and limitations

There are several strengths to this study. The search, screening and data extraction methods have been prespecified, and the study has been registered with PROSPERO (accepted 28 Apr 2020 submitted 4 Oct 2019). Further, we will search a broad range of databases, encompassing the areas of health, economics, psychology and education. This will allow us to identify a broader range of meta-analysis methods in use, not restricted to a particular discipline.

While the study will be limited by our ability to identify all potentially eligible reviews and meta-analyses of ITS studies, our search strategy attempts to capture the various ways these studies are described. However, given ITS studies are often not identified as such[16] and that our search is restricted to articles written in English, it is likely that we will not capture all reviews and meta-analyses that include ITS studies. Conversely, we may end up including reviews where no information regarding the definition of the included ITS studies is provided, or where an inappropriate label of ITS has been applied to included studies. While we will not exclude these reviews, we will record the reviewers' definition of an ITS study.

Conclusions

The ITS design is often used to examine the effects of organisational, policy change or public health interventions or exposures. Meta-analysis of results from these studies provides the opportunity to estimate the interruption's impact more precisely, and investigate factors that may modify the size of the impact. However, there is a paucity of guidance available for meta-analysing results from ITS studies. Results from this review will provide the first examination of meta-analysis methods used in practice to combine results from ITS studies. This will be used to inform future research that investigates how different methods perform, from which guidance will be developed.

Data availability

Underlying data

No underlying data are associated with this article.

Extended data

Figshare (Monash University repository, known as Bridges): Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol - Appendix 1 Search strategy. https://doi.org/10.26180/5e3b5cc4acf30[20].

Figshare (Monash University repository, known as Bridges): Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol – Appendix 2 Data extraction form (version 1). https://doi.org/10.26180/5f0682e008dda[23]

Reporting guidelines

Figshare (Monash University repository, known as Bridges): PRISMA-P checklist for 'Methods used to meta-analyse results from interrupted time series studies: A meth odological systematic review protocol'. https://doi.org/10.26180/5e3b5d75c00a6[25].

References

1. Higgins JP, Ramsay C, Reeves BC, *et al.*: **Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions.** *Res Synth Methods.* 2013; **4**(1): 12–25.
PubMed Abstract | Publisher Full Text
2. Wagner AK, Soumerai SB, Zhang F, *et al.*: **Segmented regression analysis of interrupted time series studies in medication use research.** *J Clin Pharm Ther.* 2002; **27**(4): 299–309.
PubMed Abstract | Publisher Full Text
3. Watson SI, Dixon-Woods M, Taylor CA, *et al.*: **Revising ethical guidance for the evaluation of programmes and interventions not initiated by researchers.** *J Med Ethics.* 2020; **46**(1): 26–30.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Shadish WR, Cook TD, Campbell DT: **Experimental and quasi-experimental designs for generalized causal inference**. 2002.
Reference Source
5. Bernal JL, Cummins S, Gasparrini A: **The use of controls in interrupted time series studies of public health interventions.** *Int J Epidemiol.* 2018; **47**(6): 2082–93.
PubMed Abstract | Publisher Full Text
6. McKenzie JE, Beller EM, Forbes AB: **Introduction to systematic reviews and meta-analysis.** *Respirology.* 2016; **21**(4): 626–37.
PubMed Abstract | Publisher Full Text
7. Deeks JJ: **Systematic reviews of published evidence: miracles or minefields?** *Ann Oncol.* 1998; **9**(7): 703–9.
PubMed Abstract | Publisher Full Text
8. Deeks J, Higgins J, Altman D: **Chapter 10: Analysing data and undertaking meta-analyses.** In: *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane. 2019.
Publisher Full Text
9. Caldwell DM, Dias S, Welton NJ: **Extending Treatment Networks in Health Technology Assessment: How Far Should We Go?** *Value Health.* 2015; **18**(5): 673–81.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Thompson SG, Higgins JP: **How should meta-regression analyses be undertaken and interpreted?** *Stat Med.* 2002; **21**(11): 1559–73.
PubMed Abstract | Publisher Full Text
11. Gebski V, Ellingson K, Edwards J, *et al.*: **Modelling interrupted time series to evaluate prevention and control of infection in healthcare.** *Epidemiol Infect.* 2012; **140**(12): 2131–41.
PubMed Abstract | Publisher Full Text
12. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials.* 1986; **7**(3): 177–88.
PubMed Abstract | Publisher Full Text
13. Ramsay CR, Matowe L, Grilli R, *et al.*: **Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies.** *Int J Technol Assess Health Care.* 2003; **19**(4): 613–23.
PubMed Abstract | Publisher Full Text
14. Jandoc R, Burden AM, Mamdani M, *et al.*: **Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations.** *J Clin Epidemiol.* 2015; **68**(8): 950–6.
PubMed Abstract | Publisher Full Text
15. Hudson J, Fielding S, Ramsay CR: **Methodology and reporting characteristics of studies using interrupted time series design in healthcare.** *BMC Med Res Methodol.* 2019; **19**(1): 137.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Polus S, Pieper D, Burns J, *et al.*: **Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews.** *J Clin Epidemiol.* 2017; **91**: 56–69.
PubMed Abstract | Publisher Full Text
17. Huitema BE, McKean JW: **Identifying autocorrelation generated by various error processes in interrupted time-series regression designs - A comparison of AR1 and portmanteau tests.** *Educ Psychol Meas.* 2007; **67**(3): 447–59.
Publisher Full Text
18. Turner SL, Karahalios A, Forbes AB, *et al.*: **Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review.** *J Clin Epidemiol.* 2020; **122**: 1–11.
PubMed Abstract | Publisher Full Text
19. Turner SL, Karahalios A, Forbes AB, *et al.*: **Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: protocol for a review.** *BMJ Open.* 2019; **9**(1): e024096.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Korevaar E, Karahalios E, Forbes A, *et al.*: **Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol - Appendix 1 Search strategy.** *figshare.* Journal contribution. 2020.
http://www.doi.org/10.26180/5e3b5cc4acf30
21. Harris PA, Taylor R, Minor BL, *et al.*: **The REDCap consortium: Building an international community of software platform partners.** *J Biomed Inform.* 2019; **95**: 103208.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Harris PA, Taylor R, Thielke R, *et al.*: **Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support.** *J Biomed Inform.* 2009; **42**(2): 377–81.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Korevaar E, Karahalios E, Forbes A, *et al.*: **Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol – Appendix 2 Data extraction form (version 1).** *figshare.* Journal contribution. 2020.
http://www.doi.org/10.26180/5f0682e008dda
24. StataCorp: **Stata statistical software: release 15.** Tx: College Station StataCorp LLC; 2017.
25. Korevaar E, Karahalios E, Forbes A, *et al.*: **Methods used to meta-analyse results from interrupted time series studies: A methodological systematic review protocol - Completed PRISMA-P checklist.** *figshare.* Journal contribution. 2020.
http://www.doi.org/10.26180/5e3b5d75c00a6

Open Peer Review

Current Peer Review Status: ? ✓ ✓

Version 2

Reviewer Report 02 November 2020

https://doi.org/10.5256/f1000research.27594.r73704

✓ **Tim Mathes**
Institute for Research in Operative Medicine, Faculty of Health, School of Medicine, Witten/Herdecke University, Cologne, Germany

This is a very interesting and well-written study protocol. I have only some minor suggestions for consideration.

- The search strategy assumes that the very specific study type/design label ITS is reported in the abstract of the meta-analysis. I recognized that this problem is already addressed as limitation. However, I suppose that an additional related problem would be that the search will result in an imbalance towards EPOC-Cochrane because in my experience this usually mention ITS in the abstract in contrast to other non-EPOC SR. It might be worth to adapted the sampling strategy accordingly, e.g. screening until 50 Cochrane and 50 non-Cochrane reviews.

- The search includes the term “ARIMA” but no other terms that specify possible analysis of ITS (e.g. segmented regression) and consequently would probably be biased towards ITS that are analysed using ARIMA approaches.

- “We will not restrict the meta-analysis by approach, that is, we will include both one-stage and two-stage meta-analyses. We will only include meta-analyses that combine estimates of model parameters, or combinations of these (e.g. pre-intervention fitted trend, slope change, level change)”. This statement is contradictory. Pooled estimates of one-stage models are based on pseudo IPD that are pooled in one-step. Thus, there is no model parameter for each study and consequently it will not give model parameters, that can be combined (second step in two-stage models).

- “the study is a review that includes at least two ITS studies which meet the review authors’ definition of an ITS design”. Does it mean that you will also include SR which include ITS with a control group? If yes, I think the meta-analysis methods must be analysed separately.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?
Yes

Are sufficient details of the methods provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: medical statistics, research synthesis methods, non-randomized studies

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 08 Dec 2020

Elizabeth Korevaar, Monash University, Melbourne, Australia

We would like to thank Dr Mathes for their feedback on our protocol and the suggestions for its improvement. Below, we have addressed each of the four items raised.

Minor comments:

This is a very interesting and well-written study protocol. I have only some minor suggestions for consideration.

1. The search strategy assumes that the very specific study type/design label ITS is reported in the abstract of the meta-analysis. I recognized that this problem is already addressed as limitation. However, I suppose that an additional related problem would be that the search will result in an imbalance towards EPOC-Cochrane because in my experience this usually mention ITS in the abstract in contrast to other non-EPOC SR. It might be worth to adapted the sampling strategy accordingly, e.g. screening until 50 Cochrane and 50 non-Cochrane reviews.

Thank you for your interest in our protocol and your suggestions. As noted, we acknowledged that study labels used by the systematic reviewers are a limitation of our search strategy. We attempted to circumvent this limitation by including broader, related terms (e.g. repeated measures, difference-in-difference) in our search strategy. We appreciate the suggestion of adapting the sampling strategy to increase the proportion of non-EPOC reviews. However, as the search and screening has been completed and has identified a total of 54 eligible reviews, we have extracted data from all eligible reviews.

2. The search includes the term “ARIMA” but no other terms that specify possible analysis of ITS (e.g. segmented regression) and consequently would probably be biased towards ITS that are analysed using ARIMA approaches.

In line 5 of our search strategy (Table 1), as noted by the reviewer, we have included the terms *ARIMA*, *autoregressive integrated moving average*, and *integrated moving average*. Additionally, this line in our search strategy includes the terms *piecewise regression* and *segmented regression*.

3. “We will not restrict the meta-analysis by approach, that is, we will include both one-stage and two-stage meta-analyses. We will only include meta-analyses that combine estimates of model parameters, or combinations of these (e.g. pre-intervention fitted trend, slope change, level change)”. This statement is contradictory. Pooled estimates of one-stage models are based on pseudo IPD that are pooled in one-step. Thus, there is no model parameter for each study and consequently it will not give model parameters, that can be combined (second step in two-stage models).

To address the ambiguity in our target effect measures, we have amended the wording of the highlighted text to “*We will not restrict the meta-analysis by approach, that is, we will include both one-stage and two-stage meta-analyses. We will only include meta-analyses that provide pooled estimates of model parameters that quantify the effect of the interruption*”. The pooled estimates of model parameters that we are interested in are those that quantify the effect of the interruption (whether that be parameterised as a level change, slope change, etc). We are not interested in the effects of other covariates either in a one-stage model or in the first stage of a two-stage approach.

4. “the study is a review that includes at least two ITS studies which meet the review authors’ definition of an ITS design”. Does it mean that you will also include SR which include ITS with a control group? If yes, I think the meta-analysis methods must be analysed separately.

Yes, we will include systematic reviews that include a meta-analysis of two or more ITS studies. The studies could all have an interruption (i.e. uncontrolled ITS) or at least one of the series might be a control series (i.e. controlled ITS). Controlled time series can be included by using a unified model (i.e., a one-stage meta-analysis, or the first step of a two-stage meta-analysis), a statistical comparison (i.e., separate analyses of the interrupted series and control series, with a statistical comparison of the two) or a narrative comparison. Our data collection tool will capture information on which of the three above scenarios applies to the meta-analysis, how many are controlled versus uncontrolled, and how the controlled studies were incorporated into the review.

Kind regards,
Elizabeth Korevaar

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 20 May 2020

https://doi.org/10.5256/f1000research.24513.r62669

Alexandra McAleenan
Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

This article is a protocol for a systematic review that aims to describe the meta-analysis methods (and some of the systematic review methods) used in systematic reviews that include interrupted time series (ITS) studies.

The protocol is thorough and well written. A comprehensive search of the literature will be undertaken to identify reviews of health, public health, psychology, education and economic interventions. There are clear eligibility criteria for reviews, and the data to be extracted is well described, as is the planned analysis.

I have a couple of minor suggestions. In the abstract it would be preferable to state how the 100 reviews will be selected (for e.g. "We will identify up to the 100 most recently published reviews"). You could also add in that you will be looking at the tools used to assess bias/methodological quality of ITS studies (which one could argue is more looking at systematic review methods that meta-analysis methods).

You could also clarify, by adding to the summary of data extraction items, that you will be looking at the review authors' definition of ITS studies. I presume you will also be extracting whether any of the ITS studies included a control group, and perhaps how the control group was used (especially if any re-analysis was undertaken)?

Is the rationale for, and objectives of, the study clearly described?
Yes

Is the study design appropriate for the research question?
Yes

Are sufficient details of the methods provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Systematic reviewing.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Jul 2020

Elizabeth Korevaar, Monash University, Melbourne, Australia

We would like to thank Dr McAleenan for their feedback on our protocol and the suggestions for its improvement. Below, we have addressed each of the suggestions raised.

Minor suggestions:

1. In the abstract it would be preferable to state how the 100 reviews will be selected (for e.g. "We will identify up to the 100 most recently published reviews"). You could also add in that you will be looking at the tools used to assess bias/methodological quality of ITS studies (which one could argue is more looking at systematic review methods that meta-analysis methods).

We have revised the 'Abstract' to make the selection process clearer:
"We will identify *the* 100 *most recent* reviews *(published between 1 January 2000 and 11 October 2019)* that include meta-analyses of ITS studies from a search of eight electronic databases covering several disciplines (public health, psychology, education, economics)."

In addition, we have added text to note we will be extracting information about the tools used to assess risk of bias/methodological quality:
"From eligible reviews we will extract details at the review level including discipline, type of interruption *and any tools used to assess the risk of bias / methodological quality of included ITS studies*; at the meta-analytic level we will extract type of outcome, effect measure(s), meta-analytic methods, and any methods used to re-analyse the individual ITS studies. Descriptive statistics will be used to summarise the data."

1. You could also clarify, by adding to the summary of data extraction items, that you will be looking at the review authors' definition of ITS studies. I presume you will also be extracting whether any of the ITS studies included a control group, and perhaps how the control group was used (especially if any re-analysis was undertaken)?

We have added "reviewers' definition of ITS studies" to the summary of data extraction items table (Table 2). In addition, we have included the current version of our data extraction form as an additional file in the 'Extended data' section. We will collect data on whether and how control series were incorporated into the meta-analysis (see rows 69-71 of the data extraction form, Appendix 2).

Kind regards,
Elizabeth Korevaar

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 February 2020

https://doi.org/10.5256/f1000research.24513.r60103

?

Christopher James Rose
Norwegian Institute of Public Health, Oslo, Norway

In this protocol, the authors propose to systematically survey methods used to meta-analyze results from interrupted time series (ITS) studies. ITS studies are particularly useful in systematic reviews (SRs) of interventions that cannot be studied using randomized trials (e.g., due to practical, ethical, or legal reasons). Briefly, the protocol plans to identify 100 recent systematic reviews that meta-analyze results from ITS studies and then extract and summarize characteristics of the methods used.

I agree with the authors that we currently know relatively little about the characteristics of methods used in meta-analyses of ITS results. From my point of view, I would like to have evidence that can be used to help design SRs that will meta-analyze ITS results, and to help understand potential weaknesses of such SRs. For example, I would like to have evidence to inform the number of pre- and post-interruption time points that should be required of ITS studies for inclusion in a meta-analysis, and information about whether certain ITS designs and analysis methods lead to excessive bias or imprecision. The protocol says that the resulting work will inform subsequent research that addresses these kinds of questions, using simulation and empirical evaluation. My understanding is that knowing about the landscape of methods that have been used in recent SRs will allow this subsequent work to address relevant questions (e.g., it could allow the authors to design simulation studies that usefully model what is being done in practice).

My review focuses mainly on the conceptual and statistical aspects of the protocol. I cannot comment on other aspects such as the literature search.

I have two "major" criticisms of the protocol:

1. Because I am more interested in the evidence that the subsequent research will hopefully deliver, I would like to see more detailed thought in this protocol about how the subsequent work will be performed. This should then inform what the product of the present protocol needs to deliver to ensure the success of the subsequent work. Perhaps this has already been thought through in detail and not presented here. However, if this work has not been done, I encourage the authors to do it and update the protocol.

2. The analysis is not planned in enough detail that it could be implemented without having to make important choices after having seen the data. I think this potentially leads to at two problems. First, being able to choose from among several possible analyses and means of presentation risks introducing bias. Second, more detailed planning at the protocol stage may prevent problems that would otherwise only become apparent while the work is being

done. I suggest that the authors substantially revise this section to specify in detail what they will do and how they will report their results, including preparing skeleton tables and/or figures. I think that thinking through these issues at the protocol stage will likely make this and subsequent work (see point 1 above) more efficient, and lead to higher quality papers.

I have the following "minor" comments and suggestions:

1. I suggest that the abstract is clarified to say that the authors will study the 100 **most recent** systematic reviews that include ITS analyses (and include the date range), rather than simply saying 100 (or whatever the final sample size is determined to be). My concern when reading the abstract was that the 100 SRs could be chosen arbitrarily, giving rise to potential bias.

2. It would be useful for the authors to clarify that, with respect to estimating a binomial proportion, their proposed sample size of 100 would give them a worst-case margin of error of **plus or minus** approximately ten percentage points (i.e., 40.2% to 59.8%), if the population parameter is 50%.

3. This margin of error is actually quite wide. I wonder if the authors have considered how plausible it is that the population parameter will often be close to the worst-case of 50%, and if so, whether the relatively wide confidence intervals will be informative enough for their subsequent work that will build on this paper?

4. The sample size of 100 seems to have been chosen under the assumption that a binomial proportion will be estimated for each factor studied (i.e., that each factor will have two levels). However, many of the criteria specified are factors with more than two levels (e.g., the protocol gives the example of three types of outcome that included reviews may study: continuous, count, and rate). Given that, I would encourage the authors to think about the more general case of estimating multinomial proportions. This would require a larger sample size for a worst-case scenario equivalent to that of a binomial distribution. A quick search identified Thompson 1987[1], which provides a table for estimating sample size for estimating multinomial proportions. Briefly, if the authors want to estimate multinomial proportions with 95% CIs that give a margin or error of plus or minus 10%, that paper shows that the authors should include at least 128 studies (irrespective of the number of levels of the factor studied). However, I encourage the authors to double-check this informal power analysis. The authors would then also need to use an appropriate method to estimate the proportions and their confidence intervals. Given the authors plan to use Stata, I think the correct way to do this is to estimate the proportions using syntax like "**mlogit myvar**" and then obtain the estimates and their confidence intervals via "**margins, predict()"**, but again I suggest the authors double-check this.

5. With respect to the "Selection of outcomes" section, it may be useful to rewrite item 1 in terms of "number of model parameters" rather than "number of effect measures". The meaning of the current text was not immediately clear to me.

6. I suggest the authors plan to extract a little more detail about the methods used to analyze ITS studies, so that this can be used to inform subsequent work. For example, I am interested to know how often "incorrect" model assumptions are assumed, and under what conditions meaningfully incorrect conclusions result from meta-analyses that include ITS

results from misspecified models. I'm thinking about the case where an outcome is bounded, where the observed pre- and/or post-interruption levels are close to the bounds, and the analysis assumes an error distribution whose domain includes values outside the bounds. For example, imagine the outcome "percentage of prescriptions with dosage errors", which is bounded to [0%, 100%]. If a normal error distribution is assumed in the ITS analysis, that distribution would incorrectly permit outcomes <0% and >100%. This may not be a problem if the residual standard deviation is small relative to the distance from the mean to the nearest bound, but if the observed data were close to 0% and/or 100% and the residual standard deviation was sufficiently large, the estimate of relative treatment effect estimate might be biased. I would then be interested in knowing, via empirical or simulation work from the planned subsequent research, whether these kinds of errors lead to meaningfully incorrect conclusions when such ITS results are brought into meta-analyses. If this kind of misspecification is uncommon, then perhaps that subsequent work is unnecessary, but my suspicion is that model misspecification (and this misspecification in particular) is actually very common.

7. I'm also interested in knowing what proportion of ITSs included in meta-analyses assume ITS models that are misspecified in terms of being overly simplistic (e.g., piecewise constant time series with an instantaneous post-intervention step-change). The protocol plans to survey which ITS models are used, but it's not clear whether it aims to judge whether the use is appropriate or not. This is to some extent a judgement about whether particular models are appropriate in particular contexts (all models are approximations of reality), but it would be useful to have a solid evidence base on which to make recommendations for what people should consider when they undertake ITS-based SRs.

8. The protocol talks about ITS analyses that include versus exclude controls (i.e., where the interruption does not occur), but it's not clear that the protocol will extract data on this. It would be interesting to know what proportion of SRs of ITS studies permit or include uncontrolled ITS results, and ultimately to have some information about whether inclusion of such studies leads to meaningfully incorrect conclusions (I assume it does, though I could imagine that it may be possible to include controlled and uncontrolled ITS results in a meta-regression and still reasonably estimate treatment effect). Similarly, it would be interesting to know how many controls are used (i.e., I assume that ITSs with one control are common but that 10 or 20 controls are quite rare), and then from the subsequent research, the number of controls that SR protocols should specify.

9. I suggest including the restriction to English as a possible limitation.

I wish the authors success with their research!

References

1. Thompson S: Sample Size for Estimating Multinomial Proportions. *The American Statistician*. 1987; **41** (1): 42-46 Publisher Full Text

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical modelling, biomedical research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 09 Jul 2020

Elizabeth Korevaar, Monash University, Melbourne, Australia

We would like to thank Dr Rose for their feedback on our protocol and the suggestions for its improvement. Below, we have addressed each of the items raised.
Major comments:
My review focuses mainly on the conceptual and statistical aspects of the protocol. I cannot comment on other aspects such as the literature search.

1. Because I am more interested in the evidence that the subsequent research will hopefully deliver, I would like to see more detailed thought in this protocol about how the subsequent work will be performed. This should then inform what the product of the present protocol needs to deliver to ensure the success of the subsequent work. Perhaps this has already been thought through in detail and not presented here. However, if this work has not been done, I encourage the authors to do it and update the protocol.

Thank you for your interest in our study and our future planned research. In this manuscript, we present the protocol for a systematic review of meta-analysis methods used to combine results from interrupted time series studies. While the results of our systematic review will inform subsequent research, outlining details of the future projects is beyond the scope of the current manuscript. However, we have provided more detail about the planned statistical simulation and empirical evaluation in the first paragraph of the 'Discussion': "The results of this review will inform our broader research program which aims to examine how different meta-analysis methods of ITS studies perform *and provide guidance on the methods. Specifically, the review will identify the range of statistical methods that are used in practice, characteristics of the times series studies included in the meta-analyses (e.g. number of series, length of series) and the types of effect measures used (e.g. level change, slope change). These characteristics will inform a statistical simulation study that will examine the performance of different methods for meta-analysis of ITS studies. In addition, we will identify reviews for which the raw time series data of the included ITS*

***studies are reported. These reviews will be used in an empirical evaluation to examine the impact of using different meta-analysis methods of real ITS studies.*". In addition, we have provided the current version of our data extraction form as an additional file (Appendix 2) in the 'Extended data' section. The data extraction form provides details of the data to be extracted and summarised.**

2. The analysis is not planned in enough detail that it could be implemented without having to make important choices after having seen the data. I think this potentially leads to at two problems. First, being able to choose from among several possible analyses and means of presentation risks introducing bias. Second, more detailed planning at the protocol stage may prevent problems that would otherwise only become apparent while the work is being done. I suggest that the authors substantially revise this section to specify in detail what they will do and how they will report their results, including preparing skeleton tables and/or figures. I think that thinking through these issues at the protocol stage will likely make this and subsequent work (see point 1 above) more efficient, and lead to higher quality papers.

We have added further detail in the 'Analysis' section about the general table structure we plan to use to present the results (see following). In addition, we provide the data extraction form which includes all the items we are collecting.
"We will summarise the characteristics of included systematic reviews with descriptive statistics. For categorical data (e.g. the meta-analysis approach used, the risk of bias tool used) we will present frequencies (with percentages), and for numerical data (e.g. the number of meta-analysed ITS studies, the number of pooled estimates) we will present means (with standard deviations) or medians (with interquartile range)."
The items (their response options – see the data extraction form in Appendix 2) and summary statistics will be grouped into tables using a structure such as the following:

1. **General study characteristics (e.g. research disciplines; types of interventions; definition used to classify interrupted time series; risk of bias or methodological quality assessment of included studies undertaken)**
2. **Included study characteristics (e.g. number of included ITS studies; average number of points pre- and post-interruption)**
3. **Meta-analysis methods (e.g. justification given for chosen meta-analysis method; re-analysis of primary study data undertaken, one-stage or two-stage meta-analysis used, chosen meta-analysis model, heterogeneity estimator used)**
4. **Effect measures used (e.g. number of effect measure, which effect measures, standardisation of the effect measure)**
5. **Additional analyses (e.g. sensitivity analyses, subgroup analyses)**
6. **Completeness of reporting (e.g. reporting of combined effect, reporting of measure of precision)**

Minor comments:

1. I suggest that the abstract is clarified to say that the authors will study the 100 most recent systematic reviews that include ITS analyses (and include the date range), rather than simply saying 100 (or whatever the final sample size is determined to be). My concern when

reading the abstract was that the 100 SRs could be chosen arbitrarily, giving rise to potential bias.

We have revised the text in the 'Abstract' as follows:
"We will identify the 100 most recent reviews (published between 1 January 2000 and 11 October 2019) that include meta-analyses of ITS studies from a search of eight electronic databases covering several disciplines (public health, psychology, education, economics)."

2. It would be useful for the authors to clarify that, with respect to estimating a binomial proportion, their proposed sample size of 100 would give them a worst-case margin of error of **plus or minus** approximately ten percentage points (i.e., 40.2% to 59.8%), if the population parameter is 50%.

We have revised the text in the 'Sample size' section to:
"Our sample size of 100 reviews was primarily selected for reasons of feasibility. A sample of this size will allow estimation of the percentage of reviews with a particular element (e.g. the prevalence of the reviews that re-analyse the primary study data) to within a maximum margin of error of 10% (assuming a prevalence of 50%). This margin of error represents the worst-case scenario and will decrease if the prevalence varies from 50%."

3. This margin of error is actually quite wide. I wonder if the authors have considered how plausible it is that the population parameter will often be close to the worst-case of 50%, and if so, whether the relatively wide confidence intervals will be informative enough for their subsequent work that will build on this paper?

We believe that our sample size will generally be sufficient, such that our interpretation of the confidence interval limits will be consistent. For example, if we found that the percentage of reviews that use a particular method was 10% (95%CI: 4% to 16%), our interpretation of the limits of the confidence interval would lead to the same conclusion that the method was not commonly used. As another example, if the outcome was complete reporting of effect estimates (i.e., reporting the effect estimate and a measure of precision), and 50% of the reviews were found to completely report the effect, our interpretation would not differ at the limits of the confidence interval; if the true percentage was 40% or 60%, we would be concerned.

4. The sample size of 100 seems to have been chosen under the assumption that a binomial proportion will be estimated for each factor studied (i.e., that each factor will have two levels). However, many of the criteria specified are factors with more than two levels (e.g., the protocol gives the example of three types of outcome that included reviews may study: continuous, count, and rate). Given that, I would encourage the authors to think about the more general case of estimating multinomial proportions. This would require a larger sample size for a worst-case scenario equivalent to that of a binomial distribution. A quick search identified Thompson 1987[1], which provides a table for estimating sample size for estimating multinomial proportions. Briefly, if the authors want to estimate multinomial proportions with 95% CIs that give a margin or error of plus or minus 10%, that paper

shows that the authors should include at least 128 studies (irrespective of the number of levels of the factor studied). However, I encourage the authors to double-check this informal power analysis. The authors would then also need to use an appropriate method to estimate the proportions and their confidence intervals. Given the authors plan to use Stata, I think the correct way to do this is to estimate the proportions using syntax like **"mlogit myvar"** and then obtain the estimates and their confidence intervals via **"margins, predict()"**, but again I suggest the authors double-check this.

As noted in the 'Sample size' section, our sample size of 100 reviews was primarily selected for reasons of feasibility, so we do not have the ability to increase the sample size. Importantly, however, most items are binary and not multinomial. We acknowledge that for those few multinomial items, we will have less precision.

5. With respect to the "Selection of outcomes" section, it may be useful to rewrite item 1 in terms of "number of model parameters" rather than "number of effect measures". The meaning of the current text was not immediately clear to me.

We have chosen the term "effect measures" since this is the terminology used in the Cochrane Handbook (Chapter 10: Analysing data and undertaking meta-analyses). The "number of model parameters" is not interchangeable with the "number of effect measures" because they have different meanings. Effect measures in ITS studies are generally computed as a combination of the model parameter estimates (e.g. area under the curve, long-term change in level).
Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.

6. I suggest the authors plan to extract a little more detail about the methods used to analyze ITS studies, so that this can be used to inform subsequent work. For example, I am interested to know how often "incorrect" model assumptions are assumed, and under what conditions meaningfully incorrect conclusions result from meta-analyses that include ITS results from misspecified models. I'm thinking about the case where an outcome is bounded, where the observed pre- and/or post-interruption levels are close to the bounds, and the analysis assumes an error distribution whose domain includes values outside the bounds. For example, imagine the outcome "percentage of prescriptions with dosage errors", which is bounded to [0%, 100%]. If a normal error distribution is assumed in the ITS analysis, that distribution would incorrectly permit outcomes <0% and >100%. This may not be a problem if the residual standard deviation is small relative to the distance from the mean to the nearest bound, but if the observed data were close to 0% and/or 100% and the residual standard deviation was sufficiently large, the estimate of relative treatment effect estimate might be biased. I would then be interested in knowing, via empirical or simulation work from the planned subsequent research, whether these kinds of errors lead to meaningfully incorrect conclusions when such ITS results are brought into meta-analyses. If this kind of misspecification is uncommon, then perhaps that subsequent work is unnecessary, but my suspicion is that model misspecification (and this misspecification in

particular) is actually very common.

The issue outlined is interesting and it would be useful to examine under what conditions meaningfully incorrect conclusions result from meta-analyses that include ITS results from misspecified models in a simulation study. In our study we will document whether the reviewers re-analysed the primary study data (see row 90 of the data extraction form, Appendix 2); and which analysis methods were used for the primary studies, including whether the reviewers made a judgement about the appropriateness of the analysis methods used in the primary studies (see rows 93-104 of the data extraction form).

7. I'm also interested in knowing what proportion of ITSs included in meta-analyses assume ITS models that are misspecified in terms of being overly simplistic (e.g., piecewise constant time series with an instantaneous post-intervention step-change). The protocol plans to survey which ITS models are used, but it's not clear whether it aims to judge whether the use is appropriate or not. This is to some extent a judgement about whether particular models are appropriate in particular contexts (all models are approximations of reality), but it would be useful to have a solid evidence base on which to make recommendations for what people should consider when they undertake ITS-based SRs.

We will extract information on the model structure when the reviewers re-analyse the included ITS studies (see row 107 of the data extraction form, Appendix 2). From this information, we will summarise the different model structures (e.g. level change only, slope change only). We do not plan to judge the appropriateness of the model structures given, as the reviewer pointed out, this would require content knowledge of the studies. In addition, where reviewers directly use results from the primary studies, we will collect the review authors' judgements of the appropriateness of the methods, which could include model structure misspecification (see rows 93-98).

8. The protocol talks about ITS analyses that include versus exclude controls (i.e., where the interruption does not occur), but it's not clear that the protocol will extract data on this. It would be interesting to know what proportion of SRs of ITS studies permit or include uncontrolled ITS results, and ultimately to have some information about whether inclusion of such studies leads to meaningfully incorrect conclusions (I assume it does, though I could imagine that it may be possible to include controlled and uncontrolled ITS results in a meta-regression and still reasonably estimate treatment effect). Similarly, it would be interesting to know how many controls are used (i.e., I assume that ITSs with one control are common but that 10 or 20 controls are quite rare), and then from the subsequent research, the number of controls that SR protocols should specify.

We will collect details on whether and how control series were incorporated into the meta-analysis (see rows 69-71 of the data extraction form, Appendix 2). Additionally, we will capture information regarding sensitivity analysis (e.g. comparing inclusion versus exclusion of control series, if performed). However, in this study, we will not assess whether or not the conclusion of the meta-analysis would change based on the inclusion of the control series.

9. I suggest including the restriction to English as a possible limitation.

We have amended the following paragraph in the 'Discussion' section to include the English language limitation:
"While the study will be limited by our ability to identify all potentially eligible reviews and meta-analyses of ITS studies, our search strategy attempts to capture the various ways these studies are described. However, given ITS studies are often not identified as such[16], and that our search is restricted to articles written in English, it is likely that we will not capture all reviews and meta-analyses that include ITS studies. Conversely, we may end up including reviews where no information regarding the definition of the included ITS studies is provided, or where an inappropriate label of ITS has been applied to included studies. While we will not exclude these reviews, we will record the reviewers' definition of an ITS study."

Kind regards,
Elizabeth Korevaar

Competing Interests: No competing interests were disclosed.